Online ISSN : 0975-4172
Print ISSN : 0975-4350

Global Journal of Computer Science and Technology: C

Software & Data Engineering

Highlights

Concurrent Access

Risk Level of Heart

Mobile Media SPL

Gene Expression Analysis

Discovering Thoughts, Inventing Future

VOLUME 14 ISSUE 3 VERSION 1.0

Global Journal of Computer Science and Technology: C
Software & Data Engineering

Global Journal of Computer Science and Technology: C
Software & Data Engineering

Volume 14 Issue 3 (Ver. 1.0)

Open Association of Research Society

The opinions and statements made in this book are those of the authors concerned. Ultraculture has not verified and neither confirms nor denies any of the foregoing and no warranty or fitness is implied.

Engage with the contents herein at your own risk.

The use of this journal, and the terms and conditions for our providing information, is governed by our Disclaimer, Terms and Conditions and Privacy Policy given on our website http://globaljournals.us/terms-and-condition/menu-id-1463/

By referring / using / reading / any type of association / referencing this journal, this signifies and you acknowledge that you have read them and that you accept and will be bound by the terms thereof.

All information, journals, this journal, activities undertaken, materials, services and our website, terms and conditions, privacy policy, and this journal is subject to change anytime without any prior notice.

Incorporation No.: 0423089
License No.: 42125/022010/1186
Registration No.: 430374
Import-Export Code: 1109007027
Employer Identification Number (EIN):
USA Tax ID: 98-0673427

Global Journals Inc.

*(A Delaware USA Incorporation with "Good Standing"; **Reg. Number: 0423089**)*

Sponsors: Open Association of Research Society
Open Scientific Standards

Publisher's Headquarters office

Global Journals Headquarters
301st Edgewater Place Suite, 100 Edgewater Dr.-Pl,
Wakefield MASSACHUSETTS, Pin: 01880,
United States of America
USA Toll Free: +001-888-839-7392
USA Toll Free Fax: +001-888-839-7392

Offset Typesetting

Global Journals Incorporated
2nd, Lansdowne, Lansdowne Rd., Croydon-Surrey,
Pin: CR9 2ER, United Kingdom

Packaging & Continental Dispatching

Global Journals
E-3130 Sudama Nagar, Near Gopur Square,
Indore, M.P., Pin:452009, India

Find a correspondence nodal officer near you

To find nodal officer of your country, please email us at *local@globaljournals.org*

eContacts

Press Inquiries: *press@globaljournals.org*
Investor Inquiries: *investors@globaljournals.org*
Technical Support: *technology@globaljournals.org*
Media & Releases: *media@globaljournals.org*

Pricing (Including by Air Parcel Charges):

For Authors:
22 USD (B/W) & 50 USD (Color)
Yearly Subscription (Personal & Institutional):
200 USD (B/W) & 250 USD (Color)

Integrated Editorial Board
(Computer Science, Engineering, Medical, Management, Natural Science, Social Science)

Dr. Bart Lambrecht
Director of Research in Accounting and FinanceProfessor of Finance
Lancaster University Management School
BA (Antwerp); MPhil, MA, PhD (Cambridge)

Dr. Carlos García Pont
Associate Professor of Marketing
IESE Business School, University of Navarra
Doctor of Philosophy (Management), Massachusetts Institute of Technology (MIT)
Master in Business Administration, IESE, University of Navarra
Degree in Industrial Engineering, Universitat Politècnica de Catalunya

Dr. Fotini Labropulu
Mathematics - Luther College
University of ReginaPh.D., M.Sc. in Mathematics
B.A. (Honors) in Mathematics
University of Windso

Dr. Lynn Lim
Reader in Business and Marketing
Roehampton University, London
BCom, PGDip, MBA (Distinction), PhD, FHEA

Dr. Mihaly Mezei
ASSOCIATE PROFESSOR
Department of Structural and Chemical Biology, Mount Sinai School of Medical Center
Ph.D., Etvs Lornd University
Postdoctoral Training,
New York University

Dr. Söhnke M. Bartram
Department of Accounting and FinanceLancaster University Management SchoolPh.D. (WHU Koblenz)
MBA/BBA (University of Saarbrücken)

Dr. Miguel Angel Ariño
Professor of Decision Sciences
IESE Business School
Barcelona, Spain (Universidad de Navarra)
CEIBS (China Europe International Business School).
Beijing, Shanghai and Shenzhen
Ph.D. in Mathematics
University of Barcelona
BA in Mathematics (Licenciatura)
University of Barcelona

Philip G. Moscoso
Technology and Operations Management
IESE Business School, University of Navarra
Ph.D in Industrial Engineering and Management, ETH Zurich
M.Sc. in Chemical Engineering, ETH Zurich

Dr. Sanjay Dixit, M.D.
Director, EP Laboratories, Philadelphia VA Medical Center
Cardiovascular Medicine - Cardiac Arrhythmia
Univ of Penn School of Medicine

Dr. Han-Xiang Deng
MD., Ph.D
Associate Professor and Research Department Division of Neuromuscular Medicine
Davee Department of Neurology and Clinical NeuroscienceNorthwestern University
Feinberg School of Medicine

President Editor (HON.)

Chief Author (HON.)

Dean & Editor-in-Chief (HON.)

Contents of the Volume

Global Journals Inc.

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 14 Issue 3 Version 1.0 Year 2014
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Concurrent Access Algorithms for Different Data Structures: A Research Review

By Ms. Ranjeet Kaur & Dr. Pushpa Rani Suri

Kurukshetra University, India

Abstract - Algorithms for concurrent data structure have gained attention in recent years as multi-core processors have become ubiquitous. Several features of shared-memory multiprocessors make concurrent data structures significantly more difficult to design and to verify as correct than their sequential counterparts. The primary source of this additional difficulty is concurrency. This paper provides an overview of the some concurrent access algorithms for different data structures.

Keywords: concurrency, lock-free, non-blocking, mem-ory management, compares and swap, elimination.

GJCST-C Classification : E.1

Concurrent Access Algorithms for Different Data Structures: A Research Review

Ms. Ranjeet Kaur [α] & Dr. Pushpa Rani Suri [σ]

Abstract- Algorithms for concurrent data structure have gained attention in recent years as multi-core processors have become ubiquitous. Several features of shared-memory multiprocessors make concurrent data structures significantly more difficult to design and to verify as correct than their sequential counterparts. The primary source of this additional difficulty is concurrency. This paper provides an overview of the some concurrent access algorithms for different data structures.

Keywords: concurrency, lock-free, non-blocking, memory management, compares and swap, elimination.

I. Introduction

A concurrent data structure is a particular way of storing and organizing data for access by multiple computing threads (or processes) on a computer. The proliferation of commercial shared-memory multiprocessor machines has brought about significant changes in the art of concurrent programming. Given current trends towards low cost chip multithreading (CMT), such machines are bound to become ever more widespread. Shared-memory multiprocessors are systems that concurrently execute multiple threads of computation which communicate and synchronize through data structures in shared memory. Designing concurrent data structures and ensuring their correctness is a difficult task, significantly more challenging than doing so for their sequential counterparts. The difficult of concurrency is aggravated by the fact that threads are asynchronous since they are subject to page faults, interrupts, and so on. To manage the difficulty of concurrent programming, multithreaded applications need synchronization to ensure thread-safety by coordinating the concurrent accesses of the threads. At the same time, it is crucial to allow many operations to make progress concurrently and complete without interference in order to utilize the parallel processing capabilities of contemporary architectures. The traditional way to implement shared data structures is to use mutual exclusion (locks) to ensure that concurrent operations do not interfere with one another. Locking has a number of disadvantages with respect to software engineering, fault-tolerance, and scalability. In response, researchers have investigated a variety of alternative synchronization techniques that do not employ mutual exclusion. A synchronization technique is wait-free if it ensures that every thread will continue to make progress in the face of arbitrary delay (or even failure) of other threads. It is lock-free if it ensures only that some thread always makes progress. While wait-free synchronization is the ideal behavior (thread starvation is unacceptable), lock-free synchronization is often good enough for practical purposes (as long as starvation, while possible in principle, never happens in practice).The synchronization primitives provided by most modern architectures, such as compare-and-swap (CAS) or load-locked/store-conditional (LL/SC) are powerful enough to achieve wait-free (or lock-free) implementations of any linearizable data object [23]. The remaining paper will discussed about the different data structures, concurrency control methods and various techniques given for the concurrent access to these data structures.

II. Data Structures

Data can be organized in many ways and a data structure is one of these ways. It is used to represent data in the memory of the computer so that the processing of data can be done in easier way. In other words, data structures are the logical and mathematical model of a particular organization of data. Different kinds of data structures are suited to different kinds of applications, and some are highly specialized to specific tasks. For example, B-trees are particularly well-suited for implementation of databases, while compiler implementations usually use hash tables to look up identifiers. A data structure can be broadly classified into (i) Primitive data structure (ii) Non-primitive data structure.

Primitive Data Structure: The data structures, typically those data structure that are directly operated upon by machine level instructions i.e. the fundamental data types such as int, float.

Non-Primitive Data Structure: The data structures, which are not primitive, are called non-primitive data structures. There are two types of-primitive data structures.

a) Linear Data Structures

A list, which shows the relationship of adjacency between elements, is said to be linear data structure. The most, simplest linear data structure is a 1-D array, but because of its deficiency, list is frequently used for different kinds of data.

Author α σ: Kurukshetra University, Kurukshetra.
e-mails: kaurranjeet 2203@gmail.com, pushpa.suri@yahoo.com

A [0]	A[1]	A[2]	A[3]	A[4]	A[5]

Figure 1 : A 1-D Array of 6 Elements.

b) *Non-Linear Data Structure*

A list, which doesn't show the relationship of adjacency between elements, is said to be non-linear data structure.

i. *Linear Data Structure*

A list is an ordered list, which consists of different data items connected by means of a link or pointer. This type of list is also called a linked list. A linked list may be a single list or double linked list.

- *Single linked list:* A single linked list is used to traverse among the nodes in one direction.

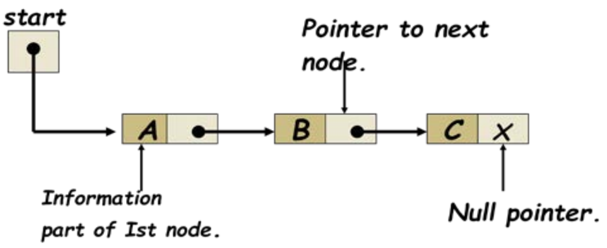


Figure 2 : A single three Nodes linked list.

Double linked list: A double linked list is used to traverse among the nodes in both the directions.
A list has two subsets. They are: -

Stack: It is also called as last-in-first-out (LIFO) system. It is a linear list in which insertion and deletion take place only at one end. It is used to evaluate different expressions.

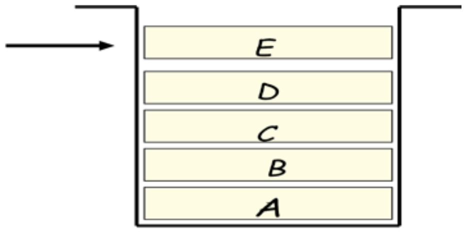


Figure 3 : A Stack with Elements.

- *Queue:* It is also called as first-in-first-out (FIFO) system. It is a linear list in which insertion takes place at once end and deletion takes place at other end. It is generally used to schedule a job in operating systems and networks.

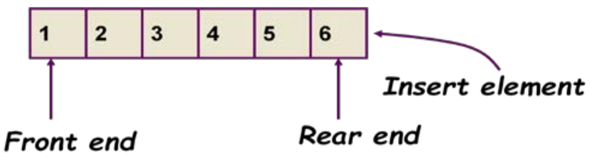


Figure 4 : A Queue with 6 Elements.

ii. *Non-Linear Data Structure*

The frequently used non-linear data structures are

- *Trees:* It maintains hierarchical relationship between various elements

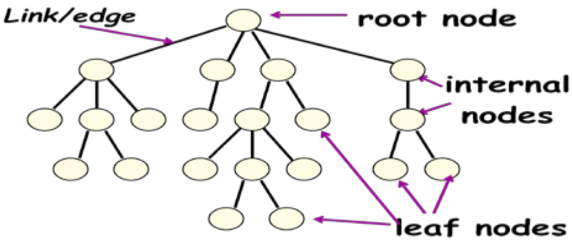


Figure 5 : A Binary Tree.

- *Graphs:* It maintains random relationship or point-to-point relationship between various elements.

III. Concurrency Control

Simultaneous execution of multiple threads/process over a shared data structure access can create several data integrity and consistency problems:

- Lost Updates.
- Uncommitted Data.
- Inconsistent retrievals

All above are the reasons for introducing the concurrency control over the concurrent access of shared data structure. Concurrent access to data structure shared among several processes must be synchronized in order to avoid conflicting updates. Synchronization is referred to the idea that multiple processes are to join up or handshake at a certain points, in order to reach agreement or commit to a certain sequence of actions. The thread synchronization or serialization strictly defined is the application of particular mechanisms to ensure that two concurrently executing threads or processes do not execute specific portions of a program at the same time. If one thread has begun to execute a serialized portion of the program, any other thread trying to execute this portion must wait until the first thread finishes.
Concurrency control techniques can be divided into two categories.

- Blocking
- Non-blocking

Both of these are discussed in below sub-sections.

a) *Blocking*

Blocking algorithms allow a slow or delayed process to prevent faster processes from completing operations on the shared data structure indefinitely. On asynchronous (especially multiprogrammed) multiprocessor systems, blocking algorithms suffer significant performance degradation when a process is halted or delayed at an inopportune moment. Many of the existing concurrent data structure algorithms that have been developed use mutual exclusion i.e. some form of locking.

Mutual exclusion degrades the system's overall performance as it causes blocking, due to that other concurrent operations cannot make any progress while the access to the shared resource is blocked by the lock. The limitation of blocking approach are given below

- *Priority Inversion:* occurs when a high-priority process requires a lock holded by a lower-priority process.
- *Convoying:* occurs when a process holding a lock is rescheduled by exhausting its quantum, by a page fault or by some other kind of interrupt. In this case, running processes requiring the lock are unable to progress.
- *Deadlock:* can occur if different processes attempt to lock the same set of objects in different orders.
- Locking techniques are not suitable in a real-time context and more generally, they suffer significant performance degradation on multiprocessors systems.

b) Non-Blocking

Non-blocking algorithm Guarantees that the data structure is always accessible to all processes and an inactive process cannot render the data structure inaccessible. Such an algorithm ensures that some active process will be able to complete an operation in a finite number of steps making the algorithm robust with respect to process failure [22]. In the following sections we discuss various non-blocking properties with different strength.

- *Wait-Freedom:* A method is wait-free if every call is guaranteed to finish in a finite number of steps. If a method is bounded wait-free then the number of steps has a finite upper bound, from this definition it follows that wait-free methods are never blocking, therefore deadlock cannot happen. Additionally, as each participant can progress after a finite number of steps (when the call finishes), wait-free methods are free of starvation.
- *Lock-Freedom:* Lock-freedom is a weaker property than wait-freedom. In the case of lock-free calls, infinitely often some method finishes in a finite number of steps. This definition implies that no deadlock is possible for lock-free calls. On the other hand, the guarantee that some call finishes in a finite number of steps is not enough to guarantee that all of them eventually finish. In other words, lock-freedom is not enough to guarantee the lack of starvation.
- *Obstruction-Freedom:* Obstruction-freedom is the weakest non-blocking guarantee discussed here. A method is called obstruction-free if there is a point in time after which it executes in isolation (other threads make no steps, e.g.: become suspended), it finishes in a bounded number of steps. All lock-free objects are obstruction-free, but the opposite is generally not true. Optimistic concurrency control (OCC) methods are usually obstruction-free. The OCC approach is that every participant tries to execute its operation on the shared object, but if a participant detects conflicts from others, it rolls back the modifications, and tries again according to some schedule. If there is a point in time, where one of the participants is the only one trying, the operation will succeed.

In the sequential setting, data structures are crucially important for the performance of the respective computation. In the parallel programming setting, their importance becomes more crucial because of the increased use of data and resource sharing for utilizing parallelism. In parallel programming, computations are split into subtasks in order to introduce parallelization at the control/computation level. To utilize this opportunity of concurrency, subtasks share data and various resources (dictionaries, buffers, and so forth). This makes it possible for logically independent programs to share various resources and data structures.

Concurrent data structure designers are striving to maintain consistency of data structures while keeping the use of mutual exclusion and expensive synchronization to a minimum, in order to prevent the data structure from becoming a sequential bottleneck. Maintaining consistency in the presence of many simultaneous updates is a complex task. Standard implementations of data structures are based on locks in order to avoid inconsistency of the shared data due to concurrent modifications. In simple terms, a single lock around the whole data structure may create a bottleneck in the program where all of the tasks serialize, resulting in a loss of parallelism because too few data locations are concurrently in use. Deadlocks, priority inversion, and convoying are also side-effects of locking. The risk for deadlocks makes it hard to compose different blocking data structures since it is not always possible to know how closed source libraries do their locking. Lock-free implementations of data structures support concurrent access. They do not involve mutual exclusion and make sure that all steps of the supported operations can be executed concurrently. Lock-free implementations employ an optimistic conflict control approach, allowing several processes to access the shared data object at the same time. They suffer delays only when there is an actual conflict between operations that causes some operations to retry. This feature allows lock-free algorithms to scale much better when the number of processes increases. An implementation of a data structure is called lock-free if it allows multiple processes/threads to access the data structure concurrently and also guarantees that at least one operation among those finishes in a finite number of its own steps regardless of the state of the other operations. A consistency (safety) requirement for lock-free data structures is linearizability [24], which ensures that each operation on the data appears to take effect instantaneously during its actual duration and the effect of all operations are consistent with the object's sequential specification. Lock-free data structures offer several advantages over their blocking counterparts, such as being immune to deadlocks, priority inversion, and convoying, and have been shown to work well in practice in many different settings [26, 25].

The remaining paper will explore the access of different data structures like stack, queue, trees, priority queue, and linked list in concurrent environment. How the sequence of data structure operations changes during concurrent access. These techniques will be based on blocking and non-blocking.

IV. Literature Review

a) Stack Data Structure

Stack is the simplest sequential data structures. Numerous issues arise in designing concurrent versions of these data structures, clearly illustrating the challenges involved in designing data structures for shared-memory multiprocessors. A concurrent stack is a data structure linearizable to a sequential stack that provides push and pop operations with the usual LIFO semantics. Various alternatives exist for the behavior of these data structures in full or empty states, including returning a special value indicating the condition, raising an exception, or blocking.

There are several lock-based concurrent stack implementations in the literature. Typically, lock-based stack algorithms are expected to offer limited robustness.

The first non-blocking implementation of concurrent link based stack was first proposed by Trieber et al [1]. It represented the stack as a singly linked list with a top pointer. It uses compare-and-swap to modify the value of Top atomically. However, this stack was very simple and can be expected to be quite efficient, but no performance results were reported for nonblocking stacks. When Michael et. al [2] compare the performance of Treiber's stack to an optimized nonblocking algorithm based on Herlihy's methodology [28], and several lock-based stacks such as an MCS lock in low load situations[29]. They concluded that Treiber's algorithm yields the best overall performance, but this performance gap increases as the degree of multiprogramming grows. All this happen due to contention and an inherent sequential bottleneck.

Hendler et al. [3] observe that any stack implementation can be made more scalable using the elimination technique [23]. Elimination allows pairs of operations with reverse semantics like pushes and pops on a stack-to complete without any central coordination, and therefore substantially aids scalability. The idea is that if a pop operation can find a concurrent push operation to "partner" with, then the pop operation can take the push operation's value, and both operations can return immediately.

b) Queue Data Structure

A concurrent queue is a data structure that provides enqueue and dequeue operations with the usual FIFO semantics. Valois et.al [4] presented a list-based nonblocking queue. The represented algorithm allows more concurrency by keeping a dummy node at the head (dequeue end) of a singly linked list, thus simplifying the special cases associated with empty and single-item. Unfortunately, the algorithm allows the tail pointer to lag behind the head pointer, thus preventing dequeuing processes from safely freeing or reusing dequeued nodes. If the tail pointer lags behind and a process frees a dequeued node, the linked list can be broken, so that subsequently enqueued items are lost. Since memory is a limited resource, prohibiting memory reuse is not an acceptable option. Valois therefore proposed a special mechanism to free and allocate memory. The mechanism associates a reference counter with each node. Each time a process creates a pointer to a node it increments the node's reference counter atomically. When it does not intend to access a node that it has accessed before, it decrements the associated reference counter atomically. In addition to temporary links from process-local variables, each reference counter reflects the number of links in the data structure that point to the node in question. For a queue, these are the head and tail pointers and linked-list links. A node is freed only when no pointers in the data structure or temporary variables point to it. Drawing ideas from the previous authors, Michel et.al [5] presented a new non-blocking concurrent queue algorithm, which is simple, fast, and practical. The algorithm implements the queue as a singly-linked list with Head and Tail pointers. Head always points to a dummy node, which is the first node in the list. Tail points to either the last or second to last node in the list. The algorithm uses compare and swap, with modification counters to avoid the ABA problem. To allow dequeuing processes to free dequeue nodes, the dequeue operation ensures that Tail does not point to the dequeued node nor to any of its predecessors. This means that dequeued nodes may safely be re-used.

The Mark et al [6] introduced a scaling technique for queue data structure which was earlier applied to LIFO data structures like stack. They transformed existing nonscalable FIFO queue implementations into scalable implementations using the elimination technique, while preserving lock-freedom and linearizability.

In all previously FIFO queue algorithms, concurrent Enqueue and Dequeue operations synchronized on a small number of memory locations, such algorithms can only allow one Enqueue and one Dequeue operation to complete in parallel, and therefore cannot scale to large numbers of concurrent operations. In the LIFO structures elimination works by allowing opposing operations such as pushes and pops to exchange values in a pair wise distributed fashion without synchronizing on a centralized data structure. This technique was straightforward in LIFO ordered structures [23]. However, this approach seemingly contradicts in a queue data structure, a Dequeue operation must take the oldest value currently waiting in the queue. It apparently cannot eliminate with a concurrent Enqueue. For example, if a queue contains a single value 1, then after an Enqueue of 2 and a Dequeue, the queue contains 2, regardless of the order of these operations.

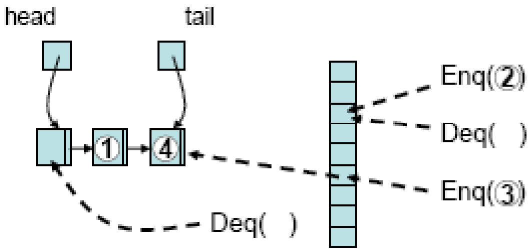


Figure 6 : Shows an Example Execution

Thus, because the queue changes, we cannot simply eliminate the Enqueue and Dequeue. In a empty queue , we could eliminate an Enqueue-Dequeue pair, because in this case the queue is unchanged by an Enqueue immediately followed by a Dequeue. In case when queue is non empty , we must be aware with linearizability correctness condition [24,25], which requires that we can order all operations in such a way that the operations in this order respect the FIFO queue semantics, but also so that no process can detect that the operations did not actually occur in this order. If one operation completes before another begins, then we must order them in this order. Otherwise, if the two are concurrent, we are free to order them however we wish. Key to their approach was the observation that they wanted to use elimination when the load on the queue is high. In such cases, if an Enqueue operation is unsuccessful in an attempt to access the queue, it will generally back off before retrying. If in the meantime all values that were in the queue when the Enqueue began are dequeued, then we can “pretend” that the Enqueue did succeed in adding its value to the tail of the queue earlier, and that it now has reached the head and can be dequeued by an eliminating Dequeue. Thus, they used time spent backing off to “age” the unsuccessful Enqueue operations so that they become “ripe” for elimination. Because this time has passed, we ensure that the Enqueue operation is concurrent with Enqueue operations that succeed on the central queue, and this allows us to order the Enqueue before some of them, even though it never succeeds on the central queue. The key is to ensure that Enqueues are eliminated only after sufficient aging.

c) Linked List Data Structure

Implementing linked lists efficiently is very important, as they act as building blocks for many other data structures. The first implementation designed for lock-free linked lists was presented by Valois et .al [19]. The main idea behind this approach was to maintain auxiliary nodes in between normal nodes of the list in order to resolve the problems that arise because of interference between concurrent operations. Also, each node in his list had a backlink pointer which was set to point to the predecessor when the node was deleted. These backlinks were then used to backtrack through the list when there was interference from a concurrent deletion. Another lock-free implementation of linked lists was given by Harris et. al[20]. His main idea was to mark a node before deleting it in order to prevent concurrent operations from changing its right pointer. The previous approach was simpler than later one. Yet another implementation of a lock-free linked list was proposed by Michael [21]. The represented Technique used [20] design to implement the lock free linked list structure. The represented algorithm was compatible with efficient memory management techniques unlike [20] algorithm.

d) Tree Data Structure

A concurrent implementation of any search tree can be achieved by protecting it using a single exclusive lock. Concurrency can be improved somewhat by using a reader-writer lock to allow all read-only (search) operations to execute concurrently with each other while holding the lock.

Kung and Lehman et al. [7] presented a concurrent binary search tree implementation in which update operations hold only a constant number of node locks at a time, and these locks only exclude other update operations: search operations are never blocked. However, this implementation makes no attempt to keep the search tree balanced.

In the context of B+-trees Lehman et al.[8] has expanded some of the ideas of previous technique. The algorithm has property that any process for manipulating the tree uses only a small number of locks at any time, no search through the tree is ever prevented from reading any node, for that purpose they have considered a variant of B* -Tree called Blink- tree.

The Blink-tree is a B*-tree modified by adding a single “link” pointer field to each node This link field points to the next node at the same level of the tree as the current node, except that the link pointer of the rightmost node on a level is a null pointer. This definition for link pointers is consistent, since all leaf nodes lie at the same level of the tree. The Blink-tree has all of the nodes at a particular level chained together into a linked list.

In fact, in [8] algorithm, update operations as well as search operations use the lock coupling technique so that no operation ever holds more than two locks at a time, which significantly improves concurrency. This technique has been further refined, so that operations never hold more than one lock at a time [9]. The presented algorithm not addressed how nodes can be merged, instead allowing delete operations to leave nodes underfull. They argue that in many cases delete operations are rare, and that if space utilization becomes a problem, the tree can occasionally be reorganized in "batch" mode by exclusively locking the entire tree. Lanin et al. [10] incorporate merging into the delete operations, similarly to how insert operations split overflowed nodes in previous implementations. Similar to [8] technique, these implementations use links to allow recovery by operations that have mistakenly reached a node that has been evacuated due to node merging. In all of the algorithms discussed above, the maintenance operations such as node splitting and merging (where applicable) are performed as part of the regular update operations.

e) Priority Queue Data Structure

The Priority Queue abstract data type is a collection of items which can efficiently support finding the item with the highest priority. Basic operations are Insert (add an item), FindMin (finds the item with minimum (or maximum) priority), and DeleteMin (removes the item with minimum (or maximum) priority). DeleteMin returns the item removed.

- *Heap-Based Priority Queues:* Many of the concurrent priority queue constructions in the literature are linearizable versions of the heap structures. Again, the basic idea is to use fine-grained locking of the individual heap nodes to allow threads accessing different parts of the data structure to do so in parallel where possible. A key issue in designing such concurrent heaps is that traditionally insert operations proceed from the bottom up and delete-min operations from the top down, which creates potential for deadlock. Biswas et al. [11] present such a lock-based heap algorithm assuming specialized "cleanup" threads to overcome deadlocks. Rao et al. [12] suggest to overcome the drawbacks of [11] using an algorithm that has both insert and delete-min operations proceed from the top down. Ayani et.al [13] improved on their algorithm by suggesting a way to have consecutive insertions be performed on opposite sides of the heap. Hunt et al. [14] present a heap based algorithm that overcomes many of the limitations of the above schemes, especially the need to acquire multiple locks along the traversal path in the heap. It proceeds by locking for a short duration a variable holding the size of the heap and a lock on either the first or last element of the heap. In order to increase parallelism, insertions traverse the heap bottom-up while deletions proceed top-down, without introducing deadlocks. Insertions also employ a left-right technique as in [13] to allow them to access opposite sides on the heap and thus minimize interference.

Unfortunately, the empirical evidence shows, the performance of [14] does not scale beyond a few tens of concurrent processors. As concurrency increases, the algorithm's locking of a shared counter location, introduces a sequential bottleneck that hurts performance. The root of the tree also becomes a source of contention and a major problem when the number of processors is in the hundreds. In summary, both balanced search trees and heaps suffer from the typical scalability impediments of centralized structures: sequential bottlenecks and increased contention. The solution proposed by lotal et.al [15] is to design concurrent priority queues based on the highly distributed SkipList data structures of Pugh [31, 32].

SkipLists are search structures based on hierarchically ordered linked-lists, with a probabilistic guarantee of being balanced. The basic idea behind SkipLists is to keep elements in an ordered list, but have each record in the list be part of up to a logarithmic number of sub-lists. These sub-lists play the same role as the levels of a binary search structure, having twice the number of items as one goes down from one level to the next. To search a list of N items, O (log N) level lists are traversed, and a constant number of items is traversed per level, making the expected overall complexity of an Insert or Delete operation on a SkipList O(logN). Author introduced the SkipQueue, a highly distributed priority queue based on a simple modification of Pugh's concurrent SkipList algorithm [31]. Inserts in the SkipQueue proceed down the levels as in [31]. For Delete-min, multiple minimal" elements are to be handed out concurrently. This means that one must coordinate the requests, with minimal contention and bottlenecking, even though Delete-mins are interleaved with Insert operations. The solution was as follows, keep a specialized delete pointer which points to the current minimal item in this list. By following the pointer, each Delete-min operation directly traverses the lowest level list, until it finds an unmarked item, which it marks as \deleted." It then proceeds to perform a regular Delete operation by searching the SkipList for the items immediately preceding the item deleted at each level of the list and then redirecting their pointers in order to remove the deleted node.

Sundell et.al [16] given an efficient and practical lock-free implementation of a concurrent priority queue that is suitable for both fully concurrent (large multiprocessor) systems as well as pre-emptive (multi-process) systems. Inspired by [15], the algorithm was based on the randomized Skiplist [28] data structure, but in contrast to [15] it is lock-free. The algorithm was based on the sequential Skiplist data structure invented by Pugh [32]. This structure uses randomization and has a probabilistic time complexity of O(logN) where N is the maximum number of elements in the list. The data structure is basically an ordered list with randomly distributed short-cuts in order to improve search times, In order to make the Skiplist construction concurrent and non-blocking; author used three of the standard atomic synchronization primitives, Test-And-Set (TAS), Fetch-And-Add (FAA) and Compare-And-Swap (CAS). To insert or delete a node from the list we have to change the respective set of next pointers. These have to be changed consistently, but not necessary all at once. The solution was to have additional information on each node about its deletion (or insertion) status. This additional information will guide the concurrent processes that might traverse into one partial deleted or inserted node. When we have changed all necessary next pointers, the node is fully deleted or inserted.

- *Tree-Based Priority Pools:* Huang and Weihl et al. [18] and Johnson et al.[17] describe concurrent priority pools: priority queues with relaxed semantics that do not guarantee linearizability of the delete-min operations. Their designs were based on a modified concurrent B+-tree implementation. Johnson introduces a "delete bin" that accumulates values to be deleted and thus reduces the load when performing concurrent delete-min operations.

V. Comparison and Analysis

Data structure	Algorithm	Merits	Demerits
Stack	Systems programming: Coping with parallelism	Simple and can be expected to be quite efficient.	Contention and an inherent sequential bottleneck.
	A scalable lock-free stack algorithm	Due to elimination technique there is high degree of parallelism.	
queue	Implementing Lock-Free queues.	Algorithm no longer needs the snapshot, only intermediate state that the queue can be in is if the tail pointer has not been updated	Required either an unaligned compare & swap or a Motorola like double-compare and swap, both of them are not supported on any architecture.
	Simple, Fast, and Practical Non-Blocking and Blocking Concurrent Queue Algorithms.	The algorithm was simple, fast and practical .it was the clear algorithm of choice for machine that provides a universal atomic primitive.	Pointers are inserted using costly CAS
	Using elimination to implement scalable and lock-free FIFO queues.	1. Scaling technique allows multiple enqueue and dequeue operations to complete in parallel. 2. The concurrent access to the head and tail of the queue do not interfere with each other as long as the queue is non-empty.	1. The elimination back off queue is practical only for very short queues as in order to keep the correct FIFO queue semantics, the enqueue operation cannot be eliminated unless all previous inserted nodes have been dequeued. 2. scalable in performance as compare to previous one but having high overhead.
Tree	Concurrent manipulation of binary search trees.	Algorithm never blocked the search operations	Search tree is not balanced
	Efficient Locking for Concurrent Operations on B-trees,	Small number of locks used	Expansive locks
	A symmetric concurrent b-tree algorithm	They involved the merging as a part of deletion.	Expansive locking

Priority queue	An efficient algorithm for concurrent priority queue heaps	Allows concurrent insertion and deletion in opposite direction.	The performance does not scale beyond a few tens of concurrent processors.
	Skip list-Based Concurrent Priority Queues	Designed a scalable concurrent priority queue for large scale multi-processor.	Algorithm based on locking approach.
	Fast and Lock-Free Concurrent Priority Queues for Multithread System.	This was a first lock-free approach for concurrent priority queue	
	A highly concurrent priority queue based on the b-link tree	Avoid the serialization bottleneck	Needs node to be locked in order to be rebalance
Linked list	Lock-free linked lists using compare-and-swap	Reduced interference of concurrent operations using backlink nodes	
	A pragmatic implementation of non-blocking linked-lists	For making successful updating of nodes, every node to be deleted was marked	Difficult to implement
	High performance dynamic lock-free hash tables and list-based sets.	Efficient with memory management techniques	Poor in performance.

VI. Conclusion

This paper reviews the different data structures and the concurrency control techniques with respect to different data structures (tree, queue, priority queue). The algorithms are categorized on the concurrency control techniques like blocking and non-blocking. Former based on locks and later one can be lock-free, wait-free or obstruction free. In the last we can see that lock free approach outperforms over locking based approach.

References Références Referencias

1. R. K. Treiber, "Systems programming: Coping with parallelism", "RJ 5118, Almaden Research Center, and ''April 1986.
2. M. Michael and M. Scott. "Nonblocking algorithms and preemption-safe locking on multiprogrammed shared - memory multiprocessors." Journal of Parallel and Distributed Computing, 51(1): 1–26, 1998.
3. D. Hendler, N. Shavit, and L. Yerushalmi. "A scalable lock-free stack algorithm." Technical Report TR-2004-128, Sun Microsystems Laboratories, 2004.
4. J. D. Valois. "Implementing Lock-Free queues." In Seventh International Conference on Parallel and Distributed Computing Systems, Las Vegas, NV, October 1994.
5. M. M. Michael and M. L. Scott. "Simple, Fast, and Practical Non-Blocking and Blocking Concurrent Queue Algorithms." 15th ACM Symp. On Principles of Distributed Computing (PODC), May 1996. pp.267 – 275.
6. Mark Moir, Daniel Nussbaum, Ori Shalev, Nir Shavit: "Using elimination to implement scalable and lock-free FIFO queues. " SPAA 2005.
7. H. Kung and P. Lehman. "Concurrent manipulation of binary search trees." ACM Transactions on Programming Languages and Systems, 5:354–382, September 1980.
8. P. Lehman and S. Yao. "Efficient Locking for Concurrent Operations on B-trees", ACM Trans. Database Systems, vol. 6,no. 4, 1981.
9. Y. Sagiv. "Concurrent operations on b-trees with overtaking." Journal of Computer and System Sciences, 33(2):275–296, October 1986.
10. V. Lanin and D. Shasha. "A symmetric concurrent b-tree algorithm." In Proceedings of the Fall Joint Computer Conference 1986, pages 380–389. IEEE Computer Society Press, November 1986.
11. J. Biswas and J. Browne. "Simultaneous update of priority structures." In Proceedings of the 1987 International Conference on Parallel Processing, pages 124–131, August 1987.
12. V. Rao and V. Kumar. "Concurrent access of priority queues." IEEE Transactions on Computers, 37:1657–1665, December 1988.
13. R. Ayani. "LR-algorithm: concurrent operations on priority queues." In Proceedings of the 2nd IEEE Symposium on Parallel and Distributed Processing, pages 22–25, 1991.
14. G. Hunt, M. Michael, S. Parthasarathy, and M. Scott. "An efficient algorithm for concurrent priority queue heaps." Information Processing Letters, 60(3): 151–157, November 1996.
15. LOTAN, N. SHAVIT. "Skiplist-Based Concurrent Priority Queues", International Parallel and Distributed Processing Symposium, 2000.

16. H.Sundell and P.tsigas. "Fast and Lock-Free Concurrent Priority Queues for Multithread System."
17. T. Johnson. "A highly concurrent priority queue based on the b-link tree." Technical Report 91-007, University of Florida, August 1991.
18. Q. Huang and W. Weihl". An evaluation of concurrent priority queue algorithms. "In IEEE Parallel and Distributed Computing Systems, pages 518–525, 1991.
19. J. D. Valois. "Lock-free linked lists using compare-and-swap." In Proceedings of the 14th ACM Symposium on Principles of Distributed Computing, pages 214–222, 1995.
20. T. L. Harris. "A pragmatic implementation of non-blocking linked-lists." In Proceedings of the 15th International Symposium on Distributed Computing, pages 300–314, 2001.
21. M. M. Michael. "High performance dynamic lock-free hash tables and list-based sets." In Proceedings of the 14th annual ACM Symposium on Parallel Algorithms and Architectures, pages 73–82, 2002.
22. M. Greenwald. "Non-Blocking Synchronization and System Design." PhD thesis, Stanford University Technical Report STAN-CS-TR-99-1624, Palo Alto, A, 8 1999.
23. N. Shavit and D. Touitou. "Elimination trees and the construction of pools and stacks." Theory of Computing Systems, 30:645–670, 1997.
24. M. Herlihy. "A methodology for implementing highly concurrent data objects." ACM Transactions on Programming Languages and Systems, 15(5): 745–770, 1993.
25. M. Herlihy and J. Wing. "Linearizability: a Correctness Condition for Concurrent Objects." ACM Transactions on Programming Languages and Systems, 12(3): 463–492, 1990.
26. H. Sundell and P. Tsigas. NOBLE: A Non-Blocking Inter-Process Communication Library. In Proceedings of the 6th Workshop on Languages, Compilers and Run-time Systems for Scalable Computers, 2002.
27. P. Tsigas and Y. Zhang. Integrating Non-blocking Synchronization in Parallel Applications: Performance Advantages and Methodologies. In Proceedings of the 3rd ACM Workshop on Software and Performance, pages 55–67. ACM Press, 2002.
28. M. Herlihy. "A methodology for implementing highly concurrent data objects." ACM Transactions on Programming Languages and Systems, 15(5): 745–770, November 1993.
29. J. Mellor-Crummey and M. Scott. "Algorithms for scalable synchronization on shared memory multiprocessors." ACM Transactions on Computer Systems, 9(1):21–65, 1991.
30. J. Turek, D. Shasha, and S. Prakash. "Locking without Blocking: Making Lock Based concurrent Data Structure Algorithms Nonblocking." In Proceedings of the 11th ACM SIGACT-SIGMOD-SIGARTSymposium on Principles of Database Systems, pages 212–222, 1992
31. W. Pugh. "Concurrent Maintenance of Skip Lists." Technical Report, Institute for Advanced Computer Studies, Department of Computer Science, University of Maryland, College Park, CS-TR-2222.1, 1989.
32. W. Pugh. Skip Lists: "A Probabilistic Alternative to Balanced Trees." In Communications of the ACM, 33(6):668{676, June 1990.

This page is intentionally left blank

Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 14 Issue 3 Version 1.0 Year 2014
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Mobilemedia SPL Creation by FeatureIDE using FODA

By Manjinder Kaur & Parveen Kumar

Lovely Professional University, India

Abstract- Software Product Lines are used in many areas, combining to form new technologies and products. A product line is a group of products that share a common development platform and vary by the composition and implementation method for the functionalities. This paper describes the implementation or creation of MobileMedia feature model using FODA (Feature Oriented Domain Analysis) methodology using FeatureIDE eclipse plug-in. The feature model created in this depicts various outlines as feature model as visual model, collaboration diagram view of model, its configuration, FeatureIDE Statistics. Basically the paper shows the concept how SPLs can be viewed as feature diagrams using various tools in order to deal with them. This modelling has been widely used by software product line communities and a number of extensions have been proposed.

Indexterms: featureide, feature model, mobilemedia, SPLS.

GJCST-C Classification : D.2

Mobilemedia SPL Creation by FeatureIDE using FODA

Manjinder Kaur [α] & Parveen Kumar [σ]

Abstract- Software Product Lines are used in many areas, combining to form new technologies and products. A product line is a group of products that share a common development platform and vary by the composition and implementation method for the functionalities. This paper describes the implementation or creation of MobileMedia feature model using FODA (Feature Oriented Domain Analysis) methodology using FeatureIDE eclipse plug-in. The feature model created in this depicts various outlines as feature model as visual model, collaboration diagram view of model, its configuration, FeatureIDE Statistics. Basically the paper shows the concept how SPLs can be viewed as feature diagrams using various tools in order to deal with them. This modelling has been widely used by software product line communities and a number of extensions have been proposed.

Indexterms: featureide, feature model, mobilemedia, SPLS.

I. Introduction

With the advancement in the software engineering, many new concepts have been introduced and changing drastically. Software Product Lines (SPLs), an innovative approach in the Software Engineering has changed many things in the industrial area. A product derived from a software product line consists of various components selected from existing component libraries; these components communicate with a common platform to perform specific functionalities.

Skilled software engineers use technologies and practices from a variety of fields to improve their productivity in creating software and to improve the quality of the delivered product called as Product Line Engineering (PLE) [6]. SPLs can be viewed as models by using the concept of feature modelling involving various methodologies. In this paper, FODA (Feature Oriented Domain Analysis) methodologies in designing the SPL using eclipse plug-in FeatureIDE.

II. Software Product Lines (SPLS)

A Software Product Lines are defined as a family of different products which shares same set of core assets or it can be said, a product line consists of multiple systems, which have same architecture and share common core assets with variability among systems. A core asset includes shared components, framework or infrastructure, tools, process, documentation, test cases as these are reused.

Basically, SPL is a family of products designed to take advantage of their common aspects and predicted variability in order to improve quality, delivery time and reduction in cost. Product line engineering (PLE) helps to design, develop, deliver, and evolve a portfolio of common products, with feature variations and functions, through which each stage of the systems and the software development lifecycle from requirements to design, development and testing. Many methods and practices are introduced that is Software Reuse, Component-Based Development and Product Line Engineering (or Product Family Engineering). Enhance the efficiency of SW development when multiple products are to be developed simultaneously:

- Higher productivity
- Higher quality
- Faster time to market
- Lower labour needs

Software Product Line Methods (SPLMs) are the software development approaches in which a set of software systems share a common set of feature produced from a set of reused core assets. Core assets are software artifacts that are re-used in the production of customized products in a software product line (SPL). The assets include the requirements, architecture, components, modelling and analysis, plans, etc. A SPL product can be quickly assembled from core assets, and hence it achieves manufacturing efficiency. SPLM supports "producing goods and services to meet individual customer's needs with near mass production efficiency". The following figure depicts the overview of Software Product Line Methods:

Author α: Student, Department of Computer Science & Engineering, Lovely Professional University, Phagwara, Punjab, India.
e-mail: maenj_chauhan@hotmail.com
Author σ: Assistant Professor, Department of Computer Science & Engineering, Lovely Professional University, Phagwara, Punjab, India.
e-mail: parveen.it@gmail.com

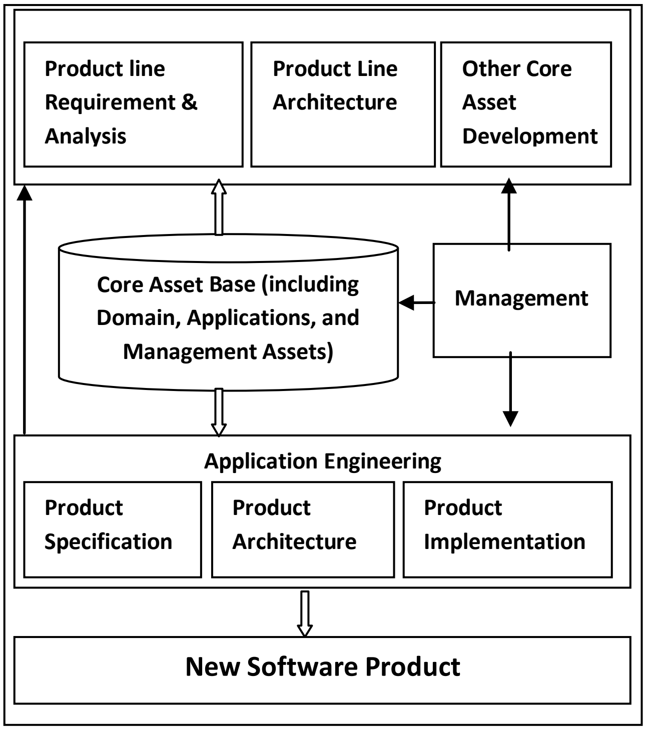

Product line Requirement & Analysis
Product Line Architecture
Other Core Asset Development
Core Asset Base (including Domain, Applications, and Management Assets)
Management
Application Engineering
Product Specification
Product Architecture
Product Implementation
New Software Product

Figure 1: Overview of SPL Methods

III. Feature Model by foda Methodology

Feature model, a representation of products of SPL in a way to express features. Feature models were first introduced in FODA (Feature-Oriented Domain Analysis) method by Kang in 1990 and since then this modeling has been widely used by software product line communities and a number of extensions have been proposed. A feature model is represented by means of feature diagrams. A feature diagram is a graphical or visual notation of a feature model in the form of AND-OR tree, and also various other extensions as feature cloning, feature attributes, collaboration diagrams, configurations etc. This model is basically used as a input to produce other different assets like documents, architecture definition or code. A feature can be defines as a quality or characteristic of software system or system. Therefore, a feature model is a model that defines features and their dependencies depicting in the feature model as cross-tree constraints. A feature configuration, a set of feature describing a member of an SPL and the member will contain a feature if and only if features are in its configuration.

IV. Basic Feature Model Notations

Basic Feature Model is a relationship between a parent feature and its child features categorized as [1]:

- Mandatory – child feature is required.
- Optional – child feature is optional.
- Or – at least one of the sub-features must be selected.
- Alternative (xor) – one of the sub-features must be selected

In addition to the parental relationships between features, cross-tree constraints are allowed. The most common are:

- A requires B – The selection of A in a product implies the selection of B.
- A excludes B – A and B cannot be part of the same product.

V. Foda Methodology

As said earlier, Feature models were first introduced in FODA (Feature-Oriented Domain Analysis) method developed by Software Engineering Institute. It is a domain analysis or product analysis method (analyzing related software systems in a domain to find commonality and variations), which involves conversion as feature model to domain engineering being used in the advance concepts in software engineering and software reuse. Domain Analysis was coined by James Neighbors in 1980s and is the first phase of domain engineering. Therefore it can be said that FODA is a domain analysis technique. The main objective of feature-oriented domain analysis if to create a domain model which represents the family of systems which is then refined into a particular desired system within a domain supporting the functional and architectural reuses [4]. FODA methodology not only identifies the systems in the domain but also the external system interacting with the domain which is known as FODA context analysis. Then further FODA feature analysis from the feature model, configures requirements and the candidate systems by analyzing the end-user's view. Configurable requirements can then be selected from the developed feature model in order to specify the final system and from this view the customer's demands can be satisfied by following this process and achieving the efficiency through technology reuse.

VI. SPL Tools Supporting Feature Models [2]

- FeatureIDE
- Clafer
- Pure::Variants
- Hydra
- S. P. L. O. T. (Software Product Line Online Tools)
- Ahead Tool suite
- Eclipse Modelling Framework Feature Model Project
- FaMa Tool suite
- ToolDay (Tool for Domain Analysis)
- BeTTy Framework
- FAMILIAR

- Feature Model Plug-in
- Gears
- Dopler
- Varmod
- PULSE-BEAT
- FeatureMapper
- MetaEdit+
- Holmes
- FAMA-FW
- MTP (Meta Programming Text processor)
- XVCL
- SPL conqueror
- Velvet
- The guidsl Tool

VII. FeatureIDE Tool for MobileMedia SPL Creation

FeatureIDE is an Eclipse plug-in or Eclipse based IDE supporting Feature-Oriented Software Development, development of SPLs which involves Domain Analysis, Software Generation and Requirement Analysis. It is the integration of various SPL techniques like Feature-Oriented programming (FOP) (integration of AHEAD, FeatureC++, FeatureHouse), Delta programmming (DOP) including DeltaJ, Aspect-oriented Programming (AOP) including AspectJ and preprocessors (includes annotations of Colligens, Munge, Antenna and Type Chef) [1]. This tool is still under development that is involving new implementations in it. Feature IDE provides a way to implement or create following feature:

- A feature model editor, graphical and text-based.
- Constraint editor with syntax and semantic checking like dead feature detection.
- Configuration Editor for creation and editing of features.
- SPL source code abstraction.
- Supports refactoring, generalizations etc.
- Statistics display of FeatureIDE project.
- Outline view of feature model.
- Collaboration Diagram view.
- Debugging and Lay outing the feature models or manyother new implementations are under development.

VIII. MobileMedia Feature Model

MobileMedia manipulates media on mobile phones that is music, photo and video. It is basically a family of multimedia management application for mobile devices and is used in research community of SPLs. In this paper, MobileMedia uses 31 features that are Mobile Selection, Mobile Management, Gallery, Internet and Games, which further involve more features as shown in the figure 2. This representation in the figure is called a feature model diagram. A node represents a feature and edges represent dependency between the features. Feature-model diagram also represents constraints on the selected feature when product is build. A white circle (hollow) indicates the optional feature and the black circle (solid) indicates the mandatory feature.

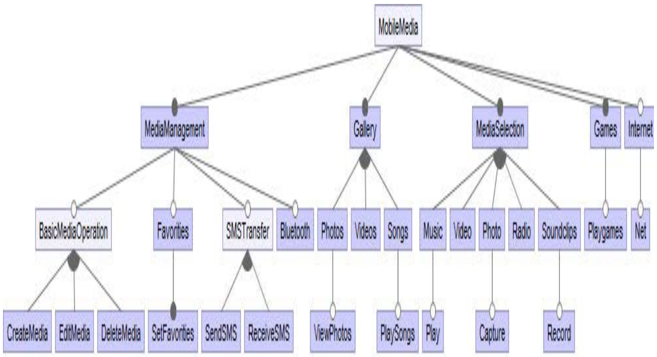


Figure 2 : A Feature Model of Mobile media

In the figure, or-relation has also been used as an arc among the children's from which at least one features from them is selected if their parent is selected. In Feature-Oriented Programming, a feature is implemented as an independent feature module. When a feature is selected, then the corresponding module is compiled together with the other feature modules which have been selected. A group of optional features may interact with another group also mandatory one does can also interact. This feature model involves Abstract and concrete features involving various mandatory and optional feature with AND, OR features. Mobile Media, Basic Media Operation and SMS Transfer are the abstract features and remaining the concrete features and terminal features.

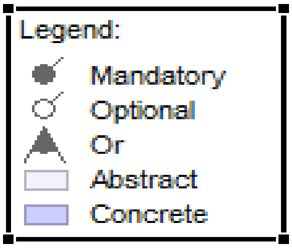
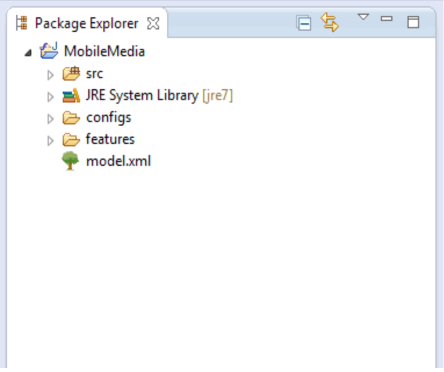


Figure 3 : Legend

Figure 3 displays the Legend shown in the Feature IDE when a feature model diagram is implemented and it shows side by side the symbols with the feature relation being used in the diagram.

a) Creation of Feature Model

First of all a feature project is created by selecting Feature IDE Project and choosing the composer as AHEAD [7]. Then project name is mentioned. When Feature IDE project will be created it will be displayed on the left side pane of window in the Package Explorer like shown in the figure and model with root and base will be displayed in the Feature Diagram IDE.

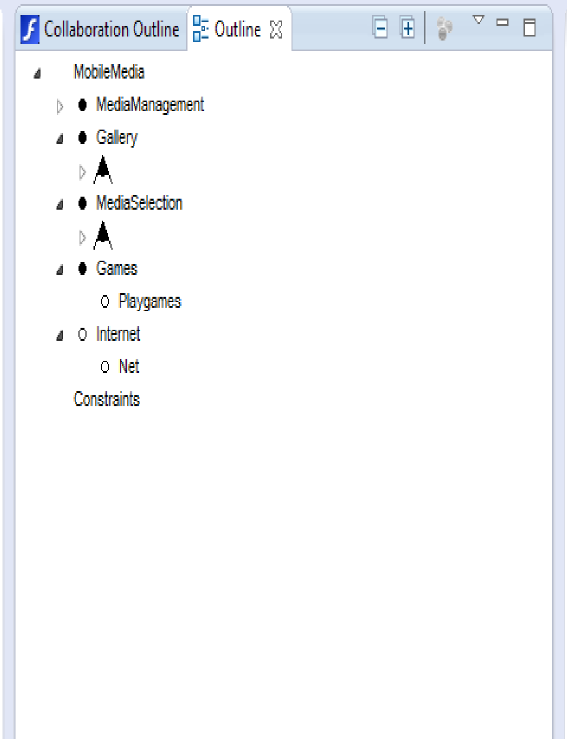


```
◢ MobileMedia
   ▷ src
   ▷ JRE System Library [jre7]
   ▷ configs
   ▷ features
     model.xml
```

Figure 4 : MobileMedia FeatureIDE Project

b) Outline View of the Feature Model

Outline view displays the overall outline of the feature model that in a brief what it involves like what are the mandatory features, optional features and constraints.



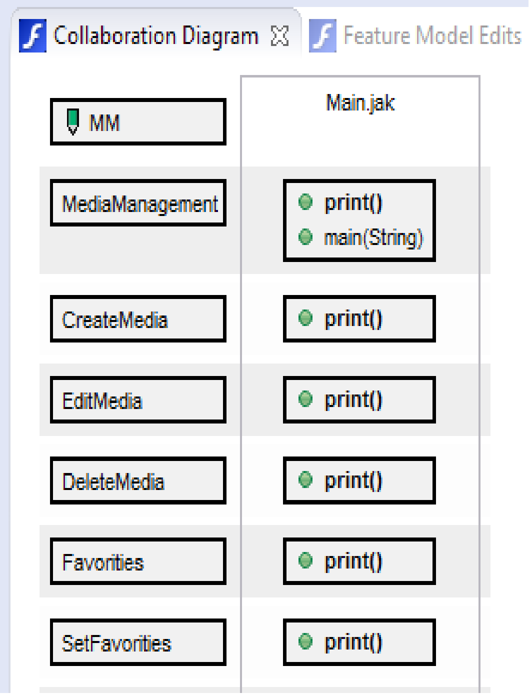
```
◢   MobileMedia
  ▷ ● MediaManagement
  ◢ ● Gallery
      ▷ A
  ◢ ● MediaSelection
      ▷ A
  ◢ ● Games
        ○ Playgames
  ◢ ○ Internet
        ○ Net
    Constraints
```

Figure 5 : Outline View

Solid circles represent the mandatory features and the hollow circles represent the optional features and another symbol represents the or-relation between the features involving it.

c) Collaboration Diagram of Featureide Project

Collaboration Diagram shows the diagram of the feature model as what is the main class being involved and in what feature it is present and shows what the other class features that refines the main feature.

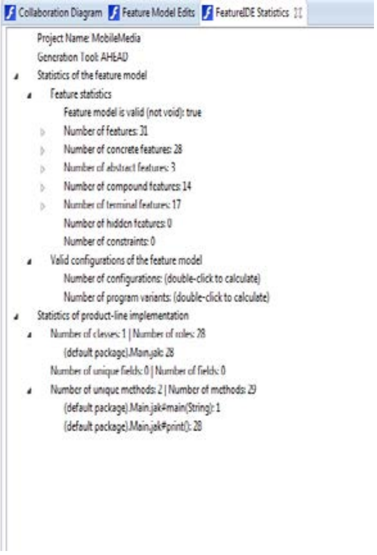


MM	Main.jak
MediaManagement	● print() ● main(String)
CreateMedia	● print()
EditMedia	● print()
DeleteMedia	● print()
Favorities	● print()
SetFavorities	● print()

Figure 6 : Collaboration Diagram

Basically it shows the Main Jak file and the functions involved in the other class features as print ( ) method.

d) Featureide Statistics of the Project



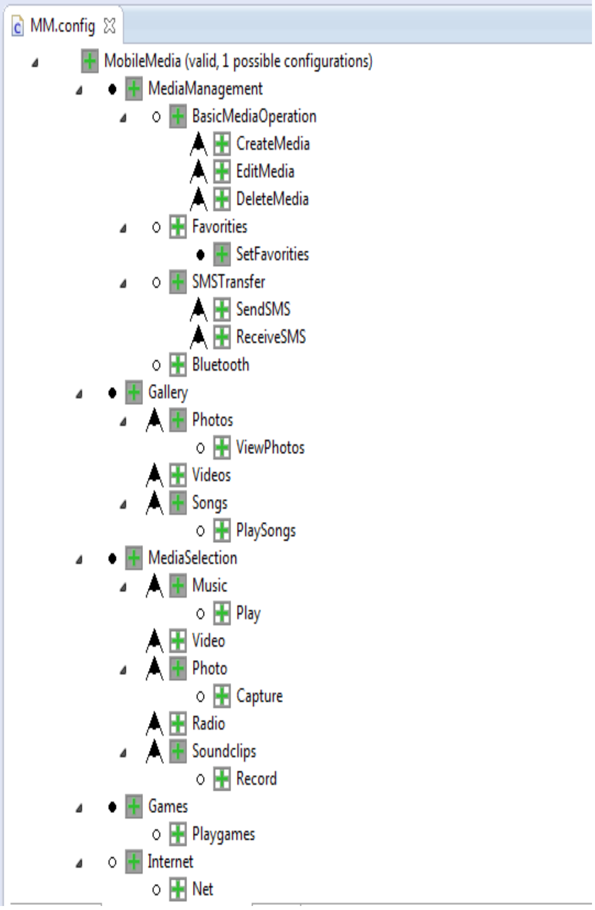
```
    Project Name: MobileMedia
    Generation Tool: AHEAD
◢   Statistics of the feature model
  ◢   Feature statistics
        Feature model is valid (not void): true
    ▷   Number of features: 31
    ▷   Number of concrete features: 28
    ▷   Number of abstract features: 3
    ▷   Number of compound features: 14
    ▷   Number of terminal features: 17
        Number of hidden features: 0
        Number of constraints: 0
  ◢   Valid configurations of the feature model
        Number of configurations: (double-click to calculate)
        Number of program variants: (double-click to calculate)
◢   Statistics of product-line implementation
  ◢   Number of classes: 1 | Number of roles: 28
        (default package).Main.jak: 28
      Number of unique fields: 0 | Number of fields: 0
  ◢   Number of unique methods: 2 | Number of methods: 29
        (default package).Main.jak#main(String): 1
        (default package).Main.jak#print(): 28
```

Figure 7 : Feature IDE Statistics

This statistics depicts the composer or generation tool AHEAD (Algebraic Hierarchical Equations for Application Design) being used and the statistics involved under Feature IDE. AHEAD composer supports the composition of Jak files where Jak extends the Java with Keywords for Feature-Oriented Programming. In the Mobile Media SPL, feature statistics explain the number of Abstract feature, Concrete features, Compound features, Terminal features, Hidden features and the Constraints involved in it. Other statistics involves the product-line implementations which include the unique methods, classes and unique fields.

e) Configuration File Creation

Configuration view displays the number of configurations the user wants to involve in its SPL feature model. It may the main single configuration involving many configurations.

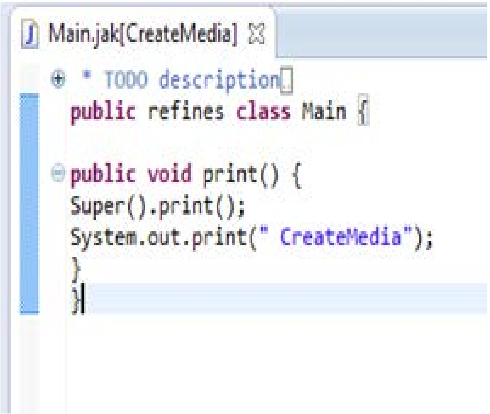


```
MM.config
MobileMedia (valid, 1 possible configurations)
  ● MediaManagement
    ○ BasicMediaOperation
        CreateMedia
        EditMedia
        DeleteMedia
    ○ Favorities
      ● SetFavorities
    ○ SMSTransfer
        SendSMS
        ReceiveSMS
    ○ Bluetooth
  ● Gallery
      Photos
      ○ ViewPhotos
      Videos
      Songs
      ○ PlaySongs
  ● MediaSelection
      Music
      ○ Play
      Video
      Photo
      ○ Capture
      Radio
      Soundclips
      ○ Record
  ● Games
    ○ Playgames
  ○ Internet
    ○ Net
```

Figure 8 : Configuration File

Here Mobile Media involves the main configuration and further its configuration that is feature under it which are refined by the main class involved in the Mobile Media. Refine is the keyword used in AHEAD composer of Feature IDE is used to specify the refinements of an existing class. In the following it has been shown how a class refines the main class.

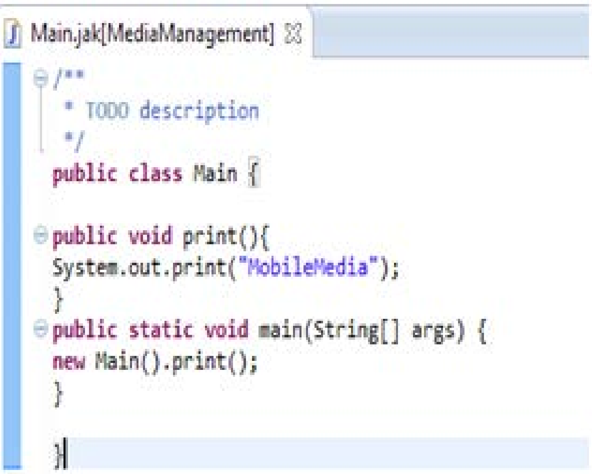
The following figure displays the Jak file creation of a single feature which uses the refine to use the main class.



```
Main.jak[CreateMedia]
 * TODO description
public refines class Main {

public void print() {
Super().print();
System.out.print(" CreateMedia");
}
}
```

Figure 9 : Source Code Showing a how a Class Main Jak is Refined by another Lower Level Class.

The main class shows the source like this file when created for the first time.

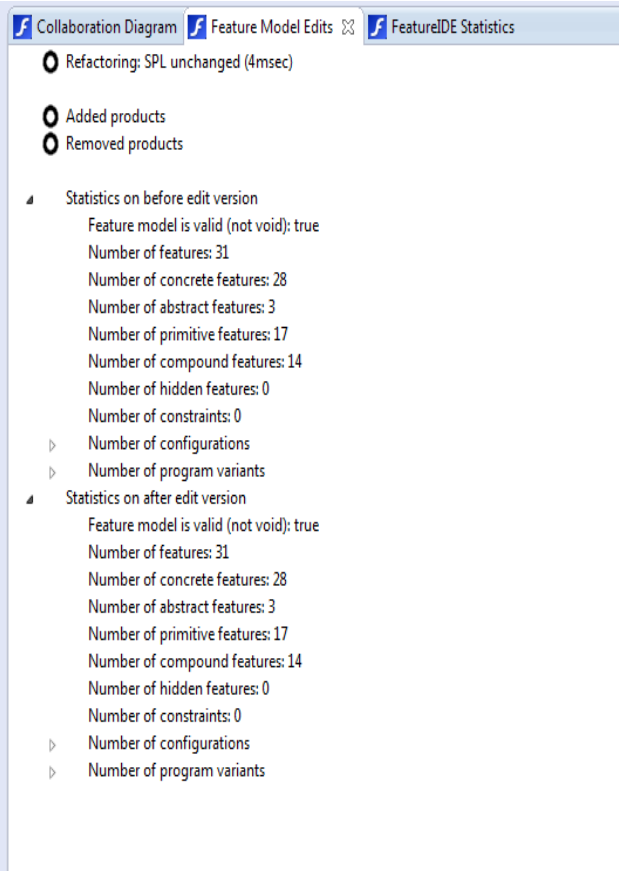


```
Main.jak[MediaManagement]
/**
 * TODO description
 */
public class Main {

public void print(){
System.out.print("MobileMedia");
}
public static void main(String[] args) {
new Main().print();
}

}
```

Figure 10 : Code Showing how a Main Jak File is Implemented.

f) Feature Model Edits

Feature Model Edit pane shows how much editing has been done to the features and the products involved in the feature model. The number of added and removed products is represented in this edit, the overall how much refactoring has been done to the SPL. In this figure, refactoring has been done to the Mobile Media SPL that before editing how many features were involved in the model and after refactoring how much features has been altered in the model.



```
Refactoring: SPL unchanged (4msec)

Added products
Removed products

Statistics on before edit version
    Feature model is valid (not void): true
    Number of features: 31
    Number of concrete features: 28
    Number of abstract features: 3
    Number of primitive features: 17
    Number of compound features: 14
    Number of hidden features: 0
    Number of constraints: 0
    Number of configurations
    Number of program variants
Statistics on after edit version
    Feature model is valid (not void): true
    Number of features: 31
    Number of concrete features: 28
    Number of abstract features: 3
    Number of primitive features: 17
    Number of compound features: 14
    Number of hidden features: 0
    Number of constraints: 0
    Number of configurations
    Number of program variants
```

Figure 11 : Feature Model Edit

IX. Conclusion

In this paper, SPLs has been discussed as how they can be presented as graphically. Number of tools has been introduced in the market to work upon it as PLE is the new trend in the software engineering. “Engineering” in product lines means the activities are taken into account involved in planning, producing, delivering, and deploying, and retiring products etc. SPLs are under research in various areas. A feature model of Mobile Media that is SPL has been proposed in this paper which is providing a way to represent software product lines in graphical manner in the Feature IDE tool as eclipse plug-in using the FODA methodology. Also defining how it works and shows various outputs of Mobile Media SPL. The feature model created in this depicts various outlines as feature model as visual model, collaboration diagram view of model, its configuration, Feature IDE Statistics and also how a main jak file is created and other files refine the main jak file.

References Références Referencias

1. Takeyama1 and Shigeru Chiba23 “Implementing Feature Interactions with Generic Feature Modules”.
2. Qaiser, Shahid Muhammad et. al., (2010) “Software Product Line: Survey of Tools”, LiU Electronic Press.
3. Bayer Joachin et. al. (1999) “PuLSE: A Methodology to Develop Software Product Lines”, ACM Press.
4. Klaus Pohl, Günter Böckle, Frank vander Linden, (2005) Software Product Line Engineering, Foundations, Principles, and Techniques, Springer.
5. Mathieu Acher, Roberto E., Rick Rabiser, “A Survey on Teaching of Software Product Lines”, Author manuscript, published in "Eight International Workshop on Variability Modelling of Software-Intensive`Systems (VaMoS'14) (2014)".
6. http://www.productlineengineering.com.
7. http://wwwiti.cs.uni-magdeburg.de/iti_db/research/featureide/.

Global Journals Inc.

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 14 Issue 3 Version 1.0 Year 2014
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

A Fuzzy Rule based Approach to Predict Risk Level of Heart Disease

By Kantesh Kumar Oad, Xu DeZhi & Pinial Khan Butt

Central South University, China

Abstract - Health care domain systems globally face lots of difficulties because of the high amount of risk factors of heart diseases in peoples (WHO, 2013). To reduce risk, improved knowledge based expert systems played an important role and has a contribution towards the development of the healthcare system for cardiovascular disease. To make use of benefits of knowledge based system, it is necessary for health organizations and users; must need to know the fuzzy rule based expert system's integrity, efficiency, and deployments, which are the open challenges of current fuzzy logic based medical systems. In our proposed system, we have designed a fuzzy rule based expert system and also by using data mining technique we have reduced the total number of attributes. Our system mainly focuses on cardiovascular disease diagnosis, and the dataset taken from UCI (Machine Learning Repository). We explored in the existing work. The majority of the researcher's experimentation was made on 14 attributes out of 76. While, in our system we took advantage of 6 attributes for system design. In the preliminary stage UCI, data participated in suggested system that will get outcomes. The performance of the system matched with Neural Network and J48 Decision Tree Algorithm.

Keywords: fuzzy reasoning, heart disease and diagnose, data mining.

GJCST-C Classification : I.5.1

A Fuzzy Rule based Approach to Predict Risk Level of Heart Disease

Kantesh Kumar Oad [α], Xu DeZhi [σ] & Pinial Khan Butt [ρ]

Abstract- Health care domain systems globally face lots of difficulties because of the high amount of risk factors of heart diseases in peoples (WHO, 2013). To reduce risk, improved knowledge based expert systems played an important role and has a contribution towards the development of the healthcare system for cardiovascular disease. To make use of benefits of knowledge based system, it is necessary for health organizations and users; must need to know the fuzzy rule based expert system's integrity, efficiency, and deployments, which are the open challenges of current fuzzy logic based medical systems. In our proposed system, we have designed a fuzzy rule based expert system and also by using data mining technique we have reduced the total number of attributes. Our system mainly focuses on cardiovascular disease diagnosis, and the dataset taken from UCI (Machine Learning Repository). We explored in the existing work. The majority of the researcher's experimentation was made on 14 attributes out of 76. While, in our system we took advantage of 6 attributes for system design. In the preliminary stage UCI, data participated in suggested system that will get outcomes. The performance of the system matched with Neural Network and J48 Decision Tree Algorithm.

Keywords: fuzzy reasoning, heart disease and diagnose, data mining.

I. Introduction

Just recently peoples are stressed over their health and wellness troubles, in the majority of the countries proportion of cardiovascular disease enhancing so quickly and it has become the leading cause / death worldwide [1][2], and it is came to be taken into consideration as a "second epidemic," changing transmittable disease[3][4]. Health domain application is one of one of the most active study area nowadays. Ideal example of health domain application is the detection system for cardiovascular disease based on computer system assisted diagnosis strategies, where the information acquired from numerous other sources and is evaluated based on computer-based application. Before it was very time consuming job to get knowledge from physician and include this knowledge to computer system program by hand into data base medical decision support system and this was totally depending upon clinical experts' concepts which may be subjective.

This trouble has really been resolved using expert systems; get physician, group, knowledge and certain human client details, intelligently. In boosting outcomes at a couple of healthcare companies and strategy internet sites, expert system has actually operated by making needed clinical knowledge quickly readily available to know-how users [5] Taking care of clinical needs, such as making certain specific medical diagnoses, evaluating in a quick manner for avoidable health problem, or avoiding undesirable drug occasions, are the most standard exploitation of Expert System [6]. Expert System could also be possibly lessened costs, progression performance, and reduce client stress. These systems are classified into 2 groups namely (1) Knowledge based and (2) non-knowledge based [7]. The knowledge based system consists of rules (if-then statements). Expert system that is implemented with the assistance of artificial intelligence has the ability to support in a new setting and to learn for instance [8][9]. Given that the concept of computer-based Clinical Decision Support System aroused at first, significant research has actually been made in both academic and practical areas. Many obstacles are longer to impede the effective application of expert systems in scientific environments, among which portrayal and reasoning concerning clinical understanding predominantly under anxiety is the locations that require improved methodologies and strategies [10][11].

In our proposed system, mainly focus on cardiovascular disease diagnosis. We have taken dataset taken from UCI (Machine Learning Repository). UCI database consists of 76 attributes; we investigated in the existing work. Most of the experiments were made by using a subset of 14 from UCI. While, in our system we reduced the number of input attributes that will reduce the number of diagnostic results and we used seven attributes to experiment. From seven attributes, we used six attributes as input, and one attribute for output. Numerical data will enter in into suggested system, and in the last; system will get prediction results. The primary objective of our research is to make and carry out fuzzy rule based system for heart disease people.

Author α: Masters of Engineering in Computer Application Technology. School of Information Science and Engineering, Central South University, Changsha Hunan, China.
e-mail: kanteshoad84@gmail.com
Author σ: Professor, School of Information Science and Engineering, China. e-mail: hunan.xu@mail.csu.edu.cn
Author ρ: Ph.D. (Computer Science) Research Scholar, China.
e-mail: pinial@yahoo.com

II. Related Work

Mehdi.N. Mehdi. Y. (2009) [12] designed a Fuzzy Expert System of diagnosing the hepatitis B intensity rate and making comparisons with Adaptive Neural. M. Neshat, M. Yaghobi, M.B. Naghibi, A. Esmaelzadeh (2008) [13] designed a fuzzy expert system for Diagnosis of liver disorders. P.K. Anooj (2012) [14] developed Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. Persi Pamela I, Gayathri. P, N. Jaisankar (2013) [15] uses a Fuzzy Optimization Technique for the Prediction of Coronary Heart Disease Using Decision Tree. Nidhi Bhatla, Kiran Jyoti (2012) [16] used Novel Approach for Heart Disease Diagnosis using Data Mining and Fuzzy Logic. Vijay Kumar Magoa, Nitin Bhatia, Ajay Bhatia, Anjali Mago (2013) [17] designed clinical decision support system for dental treatment. Syed Umar Amin, Kavita Agarwal, Dr. Rizwan Beg (2013) [18] used Data Mining in Clinical Decision Support Systems for Diagnosis, Prediction and Treatment of Heart Disease. Asha Rajkumar, Mrs. G. Sophia Reena (2010) [19] used diagnosis of Heart Disease according to the data mining Algorithm "GJCST Classification J.3. M. Anabarsi. Anupriya *, N. CH. Iyengar (2010) [20] enhanced prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm. K. Rajeswari. V. Vaithiyanathan (2011) [21] designed Heart disease diagnosis: an efficient decision support system based on fuzzy logic and genetic algorithm.

III. A Fuzzy Rule Based Approach to Predict Risk Level of Heart Disease

As you can see in figure (1) the process begins with data processing. In a second step, we reduced the number of attributes and then this processed data (symptoms) inserted into fuzzy system using MATLAB programming. After the fuzzy model is successfully developed, the prediction of the symptoms will start and lastly performance based on the result will be analyzed at the end of the development phase.

a) Data Processing

The function of data processing is to draw out the significant information from the raw data collection useful for heart disease prediction, and these data sets should be transformed into the needed style for the level of risk prediction. Due to large amount of data there are chances of errors on it so before processing heart disease datasets, the original raw datasets must be processed;as a result in data processing phase, we cleaned, transformed and analyzed it into the row column format after taking out the unnecessary ones.

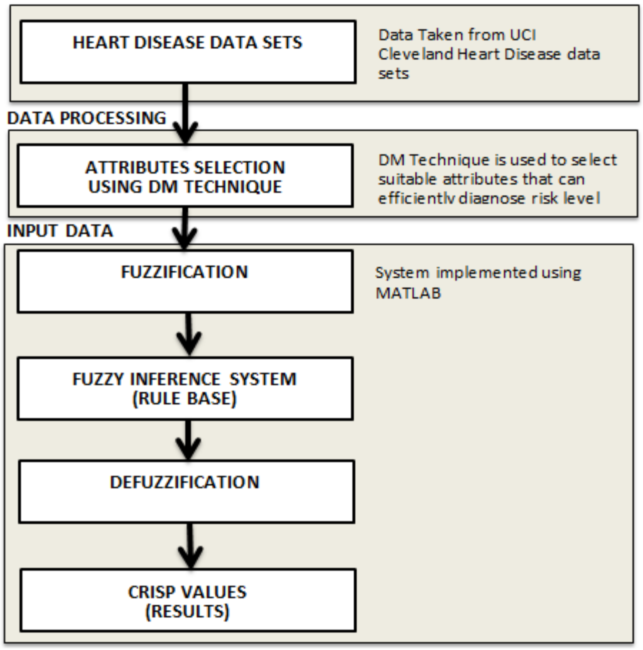

HEART DISEASE DATA SETS
Data Taken from UCI Cleveland Heart Disease data sets
DATA PROCESSING
ATTRIBUTES SELECTION USING DM TECHNIQUE
DM Technique is used to select suitable attributes that can efficiently diagnose risk level
INPUT DATA
FUZZIFICATION
System implemented using MATLAB
FUZZY INFERENCE SYSTEM (RULE BASE)
DEFUZZIFICATION
CRISP VALUES (RESULTS)

Figure 1 : Process Flow in Constructing Heart Disease Diagnose System

b) Data Formats

For recap or view purpose we should make data in a format so it can be easily reviewed. We have converted data to a .csv (Comma Separated Values) documents format also called comma delimited file. .csv file is a particularly configured text file; which establishments spreadsheet or standard database-style information in a quite easy style, with one record for each line, and each field within that document separated by a comma.

c) Attributes Selection

To minimize disease data sets (feature selection), we have used data reduction technique. Attribute selection lessens the data set size by getting rid of redundant or unnecessary and extracting with a minimized set of characteristics has an extra benefit [22]. Primarily we have actually made use of the data mining tool to get decreased collection of datasets. An attributes selection method contains four steps, named (1) subset generation; (2) subset evolution, (3) stop criterion (4) outcome validation [23] subset generation is a searching technique and we have made use of Best First Search Method using DM tool. Each new subset evaluated and matched with earlier best one according to a certain evolution criterion. It changes the previous finest subset if the new subset turns out to be much better. The process of subset generation and evolution is repeated till a given criterion is pleased.

By using selected method, we have obtained 6 attributes from a total of 14 attributes.

d) System Phases

Fuzzy system used in circumstances where we have a trouble of uncertainty. When the trouble has a dynamic behavior; Fuzzy Rule Base is an appropriate system for handling this issue. In the primary step of fuzzy expert system design; figure out the input and result variables. As it is described in section (4.3) we selected six inputs and one output variable. Then, we have actually made the membership functions for all variables used in the system. Membership function is primarily a visual portrayal of a fuzzy set; and determines the membership degree of objects to fuzzy sets. Rule evolution and defuzzification process described in the next sections.

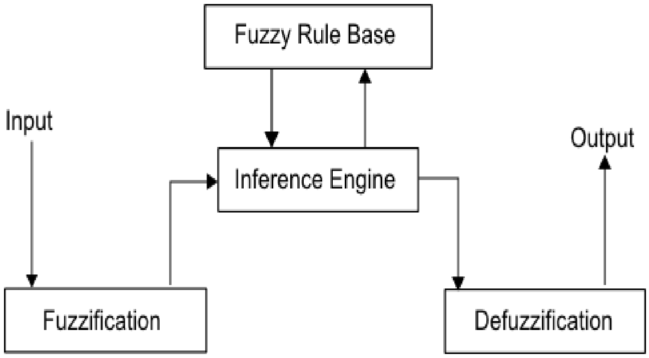


Figure 2 : General Schema of Fuzzy Logic

i. *Fuzzification and Membership Function*

Fuzzification is a process of fuzzifying all inputs and output. Determine the degree to which these inputs and outputs belong to each of the suitable fuzzy sets.

Age:

We have actually separated age into 3 fuzzy sets (Young. Middle and Old) and ranges of these fuzzy sets are determined in Table (1). We have used triangular and trapezoidal Membership function for fuzzy sets.

$\mu young(x) = 1$ when $x \in [1, 29]$
$\mu young(x) = (38- x)/(38-29)$ when $x \in [29, 38]$
$\mu young(x) = 0$ otherwise.

$\mu middle(x) = (x-33)/(39-33)$ when $x \in [33, 39]$
$\mu middle(x) = (45- x)/(45-39)$ when $x \in [39, 45]$
$\mu middle(x) = 0$ otherwise.

$\mu old(x) = 0$ when $x \in [1, 40]$
$\mu old(x) = (x-40)/(60-40)$ when $x \in [40, 60]$
$\mu old(x) = 1$ when $x \in [60, 100]$.

Blood Pressure:

We have actually separated this input fuzzy set into 4 levels called (Low, Medium, High and Very high). Trapezoidal MFs are used for (Low and Very high) and MFs of (medium and high) sets are triangular.

$\mu low(x) = 1$ when $x \in [1, 111]$
$\mu low(x) = (134- x)/(134-111)$ when $x \in [111, 134]$
$\mu low(x) = 0$ otherwise.

$\mu medium(x) = (x-126)/(139-126)$ when $x \in [126, 139]$
$\mu medium(x) = (152- x)/(152-139)$ when $x \in [139, 152]$
$\mu medium(x) = 0$ otherwise.

$\mu high(x) = (x-142)/(157-142)$ when $x \in [142, 157]$
$\mu high(x) = (172- x)/(172-157)$ when $x \in [157, 172]$
$\mu high(x) = 0$ otherwise.

$\mu veryhigh(x) = 0$ when $x \in [1, 154]$
$\mu veryhigh(x) = (x-154)/(172-154)$ when $x \in [154, 172]$
$\mu veryhigh(x) = 1$ when $x \in [172, 300]$.

Cholesterol:

Cholesterol has 3 fuzzy sets (Low, Medium and High). Ranges for these fuzzy sets are determined in Table (1).

$\mu low(x) = 1$ when $x \in [1, 151]$
$\mu low(x) = (197- x)/(197-151)$ when $x \in [151, 197]$
$\mu low(x) = 0$ otherwise.

$\mu medium(x) = (x-188)/(219-188)$ when $x \in [188, 219]$
$\mu medium(x) = (250- x)/(250-219)$ when $x \in [219, 250]$
$\mu medium(x) = 0$ otherwise.

$\mu high(x) = 0$ when $x \in [1, 217]$
$\mu high(x) = (x-217)/(263-217)$ when $x \in [217, 263]$
$\mu high(x) = 1$ when $x \in [263, 500]$.

Heart Rate:

Heart Rate split into 3 fuzzy sets named (Low, Medium and High). Ranges for these fuzzy sets are identified in table (1). MFs of (Low and High) sets are trapezoidal and MF of (Medium) is triangular.

$\mu low(x) = 1$ when $x \in [1, 100]$
$\mu low(x) = (141- x)/(141-100)$ when $x \in [100, 141]$
$\mu low(x) = 0$ otherwise.

$\mu medium(x) = (x-111)/(152-111)$ when $x \in [111, 152]$
$\mu medium(x) = (194- x)/(194-152)$ when $x \in [152, 194]$
$\mu medium(x) = 0$ otherwise.

$\mu high(x) = 0$ when $x \in [1, 152]$
$\mu high(x) = (x-152)/(216-152)$ when $x \in [152, 216]$
$\mu high(x) = 1$ when $x \in [216, 450]$.

Old Peak:

Old Peak divided into 3 fuzzy sets (Low, Risk and Terrible). These fuzzy sets have actually been shown in Table (1) with their ranges.

$\mu low(x) = 1$ when $x \in [0, 1]$
$\mu low(x) = (2- x)/(2-1)$ when $x \in [1, 2]$
$\mu low(x) = 0$ otherwise.

$\mu risk(x) = (x-1.5)/(2.8-1.53)$ when $x \in [1.5, 2.8]$
$\mu risk(x) = (4.2- x)/(4.2-2.8)$ when $x \in [2.8, 4.2]$
$\mu risk(x) = 0$ otherwise.

$\mu terrible(x) = 0$ when $x \in [0, 2.55]$
$\mu terrible(x) = (x-2.55)/(4.2-2.55)$ when $x \in [2.55, 4.2]$
$\mu terrible(x) = 1$ when $x \in [4.2, 6]$.

Thallium Scan:

This input field includes 3 fuzzy sets: (Normal, Fix Defect and Reverse Defect). For each and every fuzzy fuzzy set we have defined a value that we use them for system testing. These fuzzy sets with their values are shown in Table (1).

µnormal(x) = (x-1)/(2-1) when x ∈ [1, 2]
µnormal(x) = (3- x)/(3-2) when x ∈ [2, 3]
µnormal(x) = 0 otherwise.

µFix Defect(x) = (x-3)/(4.5-3) when x ∈ [3, 4.5]
µFix Defect(x) = (6- x)/(6-4.5) when x ∈ [4.5, 6]
µFix Defect(x) = 0 otherwise.

µRev Defect (x) = (x-6)/(6.5-6) when x ∈ [6, 6.5]
µRev Defect(x) = (7- x)/(7-6.5) when x ∈ [6.5, 7]
µRev Defect(x) = 0 otherwise.

Over we have actually selected chosen features, now we have split all inputs into fuzzy sets; we have actually utilized trapezoidal and triangular membership functions in system.

Input Field	Range	Fuzzy Sets
Age	<38 33-45 40>	Young Middle Old
Blood Pressure	< 138 126- 152 142-172 154>	Low Medium High Very High
Cholesterol	< 197 188-250 217>	Low Medium High
Heart rate	< 141 111-194 152>	Low Medium High
Old Peak	<2 1.5 - 4.2 2.5>	Low Risk Terrible
Thallium Scan	3 6 7	Normal Fixed Defect Reversible Defect

Table 1 : Risk factors and Ranges

ii. *Rules Evolution*

In Fuzzy Rule Base System, rules play an important role in the prediction. The rules deliver/provide a sense to linguistic variables and MF (membership function). So we have occupied these fuzzyfied inputs in antecedent part of the rules. In this research, we have actually utilized 19 rules to predict heart disease in the patient.

In our system antecedent part of the rule consist of only single part that will opinion result of antecedent development. Fuzzy logic will govern risk level and this prediction indeed relies on rules that we have made. The made rules we have applied in using Mat Lab R2012a in the rule editor. After then fuzzification, crisp will certainly examine by passing in rule instance.

iii. *Output (Defuzzifiction)*

In this system, we have one output variable, which divided to 2 fuzzy sets (healthy and sick). For defuzzification procedure, designed system makes use of the Centroid method, determines the area of membership functions within the range of (output) variable.

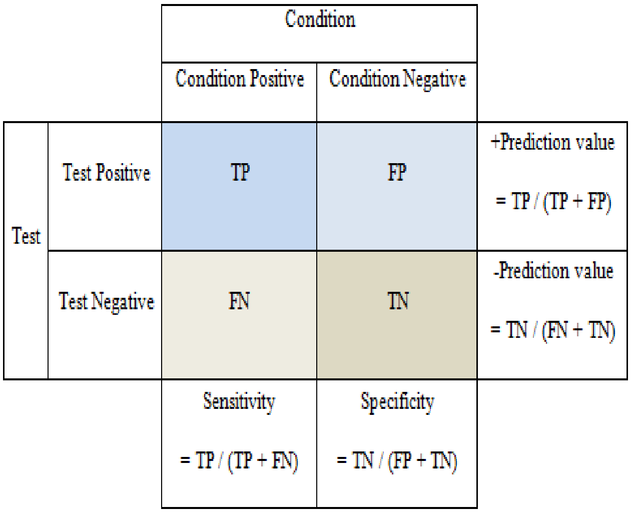
$$\mathrm{CoA} = \frac{\int_{x_{\min}}^{x_{\max}} f(\mathrm{x}).\mathrm{xdx}}{\int_{x_{\min}}^{x_{\max}} f(\mathrm{x})\,\mathrm{dx}}$$

CoA is the center of area/gravity; x is the linguistic variable and x (min) and x (max) signifies the arrays of variables.

IV. System Testing

To compare the performance of our system with Neural Network and J48 Decision Tree, we have divided Cleveland Heart Disease datasets into 2 parts e.g. training data 60% and testing data 40%. And this efficiency/performance is usually matched in term of sensitivity, specificity and accuracy. These terms normally took advantage of diagnostic approaches to enhance analysis results.

Sensitivity = TP / (TP + FN)
Specificity = TN / (FP + TN)
Accuracy = (TP+TN) / (TP+TN+FP+FN)



		Condition		
		Condition Positive	Condition Negative	
Test	Test Positive	TP	FP	+Prediction value = TP / (TP + FP)
	Test Negative	FN	TN	-Prediction value = TN / (FN + TN)
		Sensitivity = TP / (TP + FN)	Specificity = TN / (FP + TN)	

Figure 3 : Illustrates how the Positive Predictive Value, Negative Predictive Value, Sensitivity, and Specificity are Related

We use specificity to analyse and assess the amount of true positives predicted accurately. Specificity analyses and measure the amount of true negative predicted accurately. Accuracy can be obtained by sum of True Positive and True Negative divided with the total number of instances.

Example: No of healthy (True Positive) and sick (True Negative) peoples predicted correctly.

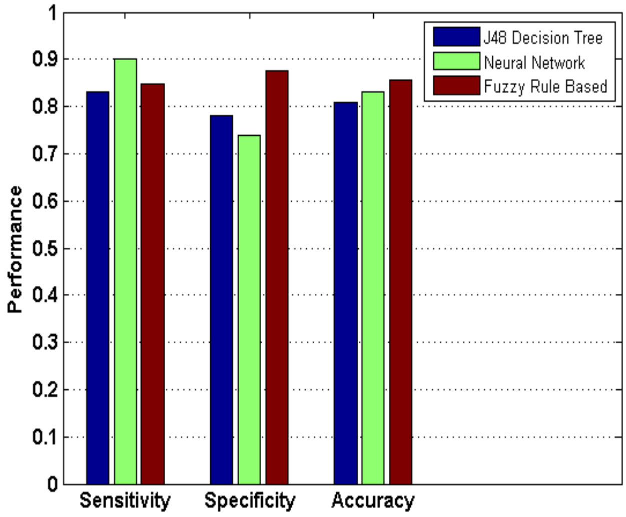


Figure 4 : Training Datasets Performance

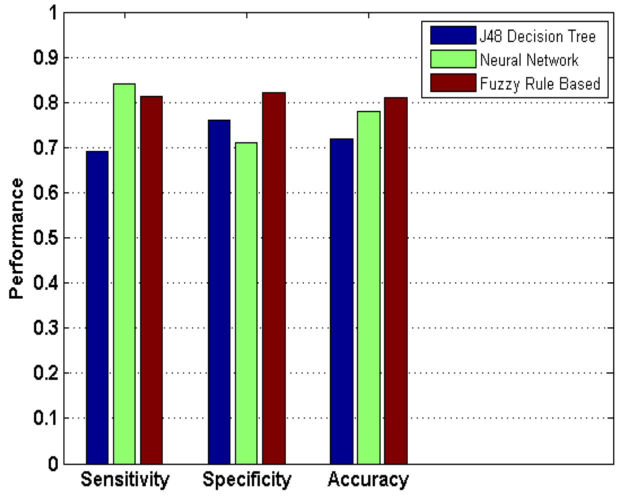


Figure 5 : Testing Datasets Performance

The performance testing has been performed on both training and testing data sets, and we merely used Cleveland heart disease datasets establishes to evaluate our system. And, in last we compared the performance of our system with Neural Network and J48 Decision Tree model using very same data sets. The final result we obtained for training and testing data are shown in figure (5-2a and 5-2b).

V. Conclusion

This proposed system "Fuzzy Rule Based Support System" modelled to predict heart disease intelligently and efficiently, and to replace manual efforts. Experts system can be more proficient and fast so it can be more accurate then manual work. Our system modelled to diagnosis and detecting cardiovascular diseases, the system involves two major phases, one that performs classification and diagnosis, the other one that detects the rate of risks of the respiratory diseases. For this system we have used mamdani inference system. In final this system tested and compared with Neural Network and J48 Decision Tree model to check performance of the system.

VI. Acknowledgement

This work is supported by National Natural Science Foundation of China under grant No.60970096 and 90818004, the Postdoctoral Science Foundation of China under Grant No.20080440988 and the Natural Science Foundation of Hunan Province, China Under Grant No.09JJ4030.

References Références Referencias

1. World Health Organization Report of the Global Atlas http://apps.who.int/globalatlas/.
2. Mendis, S., Puska, P., Norrving, B. (2011). Global Atlas on cardiovascular disease prevention and control. ISBN 978-92-4-156437-3.
3. Gale Nutrition Encyclopaedia (2011). Heartdiseasehttp://www.answers.com/topic/ischaemic-heart-disease (Accessed 25 February 2011).
4. Countries, Committee on Preventing the Global Epidemic of Cardiovascular Disease: Meeting the Challenges in Developing; Fuster, Board on Global Health; Valentin; Academies, Bridget B. Kelly, editors; Institute of Medicine of the National (2010). Promoting cardiovascular health in the developing world: a critical challenge to achieve global health. Washington, D. C.: National Academies Press. pp. Chapter 2. ISBN 978-0-309-14774-3.
5. Merijohn G. K., Bader J. D., Frantsve-Hawley J., (2008). Clinical decision support chair side tools for evidence-based dental practice. The Journal of Evidence-Based Dental Practice, 8 (3) 2008, pp. 119–132.
6. Garg N.K., Adhikari, McDonaldH. (2005). Effects of computerized clinical decision support systems on practitioner performance and patient outcomes. Journal of the American Medical Association. PubMed, 293 (10). pp. 1223–1238.
7. AbbasiM.M., KashiyarndiS. (2006). Clinical decision support systems: a discussion on different methodologies used in health care.
8. Warren J., Beliakov G., Zwaag B. (2000). Fuzzy logic in clinical practice decision support system. In: Proceedings of the 33rd Hawaii International Conference on System Sciences, Maui, Hawaii. 4–7 January 2000.

9. AndersonJ., Clearing the way for physicians use of clinical information systems Communication of the ACM (1997), pp. 83–90.
10. LinL., HuP.J.H., Sheng O.R.L., (2006). Decision Support Systems. 42.
11. MusenM., ShaharY., ShortliffeE.H., (2001). Clinical decision support systems.
12. Neshat M., Yaghobi M., (2009). Designing a Fuzzy Expert System of Diagnosing the Hepatitis B intensity Rate and comparing it with Adaptive Neural Network Fuzzy System. International Conference in Modelling Health Advances pp.1~6.
13. Neshat M., Yaghobi M., NaghibiM.B., (2008). Fuzzy Expert System Design for Diagnosis of liver disorders. IEEE Computer Society. pp. 252~256.
14. AnoojP.K., (2012). Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. "International Journal of Research and Reviews in Computer Science (IJRRCS)" Vol. 3, No. 3. ISSN: 2079-2557.
15. Persi Pamela I., Gayathri. P., JaisankarN., (2013). A Fuzzy Optimization Technique for the Prediction of Coronary Heart Disease Using Decision Tree. International Journal of Engineering and Technology (IJET). Vol 5 No 3.
16. Nidhi Bhatla, Kiran Jyoti (2012). A Novel Approach for Heart Disease Diagnosis using Data Mining and Fuzzy Logic. International Journal of Computer Applications (0975 – 8887) Volume 54– No.17.
17. Vijay Kumar Magoa, Nitin Bhatia, Ajay Bhatia, (2012). Clinical decision support system for dental treatment. Journal of Computational Science. 3 (2012) 254–261.
18. Syed Umar Amin, Kavita Agarwal, Dr. Rizwan Beg (2013). Data Mining in Clinical Decision Support Systems for Diagnosis, Prediction and Treatment of Heart Disease. International Journal of Advanced Research in Computer Engineering & Technology. (IJARCET) Volume 2, Issue 1.
19. Asha Rajkumar, Mrs. G.Sophia Reena (2010). Diagnosis of Heart Disease Using Data mining Algorithm "GJCST Classification J.3. Global Journal of Computer Science and Technology. P a g e |38 Vol. 10 Issue 10 Ver. 1.0.
20. AnbarasiM., Anupriya*E., IyengarN.CH. (2010). Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm. International Journal of Engineering Science and Technology. Vol. 2(10). 5370-5376.
21. Rajeswari*K., VaithiyanathanV., (2011). Heart disease diagnosis: an efficient decision support system based on fuzzy logic and genetic algorithm. International Journal of Decision Sciences, Risk and Management, Vol. 3, Nos. 1/2.
22. Han J. and Kamber M., (2006). Data Mining: Concepts and Techniques. Second Edition, Morgan Kaufmann Publishers, San Francisco.
23. DashM. and LiuH., (1997). Feature selection for classification. Intelligent Data Analysis: An International Journal, 1(3):131–156.

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 14 Issue 3 Version 1.0 Year 2014
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Gene Expression Analysis Methods on Microarray Data - A Review

By Prof. G. V. Padma Raju, Dr. Srinivasa Rao Peri
& Dr. Chandra Sekhar Vasamsetty

SRKR Engineering College Affiliated to Andhra University, India

Abstract- In recent years a new type of experiments are changing the way that biologists and other specialists analyze many problems. These are called high throughput experiments and the main difference with those that were performed some years ago is mainly in the quantity of the data obtained from them. Thanks to the technology known generically as microarrays, it is possible to study nowadays in a single experiment the behavior of all the genes of an organism under different conditions. The data generated by these experiments may consist from thousands to millions of variables and they pose many challenges to the scientists who have to analyze them. Many of these are of statistical nature and will be the center of this review. There are many types of microarrays which have been developed to answer different biological questions and some of them will be explained later. For the sake of simplicity we start with the most well known ones: expression microarrays.

Keywords: *micro array, classification.*

GJCST-C Classification : *H.2.8*

Gene Expression Analysis Methods on Microarray Data - A Review

Prof. G. V. Padma Raju [α], Dr. Srinivasa Rao Peri [σ] & Dr. Chandra Sekhar Vasamsetty [ρ]

Abstract- In recent years a new type of experiments are changing the way that biologists and other specialists analyze many problems. These are called high throughput experiments and the main difference with those that were performed some years ago is mainly in the quantity of the data obtained from them. Thanks to the technology known generically as microarrays, it is possible to study nowadays in a single experiment the behavior of all the genes of an organism under different conditions. The data generated by these experiments may consist from thousands to millions of variables and they pose many challenges to the scientists who have to analyze them. Many of these are of statistical nature and will be the center of this review. There are many types of microarrays which have been developed to answer different biological questions and some of them will be explained later. For the sake of simplicity we start with the most well known ones: expression microarrays.

Keywords: micro array, classification.

I. Introduction

Microarrays and other genomic data are different in nature from the classical data around which most statistical techniques have been developed. In consequence, in many cases it has been necessary to adapt existing techniques or to develop new ones in order to fit the situations encountered. We will examine some key components of microarray analysis, experimental design, quality control, preprocessing and statistical analysis. In the last section we will consider some topics where open questions still remain and which can be considered attractive for statisticians who wish to focus some of their research in this field. One of the handicaps for statisticians who may consider entering this field is how to start applying their knowledge to these problems. We will present some real examples, which we will use along the paper to illustrate some concepts [1-15].

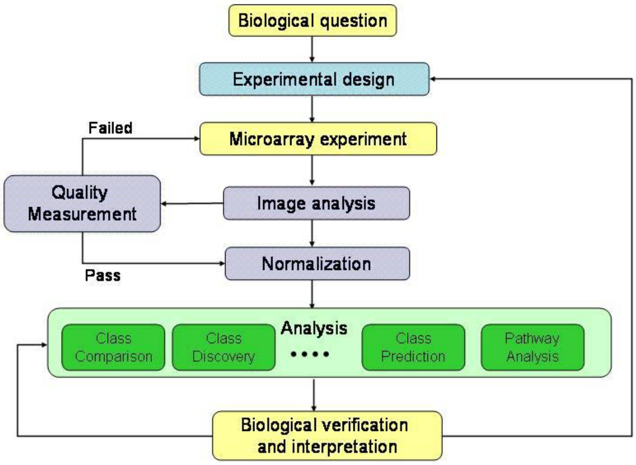


Figure 1 : The microarray analysis process

The goal of this section is to present an integrated view of the whole process of analyzing microarray data (see figure 1). Many review papers discuss the statistical techniques available for the analysis at this level.

II. Methods for Classification

Different strategies have been proposed over the last several years for feature/gene selection: filter, wrapper, embedded [16], and more recently ensemble techniques [17].

Filter techniques assess the discriminative power of features based only on intrinsic properties of the data. As a general rule, these methods estimate a relevance score and a threshold scheme is used to select the best-scoring features/ genes. Filter techniques are not necessarily used to build predictors. As stated in [18], DEGs may also be good candidates for genes which can be targeted by drugs. This group of techniques is independent of any classification scheme but under particular conditions they could give the optimal set of features for a given classifier. Saeys et al. [1] also stress on the practical advantages of these methods stating that “even when the subset of features is not optimal, they may be preferable due to their computational and statistical scalability.”

Wrapper techniques select the most discriminant subset of features by minimizing the prediction error of a particular classifier. These methods are dependent on the classifier being used and they are

Author α ρ: Dept of CSE, SRKR Engineering College, Bhimavaram, Andhra Pradesh, INDIA. e-mail: gvpadmaraju@gmail.com
Author σ: Dept of CS&SE, AU College of Engineering, Visakhapatnam, Andhra Pradesh, India.

mainly criticized because of their huge computational demands. More than that, there is no guarantee that the solution provided will be optimal if another classifier is used for prediction.

Embedded techniques represent a different class of methods in the sense that they still allow interactions with the learning algorithm but the computational time is smaller than wrapper methods.

Ensemble techniques represent a relatively new class of methods for FS. They have been proposed to cope with the instability issues observed in many techniques for FS when small perturbations in the training set occur. These methods are based on different sub sampling strategies. A particular FS method is run on a number of subsamples and the obtained features/genes are merged into a more stable subset [19].

a) Filter Methods - A Ranking Approach

Most filter methods consider the problem of FS as a ranking problem. The solution is provided by selecting the top scoring features/genes while the rest are discarded. Generally these methods follow a typical scenario described below.

1. Use a scoring function S(x) to quantify the difference in expression between different groups of samples and rank features/genes in decreasing order of the estimated scores. It is supposed that a high score is indicative for a DEG.
2. Estimate the statistical significance (e.g., p-value, confidence intervals) of the estimated scores.
3. Select the top ranked features/genes which are statistically significant as the most informative features/ genes (alternatively one could be interested in selecting the top ranked features/genes only as opposed to the top ranked significant ones).
4. Validate the selected subset of genes.

In the above-mentioned generic algorithm one can identify two aspects specific to this type of methods which play an important role in identifying informative features/genes: first, the choice of a scoring function to compute the relevance indices (or scores) and second, the assignment of statistical significance to computed scores. They will receive further consideration in order to be able to reveal the main differences between different methods and therefore helping to categorize them. As an additional remark, the reader should note that ranked lists of features/genes can also be obtained via wrapper/embedded methods not only for filters, e.g., SVM.

Recursive Feature Elimination (SVMRFE) [20] or Greedy Least Square Regression [21].Here we also outline the fact that any combination of a scoring function and a statistical significance test designed to quantify the relevance of a feature/gene for a target annotation can be transformed into a ranking method for FS. Since all steps in the generic algorithm described above are independent one from another, the users do have a lot of freedom in the way they wish to perform the selection.

b) Scoring Functions - Assigning Relevance Indices to Features

Scoring functions represent the core of ranking methods and they are used to assign a relevance index to each feature/gene. The relevance index actually quantifies the difference in expression (or the informativeness) of a particular feature/gene across the population of samples, relative to a particular target annotation. Various scoring functions are reviewed and categorized here. They cover a wide range of the literature proposed for DEGs or biomarkers discovery. The scoring functions are enumerated and categorized according to their syntactic similarities. A similar approach presenting a very comprehensive survey on distance measures between probability density functions has been employed in [22].

Several groups of scoring functions for gene ranking have been identified. In the first group, we gathered scoring functions which estimate an average rank of genes across all samples. Scoring functions from the second group quantify the divergence (or the distance) between the distributions of samples corresponding to different classes associated to a target annotation per feature/gene. The third group contains information theory-based scoring functions while the fourth group measures the degree of association between genes and a target annotation. The last group gathers a list of miscellaneous scoring functions which cannot be included in the previous four. The big majority of scoring functions presented here are usually defined to rank single genes but some of them can be easily adapted for pairs or groups of genes.

i. *Ranking Samples across Features*

This group is represented by two scoring functions: rank sum and rank-product. Supposing x1 and x2 are the expression levels of a certain gene in class c1 and class c2, respectively, the rank-sum method first combines all the samples in x1 and x2 and sorts them in ascending order. Then the ranks are assigned to samples based on that ordering. If k samples have the same value of rank i, then each of them has an average rank. If n1 and n2 denote the numbers of samples in the smaller and larger group, respectively, then the rank-sum score is computed by summing up the ranks corresponding to samples in c1. For a GEM data set, the rank-product method consists in ordering the genes across all samples in the value ascending order and then for each gene the rank product score is obtained by taking the geometrical average of the ranks of that gene in all samples.

ii. *Measuring the Divergence between the Distributions of Groups of Samples*

Another direction toward the identification of informative features/genes is to quantify the difference between the distributions of groups of samples associated to a target annotation. These scoring functions can be generically described as a function f(x1; x2) with x1; x2. For this purpose, some simple measures rely only on low-order statistics, in particular the first and second moment (mean and variance) of the distribution of expression levels in different groups. This is the simplest way to compare the distributions of two populations and implicitly imposes some more or less realistic assumptions on the distributions of samples in each population (e.g., normal distributed samples). Despite this obvious drawback they are still the most popular scoring functions used to create filters for FS in GEM analysis due to their simplicity. These scoring functions can be grouped in two families: fold-change family and t-test family. A different strategy in comparing the distributions of different populations is to rely on different estimates of the probability density function (pdf) or the cumulative density function (cdf) of populations but these methods are more expensive computationally. The different families of scoring functions mentioned here will be further presented in this section.

a. *Fold-change family*

Relative indices are assigned to features/genes based only on mean estimates of the expression levels across different groups of samples per gene. According to [23] two forms are encountered for the fold-change scoring functions: fold-change ratio and fold change difference. However, the fold-change difference is less known and usually researchers who mention fold-change in this context actually refer to fold change ratio. In practice, many packages for GEM analysis typically provide the log2 of the ratio between the means of group 1 and group 2. The numbers will be either positive or negative preserving the directionality of the expression change. t-test family. Several forms derived from the ordinary two-sample t-test are used to measure the difference in expression of genes. In the same family, we include the Z-score or the signal to noise ratio (SNR) defined as the ratio between the fold-change difference and the standardized square error of a particular gene. These scoring functions make use of both the first and second moments to assign relevance indices to genes.

b. *Bayesian scoring functions*

In several studies, the authors have defined scoring functions for informative features discovery in a Bayesian framework. The main motivation behind this is the difficulty in obtaining accurate estimates of the standard deviation of individual genes based on few measurements only. In order to cope with the weak empirical estimation of variance across a single feature/gene, several authors proposed more robust estimations of the variance by adding genes with similar expression values.

c. *PDF-based scoring functions*

Scoring functions in this category rely on different estimates of the pdfs of populations, from simple histograms to more complex estimators such as the Parzen window estimator [24]. Only few scoring functions based on this idea are used to discover informative features/genes. Here we identified Kolmogorov-Smirnov (K-S) tests [25], Kullback-Leibler divergence [26], or Bhattacharyya distance [27], but the mathematical literature abounds in measures quantifying the distance between pdfs revealing new possibilities to look for informative features/genes. We invite the reader to consult for a very comprehensive survey on this topic. Note that the use of these scoring functions for DEGs discovery is limited by the low number of samples in GEMexperiments which results in unreliable estimates of the pdf.

iii. *Information Theory-Based Scoring Functions*

These scoring functions rely on different estimates of the information contained both in the target feature c and in the gene expression x.

iv. *Measuring the Dependency between Features and Target Feature as a Function*

Scoring functions in this group have the advantage that they allow features/genes ranking when the target annotation is a continuous variable (which is not the case of the previous mentioned scoring functions). They measure the dependency between the gene's expression profile x and the target feature c as a function f(x,c). Pearson's correlation coefficient (PCCs),Its absolute value equals 1 if x and c are linearly correlated and equals 0 if they are uncorrelated. Note that PCCs is only applied if c is a continuous variable. When c is binary, PCCs comes down to the Z - score. A similar measure used for this purpose is Kendall's rank correlation coefficient (KRCCs). A variant of this measure adapted to a two-class problem is proposed in [28].

v. *Other Scoring Functions*

A list of scoring functions mentioned in the literature for informative gene discovery which cannot be grouped in the above-mentioned families is presented here. The list presented in Table 1 includes: Area Under ROC Curve (AUC), Area Between the Curve and the Rising diagonal (ABCR), Between-Within class Sum of Squares (BWSS), and Threshold Number of Miss classifications (TNoM). The reader is encouraged to consult the associated references in Table 1 for further details about these scoring functions.

Table 1 : Other scoring functions for gene ranking

AUC	$S = AUC = \sum_{k=1}^{n_0} AUC_k$ n_0 Number of individual values of gene x [29]
ABCR	$S = ABCR = \sum_{k=1}^{n_0} \|\|AUC_k - A_k\|\|$ Where $A_k = \frac{2k-1}{2n_0^2}$ [29]
BWSS	$S = BW = \frac{\sum_i \sum_k (c_i=k)(\bar{x}_k - \bar{x})2}{\sum_i \sum_k (c_i=k)(x_k - \bar{x_k})2}$ [30]
TNoM	S=TNoM= $\min_{d,t}$Err(d,t|x,c) [31]

c) *Estimating Statistical Significance for Relevance Indices*

Estimating the statistical significance for the relevance indices assigned to each feature/gene has been long addressed in the quest for DEGs. It is argued that statistical significance tests quantify the probability that a particular score or relevance index has been obtained by chance. It is common practice that features/genes ranked high in the list according to the relevance index, will be discarded if the computed scores are not statistically significant. There are different ways one can assign statistical significance despite many criticisms the most commonly used statistical significance test is the p-value. Many researchers advocate for alternative measures such as confidence intervals, especially due to the fact that p-values only bring evidence against a hypothesis (e.g., the null hypothesis of no "correlation" between features/genes and target annotation) and "confirm" a new hypothesis by rejecting the one which has been tested without bringing any evidence in supporting the new one [32]. Without entering into this debate, it is important to notice that statistical significance tests can be run either by exploring gene-wise information across all samples, either by exploring the large number of features in GEM experiments. Regardless the manner the statistical significance tests are performed, a permutation test is generally employed. It consists of running multiple tests which are identical to the original except that the target feature (or the class label) is permuted differently for each test. An important concept for estimating the statistical significance for DEGs discovery is the multiple hypotheses testing which will be described at the end of this section.

i. *Exploring Feature-Wise Information to Asses Statistical Significance*

This strategy assumes a large enough number of samples in order to infer upon the statistical significance of computed relevance indices of genes. The statistical significance is estimated for each feature/gene individually based on its intrinsic information. p-values. In statistics, the p-value is the probability of obtaining a test statistic (in our case a relevance index) at least as extreme as the one that was actually observed. The lower the p-value the more significant the result is (in the sense of statistical significance). Typical cutoff thresholds are set to 0.05 or 0.01 corresponding to a 5 or 1 percent chance that the tested hypothesis is accepted by chance. Pvalues can be estimated empirically by using a permutation test. However, standard asymptotic methods also exist, reducing substantially the computational time required by permutation tests. These methods rely on the assumption that the test statistic follows a particular distribution and the sample size is sufficiently large. When the sample size is not large enough, asymptotic results may not be valid, with the asymptotic p-values differing substantially from the exact p-values.

ii. *Exploiting the Power of Large Number of Features*

An alternative strategy to overcome the drawback of the small number of samples in GEM experiments is to take advantage of the large number of features/genes [33]. In order to illustrate this idea we will consider the following: a GEM data set containing gene information about samples originating from two populations c1 and c2, and a filter algorithm to search for DEGs between c1 and c2.

iii. *Multiple Hypothesis Testing Approach*

The study of Dudoit et al. [34] was the first work describing the multiple hypothesis testing for GEM experiments in a statistical framework. In the context of DEGs discovery, multiple hypothesis testing is seen as simultaneously testing for each gene the null hypothesis of no association between the expression level and the responses or target features [34]. According to them, any test can result in two type of errors: false positive or Type I errors and false negative or Type II errors. Multiple hypothesis testing procedures aim to provide statistically significant results by controlling the incidence rate of these errors. In other words, provide a way of setting appropriate thresholds in declaring a result statistically significant. The most popular methods for multiple hypothesis testing focus on controlling Type I error rate. This is done by imposing a certain threshold for the Type I error rate and then applying a method to produce a list of rejected hypothesis until the error rate is less than or equal with the specified threshold.

p-value with Bonferroni correction is an improved version of the classical p-value and consists in increasing the statistical threshold for declaring a gene significant by dividing the desired significance with the number of statistical tests performed [35].

False discovery rate (FDR) is a recent alternative for significance testing and has been proposed as an extension of the concept of p-values [36]. The FDR is defined as FDR =[F/G] , where F is the number of false positive genes and G is the number of genes found as being significant. In order to overcome the situations where FDR is not defined (when G = 0), Storey [37] proposed a modified version of the FDR called positive false discovery rate (pFDR) defined as Pfdr= [E/F|G > 0].

A less accurate alternative to the FDR for significance testing is the family-wise error rate (FWER) which is defined as the probability of at least one truly insignificant feature to be called significant. q-value is an extension of FDR which has been proposed to answer the need of assigning a statistical significance score to each gene in the same way that the p-value does [38]. The q-value is defined as being the minimum pFDR at which a test may be called significant. The reader should be aware that the q-value can be defined either in terms of the original statistics or in terms of the p-values.

d) Ranking Methods for FS - Examples

In this section, we discuss and review ranking methods for FS by extending the taxonomy presented in Fig. 1.

i. *Univariate Methods*

According to [16], univariate methods for FS can be either parametric or nonparametric. Here, we provide a brief description of both groups.

a. *Parametric methods*

These methods rely on some more or less explicit assumption that the data are drawn from a given probability distribution. The scoring functions used to measure the difference in expression between groups of samples for each gene provide meaningful results only if this assumption holds. In particular, many researchers state that the t-test can be used to identify DEGs only if the data in each class are drawn from some normal distribution with mean and standard deviation.

b. *Nonparametric methods*

These methods assume by definition that the data are drawn from some unknown distribution. The scoring functions used to quantify the difference in expression between classes rely either on some estimates of the pdfs or on averaged ranks of genes or samples. Obviously, these methods have a higher generalization power but for most of them (especially those relying on estimates of the pdfs), the computational cost is higher. In [16], univariate nonparametric filter techniques are split in two groups: pure model-free methods and methods based on random permutation associated to parametric tests. Pure model free methods use nonparametric scoring functions to assign a relevance index to each gene and then the statistical relevance of that index is estimated in terms of either p-value, FDR or q-value. Methods based on random permutations associated with a parametric test take advantage on the large number of genes/features in order to find genes/features which present significant changes in expression. In a first instance, they make use of a parametric scoring function to assign a relevance index to each gene and then employ a nonparametric statistical significance test to check for DEGs. The nonparametric significance test consists in comparing the distribution of relevance indices of genes estimated in the previous step and the null distribution of the test statistic (or relevance index). The null distribution of the test statistic is usually estimated using a permutation test.

ii. *Bivariate Ranking Methods*

Ranking pairs of genes according to their discrimination power between two or more conditions can be performed either using a “greedy strategy” or “all pair strategy.” Greedy strategies. Methods in this group first rank all genes by individual ranking (using one of the criteria employed by univariate ranking methods); subsequently the highest scoring gene gi is paired with the gene gj that gives the highest gene pair score. After the first pair has been selected, the next highest ranked gene remaining gs is paired with the gene gr that maximizes the pair score, and so on. In [39], a greedy gene pair ranking method has been proposed where initially the t-test was employed to first rank genes individually while the pair score measures how well the pair in combination distinguishes between two populations. Concretely, the gene pair score is the t-test of the projected coordinates of each experiment on the diagonal linear discriminant (DLD) axis, using only these two genes. For further details we invite the reader to consult [39].

All pairs strategies. Unlike greedy pairs methods, all pairs strategies examine all possible gene pairs by computing the pair score for all pairs. The pairs are then ranked by pair score, and the gene ranking list is compiled by selecting non overlapping pairs, and selecting highest scoring pairs first. This method is computationally very expensive.

e) Filter Methods - Space Search Approach

The second direction to create filters for FS is to adopt an optimization strategy which will come up with the most informative and least redundant subset of features among the whole set. This strategy implies three main steps described as follows:

1. Define a cost function to optimize.
2. Use an optimization algorithm to find the subgroup of features which optimizes the cost function.
3. Validate the selected subset of genes.

III. Our Contribution

This work categorizes the algorithms into different categories to emphasize the data structure that drives the matching. We will give in this section some characteristics of standard clustering methods in relation to microarray data analysis. Hierarchical clustering has been mainly used to find a partition of the samples more than of the genes because there are much less samples than genes so that, with genes, the resulting dendrogram is often difficult to interpret.

Algorithms Designed After 2000

In this section we survey the most classical micro array algorithms that have been designed after year 2000. In particular the algorithms based on comparisons and the algorithms based on micro array. Most of the comparison-based algorithms presented in the last ten years are obtained by improving or combining the ideas of previously published algorithms. In the following we briefly review the state-of-the-art until 2014 and the main ideas and the algorithms to which the new solutions refer.

a) During 2010

Leila Muresan et.al [40] developed an approach for the analysis of high-resolution microarray images. First, it consists of a single molecule detection step, based on undecimated wavelet transforms, and second, a spot identification step via spatial statistics approach (corresponding to the segmentation step in the classical microarray analysis). Proposed approach relies on two independent steps. First, present a wavelet-based method to detect single molecules in each subimage. Wavelet transform offers an attractive solution for the detection of small bright features, e.g., in astronomical images or in the case of microscopy, for the detection of subcellular structures. The detection is based on the property of the wavelet transform to concentrate the information in a few wavelet coefficients, and subsequently thresholding the pixels corresponding to the signal from background. Second, separate the detected molecules inside the spot of interest (the hybridization signal) from the unspecifically bound ones. This concentration estimation approaches based on spatial statistics. The first algorithm matches the empirical moments with the moments of a mixture of two Poisson distributions representing counts of molecules outside and inside the spot. The second algorithm separates spot-bound single molecules from dirt, based on nearest neighbor distances of all the detected peak locations, via an expectation-maximization (EM) approach. Since the surface was made antiadsorptive for target molecules, we can assume that the concentration of peaks outside the spot is lower than the concentration of the hybridized molecules inside the spot. The detection method was tested on simulated images with a concentration range of 0.001 to 0.5 molecules per square micrometer and signal-to-noise ratio (SNR) between 0.9 and 31.6. For SNR above 15, the false negatives relative error was below 15%. Separation of foreground/background is proved reliable, in case foreground density exceeds background by a factor of 2. The method has also been applied to real data from high-resolution microarray measurements.

Yoshinori Tamada et.al [41] presents a novel algorithm to estimate genome-wide gene networks consisting of more than 20 000 genes from gene expression data using nonparametric Bayesian networks. Due to the difficulty of learning Bayesian network structures, existing algorithms cannot be applied to more than a few thousand genes. Present algorithm overcomes this limitation by repeatedly estimating sub networks in parallel for genes selected by neighbor node sampling. Through numerical simulation, finally confirmed that proposed algorithm outperformed a heuristic algorithm in a shorter time. Proposed algorithm to microarray data from human umbilical vein endothelial cells (HUVECs) treated with siRNAs, to construct a human genome-wide gene network, which compared to a small gene network estimated for the genes extracted using a traditional bioinformatics method. The results showed that genome-wide gene network contains many features of the small network, as well as others that could not be captured during the small network estimation. The results also revealed master-regulator genes that are not in the small network but that control many of the genes in the small network. These analyses were impossible to realize without our proposed algorithm. Analysis of the result, we also constructed a gene network with 527 genes extracted. These 527 genes are selected based on the ordinal bioinformatics analysis with SAM (Significance Analysis of Microarrays) by applying it to another drug-response microarray data which were observed for HUVECs stimulated by anti-hyperlipidemia drug Fenofibrate. For this smaller gene network, performed the bootstrap method. The number of the bootstrap iterations is 1000. The final 527 gene network is generated by removing edges whose bootstrap probabilities are less than 0.5.

Tianwei Yu et.al [42] proposes an imputation scheme based on nonlinear dependencies between genes. By simulations based on real microarray data, show that incorporating non-linear relationships could improve the accuracy of missing value imputation, both in terms of normalized root mean squared error and in terms of the preservation of the list of significant genes in statistical testing. In addition, studied the impact of artificial dependencies introduced by data normalization on the simulation results. Our results suggest that methods relying on global correlation structures may yield overly optimistic simulation results when the data has been subjected to row (gene) – wise mean removal. Six datasets were used in the simulation study. They included the B-cell lymphoma profiling data , the dataset of yeast transcriptome/translatome comparison, the NCI60 cell line gene expression data, and the GSE19119 dataset on Atlantic salmon. Two yeast cell cycle time series, the alpha factor dataset and the elutriation dataset, were used to probe the effect of data normalization on simulation results in imputation studies. Four popular imputation methods were used for comparison. They included the K-nearest neighbor (KNN) method, the Bayesian PCA (BPCA) method, the local least square (LLS) method, and the SVD method. Different percentages of missing (1%, 5%, 10%, 15% and 20%) were simulated.

Jianxing Feng et.al [43] propose a novel *Graph Fragmentation Algorithm* (GFA) for protein complex identification. Adapted from a classical maxflow algorithm for finding the (weighted) densest subgraphs, GFA first finds large (weighted) dense sub graphs in a protein-protein interaction network and then breaks each such subgraph into fragments iteratively by weighting its nodes appropriately in terms of their corresponding log fold changes in the microarray data, until the fragment subgraphs are sufficiently small. Tests on three widely used protein-protein interaction datasets and comparisons with several latest methods for protein complex identification demonstrate the strong performance of proposed method in predicting novel protein complexes in terms of its specificity and efficiency. Given the high specificity (or precision) that method has achieved, finally conjecture that our prediction results imply more than 200 novel protein complexes. In this paper authors retrieved 51 sets of microarray gene expression data concerning yeast from the GEO database where the log fold changes of expression levels are provided. Each dataset contains multiple samples (or conditions). Totally, 824 samples are contained in the 51 datasets. Since the genes expressed in each sample are different and they could also be different from the genes contained in a PPI network, use a sample of the microarray data on a PPI network if it covers at least 90% of the genes in the network under consideration. For genes that have no expression data in a certain sample, treat their (log transformed) expression values as 0. Finally, chose (randomly) 500, 600, and 700 samples to be applied on the MIPS, DIP, and BioGRID PPI networks, respectively.

Jong Kyoung Kim et.al [44] develop a hybrid generative/discriminative model which enables us to make use of unlabeled sequences in the framework of discriminative motif discovery, leading to *semi-supervised discriminative motif discovery*. Numerical experiments on yeast ChIP-chip data for discovering DNA motifs demonstrate that the best performance is obtained between the purely-generative and the purely-discriminative and the semi-supervised learning improves the performance when labeled sequences are limited. This examined the yeast ChIP-chip data published to investigate the effect of α on identifying TFBSs, and the benefit of semi-supervised learning for motif discovery. The data included the intergenic binding locations of yeast TFs which were profiled under various environmental conditions. For each TF under a particular condition, defined its original positive set to be probe sequences that are bound with P-value ≤ 0.001, where the binding P-value is evaluated according to relative intensities of spots on a microarray. To establish the importance of blending generative and discriminative approaches for discovering DNA motifs, examined the ability of DMOPSH to find true motifs by varying the size of the positive set with different values of α. The top K sequences with smallest P values from the original positive set were chosen to define a positive set and the remaining sequences were defined to be unlabeled. Similarly, chose the 3K probe sequences with largest P-values for the negative set. We ran each experiment three times with different initializations and reported the means with ±1 standard error.

Xin ZHAO et.al [45] Identifying significant differentially expressed genes of a disease can help understand the disease at the genomic level. A hierarchical statistical model named multi-class kernel-imbedded Gaussian process (mKIGP) is developed under a Bayesian framework for a multi-class classification problem using microarray gene expression data. Specifically, based on a multinomial probit regression setting, an empirically adaptive algorithm with a cascading structure is designed to find appropriate featuring kernels, to discover potentially significant genes, and to make optimal tumor/cancer class predictions. A Gibbs sampler is adopted as the core of the algorithm to perform Bayesian inferences. A prescreening procedure is implemented to alleviate the computational complexity. The simulated examples show that mKIGP performed very close to the Bayesian bound and outperformed the referred state-of-the-art methods in a linear case, a non-linear case and a case with a mislabeled training sample. Its usability has great promises to problems that linear model based methods become unsatisfactory. The mKIGP was also applied to four published real microarray datasets and it was very effective for identifying significant differentially expressed genes and predicting classes in all of these datasets. This work builds a unified kernel-induced supervised learning model under a hierarchical Bayesian framework to analyze microarray gene expression patterns. With a multinomial probit regression setting, the introduction of latent variables, and a prescreening procedure, the mKIGP model was developed for a multi-class classification problem. An algorithm with a cascading structure was proposed to solve this problem and a Gibbs sampler was built as the mechanical core to do the Bayesian inference. Given a kernel type (such as a Gaussian kernel) with the training data as input, the fitted parameter(s) of the kernel and a set of significant genes can be obtained by running the algorithm. The algorithm also offers a probabilistic class prediction for each testing sample.

Alfredo Benso et.al [46] presents a new cDNA microarray data classification algorithm based on graph theory and able to overcome most of the limitations of known classification methodologies. The classifier works by analyzing gene expression data organized in an innovative data structure based on graphs, where vertices correspond to genes and edges to gene expression relationships. One of the main contributions of the classifier stems in the ability of combining in a single algorithm high accuracy in the classification process together with the ability of detecting samples not belonging to any of the trained classes, thus drastically reducing the number of false positive classification outcomes. To validate the efficiency of the proposed approach, the paper presents an experimental comparison between the GEG-based classifier and several generic state-of-the-art multi-class and one-class classification methods on a set of cDNA microarray experiments for fifteen well known and documented diseases. Experimental results show that the GEG-based classifier is able to reach the same performances reached by multi-class classifiers when dealing with samples belonging to the considered class library, while it outperforms one-class classifiers in the ability of detecting samples not belonging to any of the trained classes. To demonstrate the novelty of the proposed approach, the authors present an experimental performance comparison between the proposed classifier and several state-of-the-art classification algorithms.

Yu-Cheng Liu et.al [47] proposed a temporal dependency association rule mining method named 3D-TDAR-Mine for three-dimensional analyzing microarray datasets. The mined rules can represent the regulated-relations between genes. Through experimental evaluation, our proposed method can discover the meaningful temporal dependent association rules that are really useful for biologists. In this paper, define the Frequently Coherent Pattern as gene expressions reaction. Furthermore, Coherent Pattern is focus on one gene in one continuous time segment to compute the gene expression value similarity between any two samples. Hence, user can depend on their required feature of Coherent Pattern to choice the similarity measure method. If user wants to discover the Coherent Pattern between two samples that have identical shape in gene expression value series. They can use the PCC (Pearson correlation coefficient). But, in the real life reaction, it not always has identical shape. The expression value series between samples also have Shifting, Scale and Trend relation. Therefore, it proposed the TS3 similarity measurement to estimate the Coherent Pattern that considers the Shifting, Scale and Trend factors.

Hong-Dong Li et.al [48] presented a new approach, called Margin Influence Analysis (MIA), designed to work with support vector machines (SVM) for selecting informative genes. The rationale for performing margin influence analysis lies in the fact that the margin of support vector machines is an important factor which underlies the generalization performance of SVM models. Briefly, MIA could reveal genes which have statistically significant influence on the margin by using Mann-Whitney *U* test. The reason for using the Mann-Whitney *U* test rather than two-sample *t* test is that Mann-Whitney U test is a nonparametric test method without any distribution-related assumptions and is also a robust method. Using two publicly available cancerous microarray datasets, it is demonstrated that MIA could typically select a small number of margin-influencing genes and further achieves comparable classification accuracy compared to those reported in the literature. The method reported here, named margin influence analysis (MIA), is quite different from previous work. it is developed based model population analysis (MPA), which is a general framework for designing bioinformatics algorithms. The MIA method is currently proposed by strictly implementing the idea of MPA and specially designed for variable selection of support vector machines. It works by first computing a large number of SVM classifiers using randomly sampled variables. Each model is associated with a margin. Then the nonparametric Mann-Whitney *U* test is employed to calculate a *p*-value for each variable, aiming at uncovering the variable that can increase the margin of a SVM model significantly. The rationale behind MIA is that the performance of SVM depends heavily on the margin of the classifier. As is known, the larger the margin is, the better the prediction performance will be. For this reason, variables that can increase the margin of SVM classifiers should be regarded as informative variables or possible biomarker candidates. On the whole, the main contributions of MIA are two folds. Firstly, it is originally from model population analysis which helps statistically establish variable rank by analyzing the empirical distributions of margins of related SVM classifiers. Secondly, it explicitly utilizes the influence of each variable on the margin for variable selection. The results for two publicly available microarray datasets show that MIA typically selects a small number of margin-influencing informative genes, leading to comparable classification accuracy compared to that reported in the literature. The distinguished features and outstanding performance may make MIA a good alternative for gene selection of high dimensional microarray data.

Yang Chen, and Jinglu Hu [49] presents a constructive heuristic algorithm, featuring an accurate reconstruction guided by a set of well-defined criteria and rules. Instead of directly reconstructing the original sequence, the new algorithm first builds several accurate short fragments, which are then carefully assembled into a whole sequence. The eSBH algorithm can achieve relatively high accuracy in reconstruction from a large spectrum, than other constructive heuristics and some meta heuristics, especially for real DNA sequences in the benchmark instance sets. The experiments on benchmark instance sets demonstrate that the proposed method can reconstruct long DNA sequences with higher accuracy than current approaches in the literature.

Jong Kyoung Kim and Seungjin Choi [50] develop a hybrid generative/discriminative model which enables us to make use of unlabeled sequences in the framework of discriminative motif discovery, leading to semi-supervised discriminative motif discovery. Here the authors, assume that each subsequence is generated by a finite mixture model with two components corresponding to motif and background models. While this generative approach is useful for finding over-represented motifs in a given target set of sequences, our simple generative model has a limitation to capture the nature of labeled sequences. Numerical experiments on yeast ChIP-chip data for discovering DNA motifs demonstrate that the best performance is obtained between the purely-generative and the purely discriminative and the semi-supervised learning improves the performance when labeled sequences are limited.

Gene selection methods aim at determining biologically relevant subsets of genes in DNA microarray experiments. However, their assessment and validation represent a major difficulty since the subset of biologically relevant genes is usually unknown. To solve this problem a novel procedure for generating biologically plausible synthetic gene expression data is proposed by Marco Muselli et.al [51]. It is based on a proper mathematical model representing gene expression signatures and expression profiles through Boolean threshold functions. Here authors showed from a statistical standpoint that we may obtain artificial data reasonably close to real gene expression data. As a consequence, we may generate biologically plausible virtual gene expression data that may be easily used to evaluate gene selection methods, since, in this case, know in advance the set of "relevant" genes. On the basis of the mathematical model, we proposed an algorithmic procedure to generate artificial gene expression data, and we showed how to apply the algorithm to the analysis of the performance of statistical and machine learning based gene selection methods. The results show that the proposed procedure can be successfully adopted to analyze the quality of statistical and machine learning-based gene selection algorithms.

Leila Muresan et.al [52] developed an approach for the analysis of high-resolution microarray images. First, it consists of a single molecule detection step, based on undecimated wavelet transforms, and second, a spot identification step via spatial statistics approach (corresponding to the segmentation step in the classical microarray analysis). The detection method was tested on simulated images with a concentration range of 0.001 to 0.5 molecules per square icrometer and signal-to-noise ratio (SNR) between 0.9 and 31.6. For SNR above 15, the false negatives relative error was below 15%. Separation of foreground/background is proved reliable, in case foreground density exceeds background by a factor of 2. The method has also been applied to real data from high-resolution microarray measurements.

Banu Dost et.al [53] introduce here a new method, TCLUST, for clustering large, genome-scale data sets. The algorithm is based on measures of co-connectedness to identify dense subgraphs present in the data. The authors have applied this method to a large reference gene expression data set, and showed that the resulting clusters show strong enrichment in known biological pathways. Although TCLUST has been shown to perform as good as or better than existing methodologies, as with any methodology, certain caveats must be noted. A possible shortcoming might be that once two vertices end up in different clusters, they are never reconnected. On the one hand, this makes the algorithm converge faster, on the other hand, it might lead to some loss of sensitivity for higher error-rates. In principle, this could be adjusted, by applying the tcg thresholds more judiciously, gaining some FN edges at the cost of some FP edges, and increasing the number of iterations.

Giorgio Valentini [54] proposed a new hierarchical strategy, inspired by the true path rule, for gene function prediction extended to the overall functional taxonomy of genes. TPR-w ensembles significantly outperform both the basic TPR and *Top-down* ensembles in the genome and ontology wide prediction of gene functions in *S. cerevisiae*. The analysis of the experimental results and a theoretical investigation of the flow of information that traverses the hierarchical ensemble show the reasons why TPR-w are well-suited to the prediction of gene functions, and suggest new research lines for the development of new hierarchy-aware gene function prediction methods. The overall results show that using a single source of evidence we can obtain a high precision and recall for specific trees of the FunCat forest.

The prevalence of chronic diseases is increasing at an alarming rate. Among them the incidence of Type-2 Diabetes is rapidly increasing globally. Although genetics could play an important role in the higher prevalence of this disease, it is not clear how genetic factors interact with environmental and dietary factors to increase their incidence. In the current study, Gene Expression Analysis was performed by the authors [55,56] to find out differentially expressed genes between Type-2 Diabetes with and without parental history. For this analysis Multivariate and Univariate outlier detection methods are used. This analysis helps in identifying the potential Candidate Genes causing Type-2 Diabetes.

b) During 2011

Mohak Shah and Jacques Corbeil [57] propose a general theoretical framework for analyzing differentially expressed genes and behavior patterns from two homogenous short time-course data. The framework generalizes the recently proposed Hilbert-Schmidt Independence Criterion (HSIC)-based framework adapting it to the time-series scenario by utilizing tensor analysis for data transformation. The proposed framework is effective in yielding criteria that can identify both the differentially expressed genes and time-course patterns of interest between two time-series experiments without requiring to explicitly cluster the data. The parameters used in the framework give the user explicit control on the type of analysis to be performed. For instance, identifying genes pertaining to the time-course patterns of interest can be done simply by choosing and adjusting an apt weight vector and does not require clustering all the genes in predefined profile sets unlike traditional clustering-based methods. Moreover, the criterion is a generalization of the integer fold-change-based methods. It is more sensitive in discerning relatively small differential expressions. Hence, it enables the user to identify the cases when genes undergo less than twofold change but are or can potentially be biologically important in our understanding of a certain treatment or condition. The results, obtained by applying the proposed framework with a linear kernel formulation, on various data sets are found to be both biologically meaningful and consistent with published studies.

Xin Zhao and Leo Wang-Kit Cheung [58] developed a hierarchical statistical model named multiclass kernel-imbedded Gaussian process (mKIGP) under a Bayesian framework for a multiclass classification problem using microarray gene expression data. Specifically, based on a multinomial probit regression setting, an empirically adaptive algorithm with a cascading structure is designed to find appropriate featuring kernels, to discover potentially significant genes, and to make optimal tumor/cancer class predictions. A Gibbs sampler is adopted as the core of the algorithm to perform Bayesian inferences. A prescreening procedure is implemented to alleviate the computational complexity. The simulated examples show that mKIGP performed very close to the Bayesian bound and outperformed the referred state-of-the-art methods in a linear case, a nonlinear case, and a case with a mislabeled training sample. Its usability has great promises to problems that linear-model-based methods become unsatisfactory. The mKIGP was also applied to four published real microarray data sets and it was very effective for identifying significant differentially expressed genes and predicting classes in all of these data sets. Comparing to a regular SVM, the most popular kernel-induced learning method, the mKIGP has three key advantages. First, the probabilistic class prediction by the mKIGP could be insightful for borderline cases in real applications. Second, the mKIGP method has implemented specific procedure for tuning the kernel parameter(s) (such as the width parameter of a GK) and the model parameters (such as the variance of the noise term). Tuning parameters have always been one of the key issues for nonlinear parametric learning methods. As the gene selection procedure is imbedded into the learner, the mKIGP is also more consistent in identifying significant genes when comparing to regular UR or RFE method with a cross-validation procedure. In the simulated studies, The authors showed that the mKIGP/GK significantly outperformed its SVM or PLR counterparts with either RFE or UR as gene selection strategy in the nonlinear example and in the example with a mislabeled training sample. We also demonstrated that mKIGP functioned much better in a multiclass classification problem when comparing to another established Gaussian-Processes-based gene selection method, GP_ARD, for the real data sets. Third, the mKIGP method can provide more useful information, such as the posterior PDF of the parameters, for further statistical analysis and inference.

Argiris Sakellariou, Despina Sanoudou, and George Spyrou [59] investigate the minimum required subsets of genes, which best classify neuromuscular disease data. For this purpose, we implemented a methodology pipeline that facilitated the use of multiple feature selection methods and subsequent performance of data classification. Five feature selection methods on datasets from ten different neuromuscular diseases were utilized. Our findings reveal subsets of very small number of genes, which can successfully classify normal/disease samples. Interestingly, we observe that similar classification results may be obtained from different subsets of genes. The proposed methodology can expedite the identification of small gene subsets with high-classification accuracy that could ultimately be used in the genetics clinics for diagnostic, prognostic, and pharmacogenomic purposes. This study reveals that using appropriate bio-informatical tools, researchers can identify subsets with very small number of genes, which achieve high-classification results, as demonstrated for the neuromuscular disease datasets analyzed herein. Toward this goal, we applied five different feature selection methods on neuromuscular disease data (rare conditions for which only limited numbers of samples and microarray datasets are available), and investigated the minimum number of gene probes for highly accurate patient/sample classification.

Microarray analysis is a method for analyzing expression levels of multiple genes at once. This method is especially suitable for identifying and classifying genes whose expression level differs in two samples. The present work focuses [60,61] on identifying and classifying genes that cause type-II diabetes with two different samples, one with parental history and other without parental history. Mahalanobis Distance, Minimum Co-variance Determinant are the statistical methods used for identifying multivariate and univariate outliers for the identified inflammatory genes, the functional classification is performed by using Gene Ontology and pathway analysis. It is observed that 38 differentially expressed genes were identified out of 39400 genes tested between diabetes with and without parental history.

c) During 2012

Pradipta Maji [62] proposed supervised attribute clustering algorithm is based on measuring the similarity between attributes using the new quantitative measure, whereby redundancy among the attributes is removed. The clusters are then refined incrementally based on sample categories. The performance of the proposed algorithm is compared with that of existing supervised and unsupervised gene clustering and gene selection algorithms based on the class separability index and the predictive accuracy of naive bayes classifier, Knearest neighbor rule, and support vector machine on three cancer and two arthritis microarray data sets. The biological significance of the generated clusters is interpreted using the gene ontology. An important finding is that the proposed supervised attribute clustering algorithm is shown to be effective for identifying biologically significant gene clusters with excellent predictive capability. The main contribution of this paper is threefold, namely,

1. Defining a new quantitative measure, based on mutual information, to calculate the similarity between two genes, which incorporates the information of sample categories or class labels.
2. Development of a new supervised attribute clustering algorithm to find coregulated clusters of genes whose collective expression is strongly associated with the sample categories.
3. Comparing the performance of the proposed method and some existing methods using the class separability index and predictive accuracy of support vector machine, K-nearest neighbor rule, and naive bayes classifier.

For five microarray data, significantly better results are found for the proposed method compared to existing methods, irrespective of the classifiers used. All the results reported in this paper demonstrate the feasibility and effectiveness of the proposed method. It is capable of identifying coregulated clusters of genes whose average expression is strongly associated with the sample categories. The identified gene clusters may contribute to revealing underlying class structures, providing a useful tool for the exploratory analysis of biological data.

Ola ElBakry, M. Omair Ahmad, and M.N.S. Swamy [63] presents a general statistical method for detecting changes in microarray expression over time within a single biological group and is based on repeated measures (RM) ANOVA. In this method, unlike the classical F-statistic, statistical significance is determined taking into account the time dependency of the microarray data. A correction factor for this RM F-statistic is introduced leading to a higher sensitivity as well as high specificity. We investigate the two approaches that exist in the literature for calculating the p-values using resampling techniques of gene-wise p-values and pooled p-values. It is shown that the pooled p-values method compared to the method of the gene-wise p-values is more powerful, and computationally less expensive, and hence is applied along with the introduced correction factor to various synthetic data sets and a real data set. These results show that the proposed technique outperforms the current methods. The real data set results are consistent with the existing knowledge concerning the presence of the genes. The algorithms presented are implemented in R and are freely available upon request. In this work, RM F-statistic, which considers the dependency of measurements across the time course, has been employed for gene identification. The p-values have been computed using both the gene-wise and pooled p-values methods. Since the gene-wise p-values procedure is based on the number of permutations for each gene, this number has to be large to achieve the granularity of the pooled p-values. The synthetic data results have shown that the pooled p-values procedure is able to detect more true positives than the gene-wise p-values method does, and hence, is preferred for microarray data analysis.

Alok Sharma, Seiya Imoto, and Satoru Miyano [64] propose a feature selection algorithm in gene expression data analysis of sample classifications. The proposed algorithm first divides genes into subsets, the sizes of which are relatively small (roughly of size h), then selects informative smaller subsets of genes (of size r < h) from a subset and merges the chosen genes with another gene subset (of size r) to update the gene subset. It repeats this process until all subsets are merged into one informative subset. It illustrates the effectiveness of the proposed algorithm by analyzing three distinct gene expression data sets. The proposed algorithm explores this phenomenon and provides a way to investigate important genes. It is observed that the algorithm finds a small gene subset that provides high classification accuracy on several DNA microarray gene expression data sets. These subsets contain top-r genes. The small number of (r) genes would help to conduct biological experiments for investigating biomarkers in a time-efficient and cost-effective manner. This method shows promising classification accuracy for all the test data sets. We also show the relevance of the selected genes in terms of their biological functions.

Andrew Janowczyk et.al [65] presents a system for accurately quantifying the presence and extent of stain on account of a vascular biomarker on tissue microarrays. It demonstrate their flexible, robust, accurate, and high-throughput minimally supervised segmentation algorithm, termed hierarchical normalized cuts (HNCuts) for the specific problem of quantifying extent of vascular staining on ovarian cancer tissue microarrays. The high-throughput aspect of HNCut is driven by the use of a hierarchically represented data structure that allows us to merge two powerful image segmentation algorithms—a frequency weighted mean shift and the normalized cuts algorithm. HNCuts rapidly traverses a hierarchical pyramid, generated from the input image at various color resolutions, enabling the rapid analysis of large images (e.g., a 1500 × 1500 sized image under 6 s on a standard 2.8-GHz desktop PC). HNCut is easily generalizable to other problem domains and only requires specification of a few representative pixels (swatch) from the object of interest in order to segment the target class. Across ten runs, the HNCut algorithm was found to have average true positive, false positive, and false negative rates (on a per pixel basis) of 82%, 34%, and 18%, in terms of overlap, when evaluated with respect to a pathologist annotated ground truth of the target region of interest. By comparison, a popular supervised classifier (probabilistic boosting trees) was only able to marginally improve on the true positive and false negative rates (84% and 14%) at the expense of a higher false positive rate (73%), with an additional computation time of 62% compared to HNCut.

Blaise Hanczar and Avner Bar-Hen [66] propose a new measure of classifier performance that takes account of the uncertainty of the error. We represent the available knowledge about the costs by a distribution function defined on the ratio of the costs. The performance of a classifier is therefore computed over the set of all possible costs weighted by their probability distribution. This method is tested on both artificial and real microarray data sets. The costs are represented by a distribution function defined on the ratio of the costs. Seven new classification cost functions have been used in experiments based on both artificial and real data sets. These experiments showed that the selection of the best classifier is very depending on the used cost functions. In many cases, the best classifier can be identified by our new measure whereas the classic error measures fail.

Pradipta Maji and Chandra Das [67] proposed a gene clustering algorithm is to group genes from microarray data. It directly incorporates the information of sample categories in the grouping process for finding groups of co-regulated genes with strong association to the sample categories, yielding a supervised gene clustering algorithm. The average expression of the genes from each cluster acts as its representative. Some significant representatives are taken to form the reduced feature set to build the classifiers for cancer classification. The mutual information is used to compute both gene-gene redundancy and gene-class relevance. The performance of the proposed method, along with a comparison with existing methods, is studied on six cancer microarray data sets using the predictive accuracy of naive Bayes classifier, K-nearest neighbor rule, and support vector machine. An important finding is that the proposed algorithm is shown to be effective for identifying biologically significant gene clusters with excellent predictive capability.

d) During 2013 & 2014

Zidong Wang et.al [68] investigates the uncertainty quantification and state estimation issues. The polytopic uncertainty model (PUM) is exploited for describing the GRNs where the parameter uncertainties are constrained in a convex polytope domain. To cope with the high-dimension problem for GRN models, the principal component plane (PCP) algorithm is proposed to construct a pruned polytope in order to use as less vertices as possible to maintain the essential information from original polytope. The so-called system equivalence transformation is developed to transform the original system into a simpler canonical form and therefore facilitate the subsequent state estimation problem. For the state estimation problem, a robust stability condition is incorporated with guaranteed performance via the semi-definite programme method, and then a new sufficient condition is derived for the desired estimators with several free slack matrices. Such a condition is vertex-dependent and therefore possesses less conservatism. It is shown, via simulation from real-world microarray time-series data, that the designed estimators have strong capability of dealing with modeling and estimation problems for short but high-dimensional gene expression time series.

Anirban Mukhopadhyay [69] proposed a novel interactive genetic algorithm-based multi objective approach that simultaneously finds the clustering solution as well as evolves the set of validity measures that are to be optimized simultaneously. The proposed method interactively takes the input from the human decision maker (DM) during execution and adaptively learns from that input to obtain the final set of validity measures along with the final clustering result. The algorithm is applied for clustering real-life benchmark gene expression datasets and its performance is compared with that of several other existing clustering algorithms to demonstrate its effectiveness. The results indicate that the proposed method outperforms the other existing algorithms for all the datasets considered here. The performance of IMOC has been demonstrated for two real-life gene expression datasets and compared with that of several other existing clustering algorithms. Results indicate that IMOC produces more biologically significant clusters compared to the other algorithms and the better result provided by IMOC is statistically significant.

Ujjwal Maulik et.al [70] proposed a novel approach to combine feature (gene) selection and transductive support vector machine (TSVM). We demonstrated that 1) potential gene markers could be identified and 2) TSVMs improved prediction accuracy as compared to the standard inductive SVMs (ISVMs). A forward greedy search algorithm based on consistency and a statistic called signal-to-noise ratio were employed to obtain the potential gene markers. The selected genes of the microarray data were then exploited to design the TSVM. Experimental results confirm the effectiveness of the proposed technique compared to the ISVM and low-density separation method in the area of semi supervised cancer classification as well as gene-marker identification.

Gui-Fang Shao [71] presented a fully automatic gridding technique to break through the limitation of traditional mathematical morphology gridding methods. First, a preprocessing algorithm was applied for noise reduction. Subsequently, the optimal threshold was gained by using the improved Otsu method to actually locate each spot. In order to diminish the error, the original gridding result was optimized according to the heuristic techniques by estimating the distribution of the spots. Intensive experiments on six different data sets indicate that our method is superior to the traditional morphology one and is robust in the presence of noise.

Xiaoxiao Xu [72] analyze the statistical performance of these arrays in imaging targets at typical low signal-to-noise ratio (SNR) levels. We compute the Ziv-Zakai bound (ZZB) on the errors in estimating the unknown parameters, including the target concentrations. We find the SNR level below which the ZZB provides a more accurate prediction of the error than the posterior Cramér-Rao bound (PCRB), through numerical examples. We further apply the ZZB to select the optimal design parameters of the microsphere array device and investigate the effects of the experimental variables such as microscope point-spread function. An imaging experiment on microspheres with protein targets verifies the optimal design parameters using the ZZB.

Pablo A. Jaskowiak [73] investigate the choice of proximity measures for the clustering of microarray data by evaluating the performance of 16 proximity measures in 52 data sets from time course and cancer experiments. This method considered six correlation coefficients, four "classical" distances, and six proximity measures specifically proposed for the clustering of gene time-course data. Given their differences, we evaluated proximity measures separately for cancer and time-course experiments. Apart from the comparison of proximity measures, we introduced a set of 17 time-course benchmark data along with a new methodology (IBSA) to evaluate distances for the clustering of genes. Both data sets and methodology can be used in future research to evaluate the effectiveness of new proximity measures in this particular scenario. IBSA can be employed to evaluate proximity measures regarding any gene clustering application, i.e., it is not restricted to gene time-course data, the scenario addressed here. Results support that measures rarely employed in the gene expression literature can provide better results than commonly employed ones, such as Pearson, Spearman, and euclidean distance. Given that different measures stood out for time course and cancer data evaluations, their choice should be specific to each scenario. To evaluate measures on time-course data, we preprocessed and compiled 17 data sets from the microarray literature in a benchmark along with a new methodology, called Intrinsic Biological Separation Ability (IBSA). Both can be employed in future research to assess the effectiveness of new measures for gene time-course data.

Cosmin Lazar [74] propose GENESHIFT, a new nonparametric batch effect removal method based on two key elements from statistics: empirical density estimation and the inner product as a distance measure between two probability density functions; second we introduce a new validation index of batch effect removal methods based on the observation that samples from two independent studies drawn from a same population should exhibit similar probability density functions. This evaluated and compared the GENESHIFT method with four other state-of-the-art methods for batch effect removal: Batch-mean centering, empirical Bayes or COMBAT, distance-weighted discrimination, and cross-platform normalization. Several validation indices providing complementary information about the efficiency of batch effect removal methods have been employed in our validation framework. The results show that none of the methods clearly outperforms the others. More than that, most of the methods used for comparison perform very well with respect to some validation indices while performing very poor with respect to others. GENESHIFT exhibits robust performances and its average rank is the highest among the average ranks of all methods used for comparison.

Telmo Amaral [75] presents a computational pipeline for automatically classifying and scoring breast cancer TMA spots that have been subjected to nuclear immunostaining. Spots are classified based on a bag of visual words approach. Immunohistochemical scoring is performed by computing spot features reflecting the proportion of epithelial nuclei that are stained and the strength of that staining. These are then mapped onto an ordinal scale used by pathologists. Multilayer perceptron classifiers are compared with latent topic models and support vector machines for spot classification and with Gaussian process ordinal regression and linear models for scoring. Intra-observer variation is also reported. The use of posterior entropy to identify uncertain cases is demonstrated. Evaluation is performed using TMA images stained for progesterone receptor.

Wenjie You et.al [76] focuses on extracting the potential structure hidden in high-dimensional multi category microarray data, and interpreting and understanding the results provided by the potential structure information. First, we propose using PLS-based recursive feature elimination (PLSRFE) in multi category problems. Then, we perform feature importance analysis based on PLSRFE for high-dimensional microarray data to determine the information feature (biomarkers) subset, which relates to the studied tumor subtypes problem. Finally, PLS-based supervised feature extraction is conducted on the selected specific genes subset to extract comprehensive features that best reflect the nature of classification to have a discriminating ability. The proposed algorithm is compared with several state-of-the-art methods using multiple high-dimensional multi category microarray datasets. Our comparison is performed in terms of recognition accuracy, relevance, and redundancy. Experimental results show that the algorithm proposed by us can improve the recognition rate and computational efficiency. Furthermore, mining potential structure information improves the interpretability and understandability of recognition results. The proposed algorithm can be effectively applied to microarray data analysis for the discovery of gene co-expression and co-regulation.

IV. Conclusions

Micro array is a ubiquitous problem that arises in a wide range of applications in computing, to full fill this we need efficient techniques. In this study we concentrate on micro array data and this article gave an overview of micro array models as well as programming tools. Micro array classification will always be a challenge for programmers. Higher-level programming models and appropriate programming tools only facilitate the process but do not make it a simple task. In this we say that, this study will help the researchers to develop the better techniques in the field of microarray.

References Références Referencias

1. K. Blekas, Member, IEEE, N. P. Galatsanos, Senior Member, IEEE, A. Likas, Senior Member, IEEE, and I. E. Lagaris, "Mixture Model Analysis of DNA Microarray Images", IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. 24, NO. 7, JULY 2005.
2. Han-Yu Chuang, Hongfang Liu , Stuart Brown, Cameron McMunn-Coffran, Cheng-Yan Kao and D. Frank Hsu, "Identifying Significant Genes from Microarray Data", Proceedings of the Fourth IEEE Symposium on Bioinformatics and Bioengineering, 2004.
3. K Blekas, Nikolas P. Galatsanos and Ioannis Georgiou, "an unsupervised artifact correction approach for the analysis of dna microarray images", IEEE, 2003.
4. Christian Uehara, Ioannis Kakadiaris , "towards automatic analysis of DNA microarrays", Proceedings of WACV, 2002.
5. Li Teng, Hongyu Li, Xuping Fu, Wenbin Chen, I-Fan Shen, "Dimension Reduction of Microarray Data Based on Local Tangent Space Alignment", IEEE, ICCI, 2005.
6. Jianhua Xuan, Eric Hoffman, Robert Clarke, Yue Wang, "Normalization of Microarray Data by Iterative Nonlinear Regression", IEEE conference on IBBE, 2005.
7. Dietmar P. F. Moeller, "Business Objects as Part of a Preprocessing based Micro Array Data Analysis", IEEE conference on EIT, 2005.
8. Qingzhong Liu, Student Member, IEEE, Andrew H. Sung, "Recursive Feature Addition for Gene Selection" IEEE conference on Nueral network, 2006.
9. Wei Peng and Tao Li, "IntClust: A Software Package for Clustering Replicated Microarray Data", IEEE conference on BIBE, 2006.
10. Yijuan Lu, Qi Tian, Feng Liu, Maribel Sanchez, and Yufeng Wang, "Interactive Semisupervised Learning for Microarray Analysis", ieee/acm transactions on computational biology and bioinformatics, vol. 4, no. 2, april-june 2007.
11. Haiying Wang, Huiru Zheng, Francisco Azuaje, "Poisson-Based Self-Organizing Feature Maps and Hierarchical Clustering for Serial Analysis of Gene Expression Data", IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) , Volume 4 Issue 2, Pages 163-175, April 2007.
12. Shahar Michal, Tor Ivry, Omer Schalit-Cohen, Moshe Sipper, and Danny Barash, "Finding a Common Motif of RNA Sequences Using Genetic Programming: The GeRNAMo System" , IEEE/ACM transactions on computational biology and bioinformatics, vol. 4, no. 4, 2007.
13. Huilin Xiong, Ya Zhang, and Xue-Wen Chen, "Data-Dependent Kernel Machines for Microarray Data Classification", IEEE/ACM transactions on computational biology and bioinformatics, VOL. 4, NO. 4, 2007.
14. Nasimul Noman and Hitoshi Iba, "Inferring Gene Regulatory Networks Using Differential Evolution

with Local Search Heuristics", IEEE/ACM transactions on computational biology and bioinformatics, vol. 4, no. 4, 2007.

15. Peng Wei and Wei Pan, "Incorporating Gene Functions into Regression Analysis of DNA-Protein Binding Data and Gene Expression Data to Construct Transcriptional Networks", IEEE/ACM transactions on computational biology and bioinformatics, vol. 5, no. 3, 2008.
16. Y. Saeys, I. Inza, and P. Larran˜ aga, "A Review of Feature Selection Techniques in Bioinformatics," Bioinformatics, vol. 23, no. 19, pp. 2507-2517, 2007.
17. P. Yang et al., "A Review of Ensemble Methods in Bioinformatics," Current Bioinformatics, vol. 5, no. 4, pp. 296-308, 2010.
18. I. Guyon, "An Introduction to Variable and Feature Selection," J. Machine Learning Research, vol. 3, pp. 1157-1182, 2003.
19. A.-C. Haury, P. Gestraud, and J.-P. Vert, "The Influence of Feature Selection Methods on Accuracy, Stability and Interpretability of Molecular Signatures," PLoS ONE, vol. 6, no. 12, p. e28210, 2011.
20. I.Guyon et al., "Gene Selection for Cancer Classification Using Support Vector Machines," Machine Learning, vol. 46, nos. 1-3, pp. 389-422, 2002.
21. T. Zhang, "On the Consistency of Feature Selection Using Greedy Least Squares Regression," J. Machine Learning Research, vol. 10, pp. 555-568, 2009.
22. S.-H. Cha, "Comprehensive Survey on Distance/Similarity Measures Between Probability Density Functions," Int'l J. Math. Models and Methods in Applied Sciences, vol. 1, no. 4, pp. 300-307, 2007.
23. D. Witten and R. Tibshirani, "A Comparison of Fold-Change and the t-Statistic for Microarray Data Analysis," technical report, Stanford Univ., 2007.
24. E. Parzen, "On Estimation of a Probability Density Function and Mode," The Annals of Math. Statistics, vol. 33, no. 3, pp. 1065-1076, 1962.
25. A. Wilinski, S. Osowski, and K. Siwek, "Gene Selection for Cancer Classification through Ensemble of Methods," Proc. Ninth Int'l Conf. Adaptive and Natural Computing Algorithms (ICANNGA '09), pp. 507-516, 2009.
26. X. Yan et al., "Detecting Differentially Expressed Genes by Relative Entropy," J. Theoretical Biology, vol. 234, no. 3, pp. 395- 402, 2005.
27. J.-G. Zhang and H.-W. Deng, "Gene Selection for Classification of Microarray Data Based on the Bayes Error," BMC Bioinformatics, vol. 8, no. 1, article 370, 2007.
28. X. Liu, A. Krishnan, and A. Mondry, "An Entropy-Based Gene Selection Method for Cancer Classification Using Microarray Data," BMC Bioinformatics, vol. 6, article 76, 2005.
29. S. Parodi, V. Pistoia, and M. Muselli, "Not Proper Roc Curves as New Tool for the Analysis of Differentially Expressed Genes in Microarray Experiments," BMC Bioinformatics, vol. 9, no. 1, article 410, 2008.
30. S. Dudoit, J. Fridlyand, and T.P. Speed, "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data," J. Am. Statistical Assoc., vol. 97, no. 457, pp. 77-87, 2002.
31. A. Ben-Dor et al., "Tissue Classification with Gene Expression Profiles," J. Computational Biology, vol. 7, pp. 559-583, 2000.
32. J. Cohen, "The Earth is Round (p < .05)," Am. Psychologist, vol. 38, pp. 997-1003, 1994.
33. W. Pan, J. Lin, and C.T. Le, "A Mixture Model Approach to Detecting Differentially Expressed Genes with Microarray Data," Functional and Integrative Genomics, vol. 3, no. 3, pp. 117-124, 2003.
34. S. Dudoit, J.P. Shaffer, and J.C. Boldrick, "Multiple Hypothesis Testing in Microarray Experiments," Statistical Science, vol. 18, no. 1, pp. 71-103, 2003.
35. J.G. Thomas et al., "An Efficient and Robust Statistical Modeling Approach to Discover Differentially Expressed Genes Using Genomic Expression Profiles," Genome Research, vol. 11, no. 7, pp. 1227-1236, 2001.
36. Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," J. Royal Statistical Soc. Series B (Methodological), vol. 57, no. 1, pp. 289-300, 1995.
37. D. Storey, "The Positive False Discovery Rate: A Bayesian Interpretation and the q-Value," Annals of Statistics, vol. 31, pp. 2013-2035, 2003.
38. J.D. Storey, "A Direct Approach to False Discovery Rates," J. Royal Statistics Soc.: Series B, vol. 64, no. 3, pp. 479-498, 2002.
39. T. Bø and I. Jonassen, "New Feature Subset Selection Procedures for Classification of Expression Profiles," Genome Biology, vol. 4, no. 4, pp. research0017.1-research0017.11, 2002.
40. Leila Muresan, Jarosław Jacak, Erich Peter Klement, Jan Hesse, and Gerhard J. Schutz, "Microarray Analysis at Single-Molecule Resolution", IEEE TRANSACTIONS ON NANOBIOSCIENCE, VOL. 9, NO. 1, MARCH 2010.
41. Yoshinori Tamada, Seiya Imoto, Hiromitsu Araki, Masao Nagasaki, Cristin Print, D. Stephen Charnock-Jones, and Satoru Miyano, "Estimating Genome-Wide Gene Networks Using Nonparametric Bayesian Network Models on Massively Parallel Computers", IEEE TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMA- TICS, 2010.

42. Tianwei Yu, Hesen Peng, Wei Sun, "Incorporating nonlinear relationships in microarray missing value imputation",IEEE TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS,2010.
43. Jianxing Feng, Rui Jiang, and Tao Jiang, "A Max-Flow Based Approach to the Identification of Protein Complexes Using Protein Interaction and Microarray Data",IEEE TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, 2010.
44. Jong Kyoung Kim and Seungjin Choi, Member, IEEE, " Probabilistic Models for Semi-Supervised Discriminative Motif Discovery in DNA Sequences", IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, FEBRUARY 2, 2010.
45. Xin ZHAO and Leo Wang-Kit CHEUNG, "Multi-Class Kernel-Imbedded Gaussian Processes for Microarray Data Analysis", IEEE TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, 2010.
46. Alfredo Benso, IEEE Senior Member, Stefano Di Carlo, IEEE Member and Gianfranco Politano, "A cDNA Microarray Gene Expression Data Classifier for Clinical Diagnostics based on Graph Theory", IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, 2010.
47. Yu-Cheng Liu, Chao-Hui Lee, Wei-Chung Chen, J. W. Shin, Hui-Huang Hsu and Vincent S. Tseng, "A Novel Method for Mining Temporally Dependent Association Rules in Three-Dimensional Microarray Datasets", IEEE , 2010.
48. Hong-Dong Li, Yi-Zeng Liang, Qing-Song Xu, Dong-Sheng Cao, Bin-Bin Tan, Bai-Chuan Deng, Chen-Chen Lin, "Recipe for Uncovering Predictive Genes using Support Vector Machines based on Model Population Analysis", IEEE TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, 2010.
49. Yang Chen, and Jinglu Hu," Accurate Reconstruction for DNA Sequencing by Hybridization Based on A Constructive Heuristic", IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, 2010.
50. Jong Kyoung Kim and Seungjin Choi, Member, IEEE, " Probabilistic Models for Semi-Supervised Discriminative Motif Discovery in DNA Sequences", IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, FEBRUARY 2, 2010.
51. Marco Muselli, Member, IEEE, Alberto Bertoni, Marco Frasca, Alessandro Beghini, Francesca Ruffino, and Giorgio Valentini, "A mathematical model for the validation of gene selection methods", IEEE ACM TRANS. ON COMP. BIOL. AND BIOINFORMATICS, 2010.
52. Leila Muresan, Jarosław Jacak, Erich Peter Klement, Jan Hesse, and Gerhard J. Schutz, "Microarray Analysis at Single-Molecule Resolution", IEEE TRANSACTIONS ON NANOBIOSCIENCE, VOL. 9, NO. 1, MARCH 2010.
53. Banu Dost, Chunlei Wu, Andrew Su, Vineet Bafna, "TCLUST: A fast method for clusterin genome-scale expression data", IEEE TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, 2010.
54. Giorgio Valentini, "True Path Rule hierarchical ensembles for genome-wide gene function prediction", IEEE ACM TRANS. ON COMP. BIOL. AND BIOINFORMATICS, 2010.
55. Chandra Sekhar, V., Allam Appa Rao, and P. Srinivasa Rao. "Differential Gene Expression Analysis for Diabetes with and without parental history." InComputer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on, vol. 9, pp. 322-326. IEEE, 2010.
56. Sekhar, V. Chandra, Allam Appa Rao, P. S. Rao, and K. Srinivas. "Identification of differentially expressed genes for diabetes with parental history vs healthy using Microarray data analysis." In Advanced Computer Theory and Engineering (ICACTE), 2010 3rd International Conference on, vol. 4, pp. V4-496. IEEE, 2010.
57. Shah, Mohak, and Jacques Corbeil. "A general framework for analyzing data from two short time-series microarray experiments." Computational Biology and Bioinformatics, IEEE/ACM Transactions on 8, no. 1 (2011): 14-26.
58. Zhao, Xin, and Leo Wang-Kit Cheung. "Multiclass Kernel-Imbedded Gaussian Processes for Microarray Data Analysis." IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 8, no. 4 (2011): 1041-1053.
59. Sakellariou, Argiris, Despina Sanoudou, and George Spyrou. "Investigating the minimum required number of genes for the classification of neuromuscular disease microarray data." Information Technology in Biomedicine, IEEE Transactions on 15, no. 3 (2011): 349-355.
60. Vasamsetty, Chandra Sekhar, Srinivasa Rao Peri, Allam Appa Rao, K. Srinivas, and Chinta Someswararao. "Gene Expression Analysis for Type-2 Diabetes Mellitus--A Study on Diabetes With And Without Parental History." *Journal of Theoretical & Applied Information Technology* 27, no. 1 (2011).
61. Vasamsetty, Chandra Sekhar, Srinivasa Rao Peri, Allam Appa Rao, K. Srinivas, and Chinta Someswararao. "Gene Expression Analysis for Type-2 Diabetes Mellitus–A Case Study on Healthy vs Diabetes with Parental History." IACSIT International Journal of Engineering and Technology, Vol.3, No.3, pp.310-314, 2011.
62. Maji, Pradipta. "Mutual information-based supervised attribute clustering for microarray sample classification." Knowledge and Data

Engineering, IEEE Transactions on 24, no. 1 (2012): 127-140.

63. ElBakry, Ola, M. Omair Ahmad, and M. N. S. Swamy. "Identification of Differentially Expressed Genes for Time-Course Microarray Data Based on Modified RM ANOVA." Computational Biology and Bioinformatics, IEEE/ACM Transactions on 9, no. 2 (2012): 451-466.
64. Sharma, Alok, Seiya Imoto, and Satoru Miyano. "A top-r feature selection algorithm for microarray gene expression data." IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 9, no. 3 (2012): 754-764.
65. Janowczyk, Andrew, Sharat Chandran, Rajendra Singh, Dimitra Sasaroli, George Coukos, Michael D. Feldman, and Anant Madabhushi. "High-throughput biomarker segmentation on ovarian cancer tissue microarrays via hierarchical normalized cuts." Biomedical Engineering, IEEE Transactions on 59, no. 5 (2012): 1240-1252.
66. Hanczar, Blaise, and Avner Bar-Hen. "A new measure of classifier performance for gene expression data." IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 9, no. 5 (2012): 1379-1386.
67. Maji, Pradipta. "Mutual information-based supervised attribute clustering for microarray sample classification." Knowledge and Data Engineering, IEEE Transactions on 24, no. 1 (2012): 127-140.
68. Wang, Zidong, Huihai Wu, Jinling Liang, Jie Cao, and Xiaohui Liu. "On modeling and state estimation for genetic regulatory networks with polytopic uncertainties." NanoBioscience, IEEE Transactions on 12, no. 1 (2013): 13-20.
69. Mukhopadhyay, Anirban, Ujjwal Maulik, and Sanghamitra Bandyopadhyay. "An interactive approach to multiobjective clustering of gene expression patterns."Biomedical Engineering, IEEE Transactions on 60, no. 1 (2013): 35-41.
70. Maulik, Ujjwal, Anirban Mukhopadhyay, and Debasis Chakraborty. "gene-expression-based cancer subtypes prediction through feature selection and transductive SVM." Biomedical Engineering, IEEE Transactions on 60, no. 4 (2013): 1111-1117.
71. Shao, Gui-Fang, Fan Yang, Qian Zhang, Qi-Feng Zhou, and Lin-Kai Luo. "Using the Maximum Between-Class Variance for Automatic Gridding of cDNA Microarray Images." Computational Biology and Bioinformatics, IEEE/ACM Transactions on 10, no. 1 (2013): 181-192.
72. Xu, Xiaoxiao, Pinaki Sarder, Nalinikanth Kotagiri, Samuel Achilefu, and Arye Nehorai. "Performance analysis and design of position-encoded microsphere arrays using the Ziv-Zakai bound." NanoBioscience, IEEE Transactions on 12, no. 1 (2013): 29-40.
73. Jaskowiak, Pablo A., Ricardo JGB Campello, and Ivan G. Costa Filho. "Proximity measures for clustering gene expression microarray data: a validation methodology and a comparative analysis." IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 10, no. 4 (2013): 845-857.
74. Lazar, Cosmin, Jonatan Taminau, Stijn Meganck, David Steenhoff, Alain Coletta, David Y. Weiss Solis, Colin Molter, Robin Duque, Hugues Bersini, and Ann Nowé. "GENESHIFT: A Nonparametric Approach for Integrating Microarray Gene Expression Data Based on the Inner Product as a Distance Measure between the Distributions of Genes." IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 10, no. 2 (2013): 383-392.
75. Amaral, Telmo, Stephen J. McKenna, Katherine Robertson, and Alastair Thompson. "Classification and Immunohistochemical Scoring of Breast Tissue Microarray Spots." (2013): 1-1.
76. You, Wenjie, Zijiang Yang, Mingshun Yuan, and Guoli Ji. "TotalPLS: Local Dimension Reduction for Multicategory Microarray Data." 1-14.

This page is intentionally left blank

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 14 Issue 3 Version 1.0 Year 2014
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Maximising the Value of Missing Data

By Dr. Atai Winkler

Abstract- The subject of missing values in databases and how to handle them has received very little attention in the statistics and data mining literature1, 2, 3 and even less, if any at all, in the marketing literature. The usual attitude of practitioners is 'we'll just have to ignore records with missing values'. On the other hand, a few very advanced theoretical solutions have been developed, some of which have been applied, particularly to clinical trials data. These solutions can only be applied to small databases, not to the very large databases held by many companies on their customers. This paper describes a new method for imputing missing values in such very large databases. Two particular features of the method are that it can handle all combinations of variable type (continuous, ordinal and categorical) and that all the missing values in the database are imputed in one run of the software. It is based on the k-nearest neighbours method, a well known method in data mining. The paper concludes by presenting the results of a study of this method when used to impute the missing values in a real set of data.

This paper is only concerned with 'missing' data, i.e. data that are not known but which have real values. It does not address the problem of 'empty' data, i.e. data that are not known but which cannot have real values.

Keywords : *missing data; imputation; gaps; holes; data mining; empty data.*

GJCST-C Classification : *H.2.8*

Maximising the Value of Missing Data

Dr. Atai Winkler

Abstract- The subject of missing values in databases and how to handle them has received very little attention in the statistics and data mining literature1, 2, 3 and even less, if any at all, in the marketing literature. The usual attitude of practitioners is 'we'll just have to ignore records with missing values'. On the other hand, a few very advanced theoretical solutions have been developed, some of which have been applied, particularly to clinical trials data. These solutions can only be applied to small databases, not to the very large databases held by many companies on their customers. This paper describes a new method for imputing missing values in such very large databases. Two particular features of the method are that it can handle all combinations of variable type (continuous, ordinal and categorical) and that all the missing values in the database are imputed in one run of the software. It is based on the *k*-nearest neighbours method, a well known method in data mining. The paper concludes by presenting the results of a study of this method when used to impute the missing values in a real set of data.

This paper is only concerned with 'missing' data, i.e. data that are not known but which have real values. It does not address the problem of 'empty' data, i.e. data that are not known but which cannot have real values.

Keywords: missing data; imputation; gaps; holes; data mining; empty data.

I. Introduction

Marketers are always striving to develop effective marketing campaigns by maximising the benefits they can achieve from the data they hold in their databases. The results of different campaigns can then be assessed by comparing the return on investment of the campaigns. Whatever the nature and aims of the campaigns, they always start with some form of analysis to gain customer insight, and the results of this analysis lead to segmentation and targeting of potential customers.

As companies move through the cycle of customer acquisition to customer retention, the ability to analyse the data they hold on their customers becomes increasingly important. However, this task is often hampered by the fact that the vast majority of databases have missing information, sometimes called gaps or holes. This may be for historical reasons where the emphasis was on customer acquisition or it may be due to changes in the needs of the organisation over time. Whatever the cause, missing data are a very common and serious problem - it is not uncommon for collected lifestyle data to have 30–40% of their values missing. This is a real problem when the data are to be analysed - if the data are not there, they cannot be analysed either at record level or for the database overall. These missing values mean that the database is not as large or as rich in information as may be assumed from its overall size (number of fields and number of records). Thus, the effect of the missing data is to limit the amount and quality of new information that can be learnt from the database.

With respect to customer retention programmes and CRM, missing customer data will have a serious adverse impact on the outcome of the campaigns because they will be based on incomplete data and therefore weak analysis. The consequence of this is that the segmentation and targeting will be less accurate than they would be if the database were fully populated. This problem of missing data has also heavily affected the lifestyle data sellers who are trying to present a more complete picture of the UK adult population. Since less data are acquired now than formerly directly from individuals about their demographics as consumers, there is a greater reliance on modelled data. Unfortunately, these modelled data are often obtained from incomplete data. This just compounds the problem of missing data because any bias in the available data manifests itself, and is probably increased, in the modelled data. Thus, a vicious circle of partial data being used to obtain more partial data is started, and after this process has been repeated many times it is likely that the final data bear little resemblance to the true UK population.

Since the problem of missing values in both customer and lifestyle databases is widespread and getting worse, methods that give more accurate estimates of the missing values compared to those obtained using currently available methods have an important and significant contribution to play in improving marketing effectiveness. The desired result of all these methods is a fully populated database with all the missing values replaced by estimates of their true values - a process known as (missing value) imputation. The imputed values should be 'good' and plausible estimates of the true values so that the statistical properties of the fully and partially populated databases are as similar as possible.

II. Missing Data and Empty Data

It may be thought that all missing data can and should be imputed. This is not always the case because some data may be missing because they

Author : e-mail: atai.winkler@win-tech.co.uk

cannot have real values - such values must remain missing. Thus, it is important to understand the difference between 'missing' data and 'empty' data. A value that is not known but which has a real value is a 'missing' value. A value that is not known but which cannot have a real value is an 'empty' value. Therefore, missing data can be imputed but empty data cannot be imputed.

III. Imputation Methods

A number of imputation methods are available, and some of the factors that help determine a suitable method are:

- Is the method only suitable for small databases or can it be used on small and large databases?
- Is the method 'local' or 'global'? In a local method the missing values in each record are imputed at record level but in a global approach the missing values in each record are imputed from the summary statistics of a group of records. This is analogous to the difference between a micro and macro approach.
- Is the method simple or complicated?
- Is the method very sensitive to assumptions about the data or can any assumptions be relaxed a little?
- Is the method very demanding in terms of time and cost or does it require few resources?

a) Desirable Properties of Imputation Methods

Some desirable properties of a 'good' and practical imputation method are:

- It is easy to use and requires minimum user-intervention.
- It should only use the information in the database (no external data are required).
- It is very amenable for use on very large databases.
- All the missing values in all the fields are imputed in one run.
- The order of the records does not affect the imputed values.
- It uses all the known information in each record to impute the missing values in that record.
- The missing values in correlated variables can be imputed.
- The heterogeneity of the database is maintained.
- All the imputed values are plausible.
- Variables in different units are allowed.
- It allows all combinations of continuous, ordinal and text variables.
- The size of the database does not affect the search and imputation methods.
- Possible imputation methods include:
- case deletion
- mean or mode substitution
- cold deck substitution
- hot deck substitution
- regression
- EM algorithm
- structural models
- *k*-nearest neighbours

b) Case Deletion

Case deletion avoids rather than solves the problem of missing values because it ignores all the incomplete records. Very often it is the default method of handing missing data. Although very easy to implement, two immediate and severe disadvantages of the method are firstly that a very large proportion of the records may be ignored and secondly that the remaining records may not be representative of the population. The commercial and financial implications of this bias in the data that are actually analysed are easy to imagine.

c) Mean Substitution and Cold Deck Imputation

Mean substitution and cold deck imputation are two frequently used imputation methods. Mean substitution involves replacing all the missing values in each field by the field's mean or mode as appropriate, and in cold deck imputation the missing values are replaced by external constants, one for each field. These methods are easy and quick to implement but being global methods they are very unlikely to maintain the statistical properties of the database. In the case of mean substitution, the mean (mode) values of the fields in the partially and fully populated databases are, by definition, the same, but the variation of each field in the fully populated database is much smaller than the corresponding variation in the partially populated database. The result is that the records are not as clearly differentiated as they should be and so it is harder to understand how people's individual characteristics determine their actions and behaviour from a database which has been fully populated in either of these ways.

Another major problem with mean substitution and cold deck imputation is that unrealistic or even impossible values can be easily imputed. This is because the value imputed in any one field is the mean of the known values in that field. Therefore, if a database contains people across a wide range of age, income and lifestyle attributes and the data can be segmented into a finite number of homogeneous clusters with high inter-cluster heterogeneity, the mean value of any field across all clusters does not have meaning or significance for any single cluster or all the clusters. Therefore, using values imputed in this way as the basis for marketing campaigns and other commercial activities may not yield the desired outcomes because the targeting and segmentation are based on poor quality dat.

d) Hot Deck Imputation

A slightly more advanced method of imputation is hot deck imputation. This is similar to cold deck imputation except that the missing values are replaced by values from 'similar' records in the database. These similar records are obtained by clustering the complete records and then assigning a cluster to each incomplete record. The missing values in each incomplete record are replaced by values calculated from its associated cluster. Like mean substitution and cold deck imputation, hot deck imputation is a global method.

e) Regression

In regression imputation the missing values are replaced by values calculated from a regression equation, for example

$$y = a + bx_1 + cx_2 \tag{1}$$

y is the variable to be imputed, and $x1$ and $x2$ are other variables (a , b and c are known constants).

Implicit in using (1) is that the values of the variables on the right hand side of it ($x1$ and $x2$) in records whose values of y are to be imputed are known. This problem can be overcome by developing the models only from complete records - but this raises a fundamental problem, namely what happens if the complete records are either a small proportion of the database or they are a distinct group in the database rather than being a fair reflection of the database as a whole? On the other hand regression imputation is a local method because the missing values in each record are calculated from the data in that record - a significant advantage. Notwithstanding this advantage, regression imputation has a number of practical and theoretical problems, including:

- since a regression equation must be developed for each variable with missing values, regression imputation is very time consuming, especially in large databases and in databases many of whose fields have missing values;
- working out the equations may be difficult, not least because the correlations between the variables may be weak;
- different relationships may exist for different homogeneous groups in the database and so trying to find one relationship across all groups will yield an unsatisfactory compromise that is not an accurate portrayal of the relationship in any one group - the single equation will predict values that do not reflect any individual's unique characteristics, and so the same problems as those associated with mean substitution, namely reduction in the heterogeneity of the database, may arise;
- the relationships between the variables are artificially and falsely inflated because the missing values are estimated by substituting into the regression equations.

f) EM Algorithm and Structural Models

These methods are very advanced and demanding in terms of the time and expertise required. They are not amenable for use on large databases.

g) K-Nearest Neighbours and Imputation

k-nearest neighbours is a data mining method used in estimation and classification problems. Unlike many other methods used in statistical data analysis and modelling, it does not require a model to be developed for each field. Rather, it is based on the simple concept that the (statistical) similarity between two records is calculated from the multivariate distance between them. If two records are similar, i.e. their corresponding fields have similar values, they will be close to one another and so the distance between them will be 'small' when their common known data are plotted. These records are more similar to one another than are other records with larger distances between them. This geometric way in which the most similar records are found explains why the method is called *k*-nearest neighbours. Thus, the method involves mapping all the data into multi-dimensional space and then calculating the distances between all pairs of records (each dimension is a variable).

The method works by firstly finding a pool of donors, i.e. complete records, for each recipient, i.e. incomplete record. It then uses the values in the donors of each field that is missing in the recipient to impute the missing data in the recipient. There are three stages to the method.

For each incomplete record:

1. Search the entire database for similar complete records using the values in the selected fields in the incomplete record.
2. Rank the complete records by distance to the incomplete record.
3. Use the specified number of complete records in the ranked set to impute the missing values in the incomplete record.

By using all the known data in each recipient to search for its most similar donors and then using these donors to impute the recipient's missing values, the statistical properties of the partially populated database are maintained. This process of searching for similar donors and then using them to impute the missing values is repeated for each recipient. This recipient-by-recipient approach means that each recipient has its own donors, i.e. Nearest Neighbours (NNs), from which its missing values are imputed. It is this feature of *k*-nearest neighbours that helps maintain the heterogeneity of the database.

This very localised approach to imputation is in marked contrast to global methods where the imputation is based on groups of recipients and each group has the same donors and therefore the same imputed values. Thus, the variation of the variables in a database which has been fully populated using a global method is lower than it is in a database which has been fully populated using *k*-nearest neighbours. Furthermore, since each recipient record is treated individually, the method obtains the most accurate imputed values for each recipient record rather than attempting to obtain the most accurate average imputed values across a group of records.

The main reason for the limited use up to now of the *k*-nearest neighbours method with very large databases is that the number of distances that have to be stored and then ranked made it impractical to use on such databases. This problem has now been overcome so that it does not store the distance from the incomplete record being processed to each complete record, rank all the distances and then select the specified number of NNs. This means that only a fraction of the number of complete records in the database are stored at any one time.

A good example of the type of data that can be imputed using *k*-nearest neighbours is lifestyle data. However, variables such as ownership of pets, type of credit card owned and participation in hobbies, for example stamp collecting, should not be imputed because they are generally independent of other variables and do not define people.

To show how *k*-nearest neighbours works, consider the data in Table 1. The data come from a survey and the fields are:

mar_stat: marital status: D (divorced); M (married); S (single)

res_stat: residential status: P (owner-occupier); T (rent alone); Z (multiple rent)

age: age (months)

bank: time with bank (months)

cheq_card: own a cheque guarantee card: N (no); Y (yes)

add: time at current address (months)

emp: time with current employer (months)

occup: occupation code

Table 1

Record No.	Mar Stat	Res Stat	Age (mths)	Bank (mths)	Cheq Card	Add (mths)	Emp (mths)	Occup
1	M	T	334	18	Y	20	12	ES
2	M	T	308	24	Y	24	66	ES
3	D		317		N	36		EO
4	M	T	271	60	N	36	60	EM
5	M			132	Y		0	
6	D	T	516	72	N	6	11	S
7	S	P	314	14	N	54	42	EB
8			338	12			66	SB
9	M	Z	448	126	Y	82	120	EP
10			749		Y	12		

This small database has 10 records of which 4 (records 3, 5, 8 and 10) have missing values and the other 6 are complete. The three NNs for each incomplete record were calculated using *k*-nearest neighbours and are shown in Table 2.

Table 2

Recipient Record	Nearest Neighbour	Next Nearest Neighbour	Next Nearest Neighbour
3	4	7	2
5	1	2	4
8	2	7	4
10	9	1	2

The table shows that the three NNs for record 3 are records 4, 7 and 2, with record 4 being the most similar and record 2 the least similar of the three NNs. Since there are six complete records in the database, there can be up to six NNs.

If only one donor (*the* NN) is to be used, the missing values in records 3, 5, 8 and 10 are copied directly from the corresponding fields in records 4, 1, 2 and 9 respectively. If two NNs are to be used, the missing values in records 3, 5, 8 and 10 are calculated from the corresponding fields in records 4 and 7, 1 and 2, 2 and 7, and 9 and 1 respectively. If three NNs are to be used, the missing values in record 3 are calculated from the corresponding fields in records 4, 7 and 2, record 5 from records 1, 2 and 4, record 8 from records 2, 7 and 4, and record 10 from records 9, 1 and 2.

The fully populated database shown in Table 3 was obtained by using the three NNs shown in Table 2.

Table 3

Record No.	Mar Stat	Res Stat	Age (mths)	Bank (mths)	Cheq Card	Add (mths)	Emp (mths)	Occup
1	M	T	334	18	Y	20	12	ES
2	M	T	308	24	Y	24	66	ES
3	D	T	317	33	N	36	56	EO
4	M	T	271	60	N	36	60	EM
5	M	T	304	132	Y	27	0	ES
6	D	T	516	72	N	6	11	S
7	S	P	314	14	N	54	42	EB
8	M	T	338	12	N	38	66	SB
9	M	Z	448	126	Y	82	120	EP
10	M	T	749	56	Y	12	66	ES

As with many methods of data analysis, the trade-off between run-time and accuracy must be considered - a big increase in run-time is not always accompanied by a significant improvement in accuracy. For *k*-nearest neighbours an obvious question is how the number of donors affects the accuracy of the imputed values. It is reasonable to assume that as the number of donors increases, the 'better' will be the imputed values. However, increasing the number of donors has two adverse effects: firstly, as the distance between the recipient and the donors increases, the donors become more dissimilar from the recipient; and secondly, the run-time increases. This and other questions are discussed later in this paper.

Example of Using k-Nearest Neighbours For Imputation:

The use of the *k*-nearest neighbours method for imputation was tested on a database of 20,000 records, of which 13,303 (66.5%) were fully populated and the remaining 6,697 (33.5%) had at least one missing value. The actual values of these missing values were known so that after the imputation the imputed values could be compared with the actual values. The results of this comparison are presented. The imputation and validation were carried out on a Dell Precision 650 Workstation with 1 Xeon 3.2GHz processor and 3Gb RAM.

The distribution of missing values in the 6,697 records is shown in Table 4.

Table 4

No. Missing	No. Records	Cumulative %
1	2,947	44.00
2	2,442	80.47
3	1,026	95.79
4	238	99.34
5	40	99.94
6	3	99.99
7	1	100.00
8	0	100.00

The table had 8 fields, as described in Table 5.

Table 5

Field	Description	Type
p_hhld	head of household	text
h_comp	household composition	text
h_shrs	value of shares held	text
h_prop	type of property	text
h_inc	household income	ordinal
h_res	residence type	text
h_ten	household tenure	text
age	age	continuous

Four runs were carried out. The settings of the runs are shown in Table 6.

Table 6

Run No.	Sampling Percentage	No. of Complete Records Sampled	No. of Nearest Neighbours
1	10	1,300	10
2	10	1,300	20
3	5	665	10
4	5	665	1

h) Imputation Results

The results of the imputation for the text variables are shown in Table 7.

Table 7

Field	Null Count	No. of Categrs.	% Correctly Imp. (Run 1)	% Correctly Imp. (Run 2)	%Correctly Imp. (Run 3)	% Correctly Imp. (Run 4)
p_hhld	991	2	90.62	90.41	90.31	88.50
h_comp	1,090	12	37.43	35.23	33.94	30.83
h_shrs	1,090	3	76.97	77.34	76.79	75.32
h_prop	1,189	5	83.52	83.35	82.51	79.14
h_res	1,255	5	44.94	45.10	45.82	40.24
h_ten	1,189	3	68.04	66.69	68.29	60.64

Age was the only continuous field and its null count was 4,027. The results of the four runs are shown in Table 8. (ME stands for Mean Error and MAE stands for Mean Absolute Error.)

Table 8

Run No.	ME	MAE
1	0.64	12.69
2	1.15	12.82
3	1.07	12.90
4	0.71	14.21

Income was the only ordinal field. It was divided into 10 levels, and its null count was 1,255. The results of the four runs are shown in Table 9 as the percentage of entries in cells off the Leading Diagonal (LD) in the cross classification matrix (this is just a crosstab of actual values against imputed values).

Table 9

Run No.	% Corr Imputed	% 1 off LD	% 2 off LD	% 3 off LD	% 4 off LD	% 5 off LD	% 6 off LD	% 7 off LD	% ≥8 off LD
1	17.53	32.91	25.02	14.02	6.93	2.63	0.72	0.24	
2	17.05	31.39	24.86	16.10	7.09	2.71	0.80		
3	15.62	31.08	23.59	16.02	8.84	3.82	0.80	0.23	
4	17.13	29.72	20.88	15.30	9.24	4.46	2.07	0.88	0.32

Table 10 shows the run-times of the four runs.

Table 10

Run No.	Run-Time (Mins)
1	47
2	48
3	24
4	24

IV. Discussion of Results

Table 7 shows that the percentage of correctly imputed text variables is not very sensitive to the sampling percentage or to the number of NNs. This is because the records are randomly distributed in the database rather than there being a number of distinct and homogeneous groups in the database each of which is concentrated in adjoining records, and the variation of the fields in the NNs hardly increases as the number of NNs increases.

It is expected that as the number of NNs increases the variation of the fields in the NNs would also increase because the (complete) records in the search domains are ranked by similarity to the incomplete record. This means that the record most recently added to the search domain is always most similar to the previously added record and least similar to the first record in the search domain. If the variation of the fields in the NNs were much greater, the success of the imputation would be much more sensitive to the number of NNs. In general, the rate at which the variation of the fields in the NNs increases depends on the data and the number of NNs.

Another interesting result in Table 7 is the variation of the percentage of values correctly imputed across the fields; for example in run 1 it is between 37.43 and 90.62. Now, it is reasonable to assume that as the number of categories of a text field increases, the percentage of values correctly imputed would decrease (all other things being equal). However, there is another more important factor at play, and that is the standard deviation of the relative frequencies of the categories. Table 11 shows this standard deviation, the rank of the standard deviations, the rank of the percentage of values correctly imputed for any of the four runs in Table 7 and the rank of the number of categories for all the text fields (1 is the smallest rank and 6 is the largest rank).

Table 11

Field	St. Dev	Rank of St. Dev	Rank of % Corr. Imp.	Rank of No. Categories
p_hhld	57.13	6	6	1
h_comp	9.46	1	1	6
h_shrs	39.78	5	4	2.5
h_prop	34.31	4	5	4.5
h_res	11.47	2	2	4.5
h_ten	28.49	3	3	2.5

There is an almost perfect 1 to 1 correspondence between the third and fourth columns in Table 11.

The only discrepancy occurs with the ranks of h_shrs and h_prop - their standard deviations are more similar to one another than are other pairs of standard deviations. This suggests that the variation in the relative frequencies of the categories of the fields is a significant factor in determining the percentage of values correctly imputed. However, the fourth and fifth columns in Table 11 appear to follow a weak inverse relationship - the larger the number of categories a text field has, the smaller is its percentage of values correctly imputed likely to be.

The results in Table 8 show that for none of the runs is the absolute value of the mean error equal to the mean absolute error. This is because the values of age in the NNs are not all positively or all negatively biased, i.e. each set of NNs has values of age both above and below the true values. Once again, this shows that the records are randomly distributed in the database.

The results in Table 9 appear to suggest that the imputed ordinal values are biased because the values in the % 1 off LD, % 2 off LD and % 3 off LD columns are mostly greater than the corresponding values in the % Correctly Imputed column. This impression is incorrect because the entries in each column are obtained by adding a different number of values. Table 12 shows the number of values used to calculate the entries in each position in Table 9. Thus, for run 1 17.53 was obtained by adding 10 numbers, 32.91 was obtained by adding 18 numbers and 25.02 was obtained by adding 16 numbers. One approximate way of determining if the imputed values are biased is to normalise each entry in Table 9 by dividing it by the number of values for that position as shown in Table 12.

Table 12

Position	No. of Values
LD	10
1 off LD	18
2 off LD	16
3 off LD	14
4 off LD	12
5 off LD	10
6 off LD	8
7 off LD	6
8 off LD	4
9 off LD	2

Table 9 shows that for all the runs about half the values were imputed correctly or within one level either side of the LD, and that the percentages fall off very quickly as the cells move away from the LD.

Comparing Tables 6 and 10, it is immediately apparent that the run-time is strongly influenced by the number of complete records from which the imputed values are calculated and not affected at all by the number of NNs. This is as expected because the number of distances calculated is given by the product of the number of incomplete records and the number of complete records used. The number of NNs does not determine the run-time because the software used in this study has a very powerful sorting algorithm which sorts the distances as they are processed. This means that the number of distances actually stored as each incomplete record is processed is kept to a minimum and is very close to or equal to the number of NNs specified at input. The alternative solution to this sorting problem is to store the distance for each complete record and then when all the complete records have been processed rank all the distances from smallest to largest. This approach has two big computational problems: firstly, the larger the database the more distances have to be stored; and secondly the time required to sort these distances is not insignificant, and indeed may be greater than the time required to calculate the distances and impute the missing values.

V. Conclusion

This paper has presented the results of a study on how a powerful data mining technique, *k*-nearest neighbours, has been enhanced and then applied to the ever-present problem of how to impute missing values in large databases. The enhancements make the method amenable for use on very large commercial databases. The results of a study presented in this paper show that its overall accuracy is very high. The advantage of having more accurate imputed data is that campaigns can be better targeted. In turn, this will generate higher response rates and a greater return on investment.

References Références Referencias

1. Manly B. (1994) ‘Multivariate Statistical Methods, A Primer’, Chapman and Hall, 0-412-60300-4.
2. G. E. A. P. A. Batista, M. C. Monard, ‘K-Nearest Neighbour as Imputation Method: Experimental Results’. Technical report ICMC-USP (University of Sao Paulo), (2002), ISSN–0103-2569.
3. G. E. A. P. A. Batista, M. C. Monard, ‘An Analysis of Four Missing Data Treatment Methods For Supervised Learning’. Applied Artificial Intelligence, Vol. 17 (5-6), pages 519-533 (2003).

Global Journals Inc. (US) Guidelines Handbook 2014

www.GlobalJournals.org

FELLOW OF ASSOCIATION OF RESEARCH SOCIETY IN COMPUTING (FARSC)

Global Journals Incorporate (USA) is accredited by Open Association of Research Society (OARS), U.S.A and in turn, awards "FARSC" title to individuals. The 'FARSC' title is accorded to a selected professional after the approval of the Editor-in-Chief/Editorial Board Members/Dean.

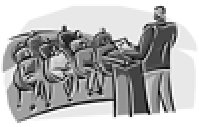
> The "FARSC" is a dignified title which is accorded to a person's name viz. Dr. John E. Hall, Ph.D., FARSC or William Walldroff, M.S., FARSC.

FARSC accrediting is an honor. It authenticates your research activities. After recognition as FARSC, you can add 'FARSC' title with your name as you use this recognition as additional suffix to your status. This will definitely enhance and add more value and repute to your name. You may use it on your professional Counseling Materials such as CV, Resume, and Visiting Card etc.

The following benefits can be availed by you only for next three years from the date of certification:

FARSC designated members are entitled to avail a 40% discount while publishing their research papers (of a single author) with Global Journals Incorporation (USA), if the same is accepted by Editorial Board/Peer Reviewers. If you are a main author or co-author in case of multiple authors, you will be entitled to avail discount of 10%.

Once FARSC title is accorded, the Fellow is authorized to organize a symposium/seminar/conference on behalf of Global Journal Incorporation (USA).The Fellow can also participate in conference/seminar/symposium organized by another institution as representative of Global Journal. In both the cases, it is mandatory for him to discuss with us and obtain our consent.

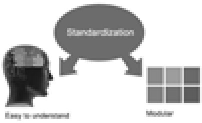


You may join as member of the Editorial Board of Global Journals Incorporation (USA) after successful completion of three years as Fellow and as Peer Reviewer. In addition, it is also desirable that you should organize seminar/symposium/conference at least once.

We shall provide you intimation regarding launching of e-version of journal of your stream time to time. This may be utilized in your library for the enrichment of knowledge of your students as well as it can also be helpful for the concerned faculty members.

The FARSC can go through standards of OARS. You can also play vital role if you have any suggestions so that proper amendment can take place to improve the same for the benefit of entire research community.

As FARSC, you will be given a renowned, secure and free professional email address with 100 GB of space e.g. johnhall@globaljournals.org. This will include Webmail, Spam Assassin, Email Forwarders, Auto-Responders, Email Delivery Route tracing, etc.



The FARSC will be eligible for a free application of standardization of their researches. Standardization of research will be subject to acceptability within stipulated norms as the next step after publishing in a journal. We shall depute a team of specialized research professionals who will render their services for elevating your researches to next higher level, which is worldwide open standardization.

The FARSC member can apply for grading and certification of standards of their educational and Institutional Degrees to Open Association of Research, Society U.S.A. Once you are designated as FARSC, you may send us a scanned copy of all of your credentials. OARS will verify, grade and certify them. This will be based on your academic records, quality of research papers published by you, and some more criteria. After certification of all your credentials by OARS, they will be published on your Fellow Profile link on website https://associationofresearch.org which will be helpful to upgrade the dignity.

The FARSC members can avail the benefits of free research podcasting in Global Research Radio with their research documents. After publishing the work, (including published elsewhere worldwide with proper authorization) you can upload your research paper with your recorded voice or you can utilize chargeable services of our professional RJs to record your paper in their voice on request.

The FARSC member also entitled to get the benefits of free research podcasting of their research documents through video clips. We can also streamline your conference videos and display your slides/ online slides and online research video clips at reasonable charges, on request.

The FARSC is eligible to earn from sales proceeds of his/her researches/reference/review Books or literature, while publishing with Global Journals. The FARSC can decide whether he/she would like to publish his/her research in a closed manner. In this case, whenever readers purchase that individual research paper for reading, maximum 60% of its profit earned as royalty by Global Journals, will be credited to his/her bank account. The entire entitled amount will be credited to his/her bank account exceeding limit of minimum fixed balance. There is no minimum time limit for collection. The FARSC member can decide its price and we can help in making the right decision.

The FARSC member is eligible to join as a paid peer reviewer at Global Journals Incorporation (USA) and can get remuneration of 15% of author fees, taken from the author of a respective paper. After reviewing 5 or more papers you can request to transfer the amount to your bank account.

MEMBER OF ASSOCIATION OF RESEARCH SOCIETY IN COMPUTING (MARSC)

The ' MARSC ' title is accorded to a selected professional after the approval of the Editor-in-Chief / Editorial Board Members/Dean.
The “MARSC” is a dignified ornament which is accorded to a person’s name viz. Dr. John E. Hall, Ph.D., MARSC or William Walldroff, M.S., MARSC.

MARSC accrediting is an honor. It authenticates your research activities. After becoming MARSC, you can add 'MARSC' title with your name as you use this recognition as additional suffix to your status. This will definitely enhance and add more value and repute to your name. You may use it on your professional Counseling Materials such as CV, Resume, Visiting Card and Name Plate etc.

The following benefitscan be availed by you only for next three years from the date of certification.

MARSC designated members are entitled to avail a 25% discount while publishing their research papers (of a single author) in Global Journals Inc., if the same is accepted by our Editorial Board and Peer Reviewers. If you are a main author or co-author of a group of authors, you will get discount of 10%.

As MARSC, you will be given a renowned, secure and free professional email address with 30 GB of space e.g. johnhall@globaljournals.org. This will include Webmail, Spam Assassin, Email Forwarders, Auto-Responders, Email Delivery Route tracing, etc.

We shall provide you intimation regarding launching of e-version of journal of your stream time to time. This may be utilized in your library for the enrichment of knowledge of your students as well as it can also be helpful for the concerned faculty members.

The MARSC member can apply for approval, grading and certification of standards of their educational and Institutional Degrees to Open Association of Research, Society U.S.A.

Once you are designated as MARSC, you may send us a scanned copy of all of your credentials. OARS will verify, grade and certify them. This will be based on your academic records, quality of research papers published by you, and some more criteria.

It is mandatory to read all terms and conditions carefully.

Auxiliary Memberships

Institutional Fellow of Open Association of Research Society (USA)-OARS (USA)

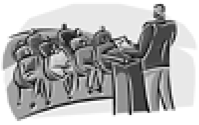
Global Journals Incorporation (USA) is accredited by Open Association of Research Society, U.S.A (OARS) and in turn, affiliates research institutions as “Institutional Fellow of Open Association of Research Society” (IFOARS).

The “FARSC” is a dignified title which is accorded to a person’s name viz. Dr. John E. Hall, Ph.D., FARSC or William Walldroff, M.S., FARSC.

The IFOARS institution is entitled to form a Board comprised of one Chairperson and three to five board members preferably from different streams. The Board will be recognized as “Institutional Board of Open Association of Research Society”-(IBOARS).

The Institute will be entitled to following benefits:

The IBOARS can initially review research papers of their institute and recommend them to publish with respective journal of Global Journals. It can also review the papers of other institutions after obtaining our consent. The second review will be done by peer reviewer of Global Journals Incorporation (USA) The Board is at liberty to appoint a peer reviewer with the approval of chairperson after consulting us.
The author fees of such paper may be waived off up to 40%.



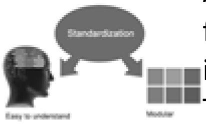
The Global Journals Incorporation (USA) at its discretion can also refer double blind peer reviewed paper at their end to the board for the verification and to get recommendation for final stage of acceptance of publication.

The IBOARS can organize symposium/seminar/conference in their country on behalf of Global Journals Incorporation (USA)-OARS (USA). The terms and conditions can be discussed separately.

The Board can also play vital role by exploring and giving valuable suggestions regarding the Standards of “Open Association of Research Society, U.S.A (OARS)” so that proper amendment can take place for the benefit of entire research community. We shall provide details of particular standard only on receipt of request from the Board.

The board members can also join us as Individual Fellow with 40% discount on total fees applicable to Individual Fellow. They will be entitled to avail all the benefits as declared. Please visit Individual Fellow-sub menu of GlobalJournals.org to have more relevant details.

We shall provide you intimation regarding launching of e-version of journal of your stream time to time. This may be utilized in your library for the enrichment of knowledge of your students as well as it can also be helpful for the concerned faculty members.



After nomination of your institution as “Institutional Fellow” and constantly functioning successfully for one year, we can consider giving recognition to your institute to function as Regional/Zonal office on our behalf.
The board can also take up the additional allied activities for betterment after our consultation.

The following entitlements are applicable to individual Fellows:

Open Association of Research Society, U.S.A (OARS) By-laws states that an individual Fellow may use the designations as applicable, or the corresponding initials. The Credentials of individual Fellow and Associate designations signify that the individual has gained knowledge of the fundamental concepts. One is magnanimous and proficient in an expertise course covering the professional code of conduct, and follows recognized standards of practice.

Open Association of Research Society (US)/ Global Journals Incorporation (USA), as described in Corporate Statements, are educational, research publishing and professional membership organizations. Achieving our individual Fellow or Associate status is based mainly on meeting stated educational research requirements.

Disbursement of 40% Royalty earned through Global Journals : Researcher = 50%, Peer Reviewer = 37.50%, Institution = 12.50% E.g. Out of 40%, the 20% benefit should be passed on to researcher, 15 % benefit towards remuneration should be given to a reviewer and remaining 5% is to be retained by the institution.

We shall provide print version of 12 issues of any three journals [as per your requirement] out of our 38 journals worth $ 2376 USD.

Other:

The individual Fellow and Associate designations accredited by Open Association of Research Society (US) credentials signify guarantees following achievements:

- The professional accredited with Fellow honor, is entitled to various benefits viz. name, fame, honor, regular flow of income, secured bright future, social status etc.

- In addition to above, if one is single author, then entitled to 40% discount on publishing research paper and can get 10%discount if one is co-author or main author among group of authors.
- The Fellow can organize symposium/seminar/conference on behalf of Global Journals Incorporation (USA) and he/she can also attend the same organized by other institutes on behalf of Global Journals.
- The Fellow can become member of Editorial Board Member after completing 3yrs.
- The Fellow can earn 60% of sales proceeds from the sale of reference/review books/literature/publishing of research paper.
- Fellow can also join as paid peer reviewer and earn 15% remuneration of author charges and can also get an opportunity to join as member of the Editorial Board of Global Journals Incorporation (USA)
- • This individual has learned the basic methods of applying those concepts and techniques to common challenging situations. This individual has further demonstrated an in–depth understanding of the application of suitable techniques to a particular area of research practice.

Note :

> 〃
>
> - In future, if the board feels the necessity to change any board member, the same can be done with the consent of the chairperson along with anyone board member without our approval.
> - In case, the chairperson needs to be replaced then consent of 2/3rd board members are required and they are also required to jointly pass the resolution copy of which should be sent to us. In such case, it will be compulsory to obtain our approval before replacement.
> - In case of “Difference of Opinion [if any]” among the Board members, our decision will be final and binding to everyone. 〃

Process of submission of Research Paper

The Area or field of specialization may or may not be of any category as mentioned in 'Scope of Journal' menu of the GlobalJournals.org website. There are 37 Research Journal categorized with Six parental Journals GJCST, GJMR, GJRE, GJMBR, GJSFR, GJHSS. For Authors should prefer the mentioned categories. There are three widely used systems UDC, DDC and LCC. The details are available as 'Knowledge Abstract' at Home page. The major advantage of this coding is that, the research work will be exposed to and shared with all over the world as we are being abstracted and indexed worldwide.

The paper should be in proper format. The format can be downloaded from first page of 'Author Guideline' Menu. The Author is expected to follow the general rules as mentioned in this menu. The paper should be written in MS-Word Format (*.DOC,*.DOCX).

The Author can submit the paper either online or offline. The authors should prefer online submission.Online Submission: There are three ways to submit your paper:

(A) (I) First, register yourself using top right corner of Home page then Login. If you are already registered, then login using your username and password.

(II) Choose corresponding Journal.

(III) Click 'Submit Manuscript'. Fill required information and Upload the paper.

(B) If you are using Internet Explorer, then Direct Submission through Homepage is also available.

(C) If these two are not convenient, and then email the paper directly to dean@globaljournals.org.

Offline Submission: Author can send the typed form of paper by Post. However, online submission should be preferred.

Preferred Author Guidelines

MANUSCRIPT STYLE INSTRUCTION (Must be strictly followed)

Page Size: 8.27" X 11'"

- Left Margin: 0.65
- Right Margin: 0.65
- Top Margin: 0.75
- Bottom Margin: 0.75
- Font type of all text should be Swis 721 Lt BT.
- Paper Title should be of Font Size 24 with one Column section.
- Author Name in Font Size of 11 with one column as of Title.
- Abstract Font size of 9 Bold, “Abstract” word in Italic Bold.
- Main Text: Font size 10 with justified two columns section
- Two Column with Equal Column with of 3.38 and Gaping of .2
- First Character must be three lines Drop capped.
- Paragraph before Spacing of 1 pt and After of 0 pt.
- Line Spacing of 1 pt
- Large Images must be in One Column
- Numbering of First Main Headings (Heading 1) must be in Roman Letters, Capital Letter, and Font Size of 10.
- Numbering of Second Main Headings (Heading 2) must be in Alphabets, Italic, and Font Size of 10.

You can use your own standard format also.
Author Guidelines:

1. General,

2. Ethical Guidelines,

3. Submission of Manuscripts,

4. Manuscript’s Category,

5. Structure and Format of Manuscript,

6. After Acceptance.

1. GENERAL

Before submitting your research paper, one is advised to go through the details as mentioned in following heads. It will be beneficial, while peer reviewer justify your paper for publication.

Scope

The Global Journals Inc. (US) welcome the submission of original paper, review paper, survey article relevant to the all the streams of Philosophy and knowledge. The Global Journals Inc. (US) is parental platform for Global Journal of Computer Science and Technology, Researches in Engineering, Medical Research, Science Frontier Research, Human Social Science, Management, and Business organization. The choice of specific field can be done otherwise as following in Abstracting and Indexing Page on this Website. As the all Global

Journals Inc. (US) are being abstracted and indexed (in process) by most of the reputed organizations. Topics of only narrow interest will not be accepted unless they have wider potential or consequences.

2. ETHICAL GUIDELINES

Authors should follow the ethical guidelines as mentioned below for publication of research paper and research activities.

Papers are accepted on strict understanding that the material in whole or in part has not been, nor is being, considered for publication elsewhere. If the paper once accepted by Global Journals Inc. (US) and Editorial Board, will become the copyright of the Global Journals Inc. (US).

Authorship: The authors and coauthors should have active contribution to conception design, analysis and interpretation of findings. They should critically review the contents and drafting of the paper. All should approve the final version of the paper before submission

The Global Journals Inc. (US) follows the definition of authorship set up by the Global Academy of Research and Development. According to the Global Academy of R&D authorship, criteria must be based on:

1) Substantial contributions to conception and acquisition of data, analysis and interpretation of the findings.

2) Drafting the paper and revising it critically regarding important academic content.

3) Final approval of the version of the paper to be published.

All authors should have been credited according to their appropriate contribution in research activity and preparing paper. Contributors who do not match the criteria as authors may be mentioned under Acknowledgement.

Acknowledgements: Contributors to the research other than authors credited should be mentioned under acknowledgement. The specifications of the source of funding for the research if appropriate can be included. Suppliers of resources may be mentioned along with address.

Appeal of Decision: The Editorial Board’s decision on publication of the paper is final and cannot be appealed elsewhere.

Permissions: It is the author's responsibility to have prior permission if all or parts of earlier published illustrations are used in this paper.

Please mention proper reference and appropriate acknowledgements wherever expected.

If all or parts of previously published illustrations are used, permission must be taken from the copyright holder concerned. It is the author's responsibility to take these in writing.

Approval for reproduction/modification of any information (including figures and tables) published elsewhere must be obtained by the authors/copyright holders before submission of the manuscript. Contributors (Authors) are responsible for any copyright fee involved.

3. SUBMISSION OF MANUSCRIPTS

Manuscripts should be uploaded via this online submission page. The online submission is most efficient method for submission of papers, as it enables rapid distribution of manuscripts and consequently speeds up the review procedure. It also enables authors to know the status of their own manuscripts by emailing us. Complete instructions for submitting a paper is available below.

Manuscript submission is a systematic procedure and little preparation is required beyond having all parts of your manuscript in a given format and a computer with an Internet connection and a Web browser. Full help and instructions are provided on-screen. As an author, you will be prompted for login and manuscript details as Field of Paper and then to upload your manuscript file(s) according to the instructions.

To avoid postal delays, all transaction is preferred by e-mail. A finished manuscript submission is confirmed by e-mail immediately and your paper enters the editorial process with no postal delays. When a conclusion is made about the publication of your paper by our Editorial Board, revisions can be submitted online with the same procedure, with an occasion to view and respond to all comments.

Complete support for both authors and co-author is provided.

4. MANUSCRIPT'S CATEGORY

Based on potential and nature, the manuscript can be categorized under the following heads:

Original research paper: Such papers are reports of high-level significant original research work.

Review papers: These are concise, significant but helpful and decisive topics for young researchers.

Research articles: These are handled with small investigation and applications.

Research letters: The letters are small and concise comments on previously published matters.

5. STRUCTURE AND FORMAT OF MANUSCRIPT

The recommended size of original research paper is less than seven thousand words, review papers fewer than seven thousands words also.Preparation of research paper or how to write research paper, are major hurdle, while writing manuscript. The research articles and research letters should be fewer than three thousand words, the structure original research paper; sometime review paper should be as follows:

Papers: These are reports of significant research (typically less than 7000 words equivalent, including tables, figures, references), and comprise:

(a)Title should be relevant and commensurate with the theme of the paper.

(b) A brief Summary, "Abstract" (less than 150 words) containing the major results and conclusions.

(c) Up to ten keywords, that precisely identifies the paper's subject, purpose, and focus.

(d) An Introduction, giving necessary background excluding subheadings; objectives must be clearly declared.

(e) Resources and techniques with sufficient complete experimental details (wherever possible by reference) to permit repetition; sources of information must be given and numerical methods must be specified by reference, unless non-standard.

(f) Results should be presented concisely, by well-designed tables and/or figures; the same data may not be used in both; suitable statistical data should be given. All data must be obtained with attention to numerical detail in the planning stage. As reproduced design has been recognized to be important to experiments for a considerable time, the Editor has decided that any paper that appears not to have adequate numerical treatments of the data will be returned un-refereed;

(g) Discussion should cover the implications and consequences, not just recapitulating the results; conclusions should be summarizing.

(h) Brief Acknowledgements.

(i) References in the proper form.

Authors should very cautiously consider the preparation of papers to ensure that they communicate efficiently. Papers are much more likely to be accepted, if they are cautiously designed and laid out, contain few or no errors, are summarizing, and be conventional to the approach and instructions. They will in addition, be published with much less delays than those that require much technical and editorial correction.

The Editorial Board reserves the right to make literary corrections and to make suggestions to improve briefness.

It is vital, that authors take care in submitting a manuscript that is written in simple language and adheres to published guidelines.

Format

Language: The language of publication is UK English. Authors, for whom English is a second language, must have their manuscript efficiently edited by an English-speaking person before submission to make sure that, the English is of high excellence. It is preferable, that manuscripts should be professionally edited.

Standard Usage, Abbreviations, and Units: Spelling and hyphenation should be conventional to The Concise Oxford English Dictionary. Statistics and measurements should at all times be given in figures, e.g. 16 min, except for when the number begins a sentence. When the number does not refer to a unit of measurement it should be spelt in full unless, it is 160 or greater.

Abbreviations supposed to be used carefully. The abbreviated name or expression is supposed to be cited in full at first usage, followed by the conventional abbreviation in parentheses.

Metric SI units are supposed to generally be used excluding where they conflict with current practice or are confusing. For illustration, 1.4 l rather than 1.4 × 10-3 m3, or 4 mm somewhat than 4 × 10-3 m. Chemical formula and solutions must identify the form used, e.g. anhydrous or hydrated, and the concentration must be in clearly defined units. Common species names should be followed by underlines at the first mention. For following use the generic name should be constricted to a single letter, if it is clear.

Structure

All manuscripts submitted to Global Journals Inc. (US), ought to include:

Title: The title page must carry an instructive title that reflects the content, a running title (less than 45 characters together with spaces), names of the authors and co-authors, and the place(s) wherever the work was carried out. The full postal address in addition with the e-mail address of related author must be given. Up to eleven keywords or very brief phrases have to be given to help data retrieval, mining and indexing.

Abstract, used in Original Papers and Reviews:

Optimizing Abstract for Search Engines

Many researchers searching for information online will use search engines such as Google, Yahoo or similar. By optimizing your paper for search engines, you will amplify the chance of someone finding it. This in turn will make it more likely to be viewed and/or cited in a further work. Global Journals Inc. (US) have compiled these guidelines to facilitate you to maximize the web-friendliness of the most public part of your paper.

Key Words

A major linchpin in research work for the writing research paper is the keyword search, which one will employ to find both library and Internet resources.

One must be persistent and creative in using keywords. An effective keyword search requires a strategy and planning a list of possible keywords and phrases to try.

Search engines for most searches, use Boolean searching, which is somewhat different from Internet searches. The Boolean search uses "operators," words (and, or, not, and near) that enable you to expand or narrow your affords. Tips for research paper while preparing research paper are very helpful guideline of research paper.

Choice of key words is first tool of tips to write research paper. Research paper writing is an art.A few tips for deciding as strategically as possible about keyword search:

- One should start brainstorming lists of possible keywords before even begin searching. Think about the most important concepts related to research work. Ask, "What words would a source have to include to be truly valuable in research paper?" Then consider synonyms for the important words.
- It may take the discovery of only one relevant paper to let steer in the right keyword direction because in most databases, the keywords under which a research paper is abstracted are listed with the paper.
- One should avoid outdated words.

Keywords are the key that opens a door to research work sources. Keyword searching is an art in which researcher's skills are bound to improve with experience and time.

Numerical Methods: Numerical methods used should be clear and, where appropriate, supported by references.

Acknowledgements: Please make these as concise as possible.

References

References follow the Harvard scheme of referencing. References in the text should cite the authors' names followed by the time of their publication, unless there are three or more authors when simply the first author's name is quoted followed by et al. unpublished work has to only be cited where necessary, and only in the text. Copies of references in press in other journals have to be supplied with submitted typescripts. It is necessary that all citations and references be carefully checked before submission, as mistakes or omissions will cause delays.

References to information on the World Wide Web can be given, but only if the information is available without charge to readers on an official site. Wikipedia and Similar websites are not allowed where anyone can change the information. Authors will be asked to make available electronic copies of the cited information for inclusion on the Global Journals Inc. (US) homepage at the judgment of the Editorial Board.

The Editorial Board and Global Journals Inc. (US) recommend that, citation of online-published papers and other material should be done via a DOI (digital object identifier). If an author cites anything, which does not have a DOI, they run the risk of the cited material not being noticeable.

The Editorial Board and Global Journals Inc. (US) recommend the use of a tool such as Reference Manager for reference management and formatting.

Tables, Figures and Figure Legends

Tables: Tables should be few in number, cautiously designed, uncrowned, and include only essential data. Each must have an Arabic number, e.g. Table 4, a self-explanatory caption and be on a separate sheet. Vertical lines should not be used.

Figures: Figures are supposed to be submitted as separate files. Always take in a citation in the text for each figure using Arabic numbers, e.g. Fig. 4. Artwork must be submitted online in electronic form by e-mailing them.

Preparation of Electronic Figures for Publication

Even though low quality images are sufficient for review purposes, print publication requires high quality images to prevent the final product being blurred or fuzzy. Submit (or e-mail) EPS (line art) or TIFF (halftone/photographs) files only. MS PowerPoint and Word Graphics are unsuitable for printed pictures. Do not use pixel-oriented software. Scans (TIFF only) should have a resolution of at least 350 dpi (halftone) or 700 to 1100 dpi (line drawings) in relation to the imitation size. Please give the data for figures in black and white or submit a Color Work Agreement Form. EPS files must be saved with fonts embedded (and with a TIFF preview, if possible).

For scanned images, the scanning resolution (at final image size) ought to be as follows to ensure good reproduction: line art: >650 dpi; halftones (including gel photographs) : >350 dpi; figures containing both halftone and line images: >650 dpi.

Color Charges: It is the rule of the Global Journals Inc. (US) for authors to pay the full cost for the reproduction of their color artwork. Hence, please note that, if there is color artwork in your manuscript when it is accepted for publication, we would require you to complete and return a color work agreement form before your paper can be published.

Figure Legends: Self-explanatory legends of all figures should be incorporated separately under the heading 'Legends to Figures'. In the full-text online edition of the journal, figure legends may possibly be truncated in abbreviated links to the full screen version. Therefore, the first 100 characters of any legend should notify the reader, about the key aspects of the figure.

6. AFTER ACCEPTANCE

Upon approval of a paper for publication, the manuscript will be forwarded to the dean, who is responsible for the publication of the Global Journals Inc. (US).

6.1 Proof Corrections

The corresponding author will receive an e-mail alert containing a link to a website or will be attached. A working e-mail address must therefore be provided for the related author.

Acrobat Reader will be required in order to read this file. This software can be downloaded

(Free of charge) from the following website:

www.adobe.com/products/acrobat/readstep2.html. This will facilitate the file to be opened, read on screen, and printed out in order for any corrections to be added. Further instructions will be sent with the proof.

Proofs must be returned to the dean at dean@globaljournals.org within three days of receipt.

As changes to proofs are costly, we inquire that you only correct typesetting errors. All illustrations are retained by the publisher. Please note that the authors are responsible for all statements made in their work, including changes made by the copy editor.

6.2 Early View of Global Journals Inc. (US) (Publication Prior to Print)

The Global Journals Inc. (US) are enclosed by our publishing's Early View service. Early View articles are complete full-text articles sent in advance of their publication. Early View articles are absolute and final. They have been completely reviewed, revised and edited for publication, and the authors' final corrections have been incorporated. Because they are in final form, no changes can be made after sending them. The nature of Early View articles means that they do not yet have volume, issue or page numbers, so Early View articles cannot be cited in the conventional way.

6.3 Author Services

Online production tracking is available for your article through Author Services. Author Services enables authors to track their article - once it has been accepted - through the production process to publication online and in print. Authors can check the status of their articles online and choose to receive automated e-mails at key stages of production. The authors will receive an e-mail with a unique link that enables them to register and have their article automatically added to the system. Please ensure that a complete e-mail address is provided when submitting the manuscript.

6.4 Author Material Archive Policy

Please note that if not specifically requested, publisher will dispose off hardcopy & electronic information submitted, after the two months of publication. If you require the return of any information submitted, please inform the Editorial Board or dean as soon as possible.

6.5 Offprint and Extra Copies

A PDF offprint of the online-published article will be provided free of charge to the related author, and may be distributed according to the Publisher's terms and conditions. Additional paper offprint may be ordered by emailing us at: editor@globaljournals.org .

You must strictly follow above Author Guidelines before submitting your paper or else we will not at all be responsible for any corrections in future in any of the way.

Before start writing a good quality Computer Science Research Paper, let us first understand what is Computer Science Research Paper? So, Computer Science Research Paper is the paper which is written by professionals or scientists who are associated to Computer Science and Information Technology, or doing research study in these areas. If you are novel to this field then you can consult about this field from your supervisor or guide.

TECHNIQUES FOR WRITING A GOOD QUALITY RESEARCH PAPER:

1. Choosing the topic: In most cases, the topic is searched by the interest of author but it can be also suggested by the guides. You can have several topics and then you can judge that in which topic or subject you are finding yourself most comfortable. This can be done by asking several questions to yourself, like Will I be able to carry our search in this area? Will I find all necessary recourses to accomplish the search? Will I be able to find all information in this field area? If the answer of these types of questions will be "Yes" then you can choose that topic. In most of the cases, you may have to conduct the surveys and have to visit several places because this field is related to Computer Science and Information Technology. Also, you may have to do a lot of work to find all rise and falls regarding the various data of that subject. Sometimes, detailed information plays a vital role, instead of short information.

2. Evaluators are human: First thing to remember that evaluators are also human being. They are not only meant for rejecting a paper. They are here to evaluate your paper. So, present your Best.

3. Think Like Evaluators: If you are in a confusion or getting demotivated that your paper will be accepted by evaluators or not, then think and try to evaluate your paper like an Evaluator. Try to understand that what an evaluator wants in your research paper and automatically you will have your answer.

4. Make blueprints of paper: The outline is the plan or framework that will help you to arrange your thoughts. It will make your paper logical. But remember that all points of your outline must be related to the topic you have chosen.

5. Ask your Guides: If you are having any difficulty in your research, then do not hesitate to share your difficulty to your guide (if you have any). They will surely help you out and resolve your doubts. If you can't clarify what exactly you require for your work then ask the supervisor to help you with the alternative. He might also provide you the list of essential readings.

6. Use of computer is recommended: As you are doing research in the field of Computer Science, then this point is quite obvious.

7. Use right software: Always use good quality software packages. If you are not capable to judge good software then you can lose quality of your paper unknowingly. There are various software programs available to help you, which you can get through Internet.

8. Use the Internet for help: An excellent start for your paper can be by using the Google. It is an excellent search engine, where you can have your doubts resolved. You may also read some answers for the frequent question how to write my research paper or find model research paper. From the internet library you can download books. If you have all required books make important reading selecting and analyzing the specified information. Then put together research paper sketch out.

9. Use and get big pictures: Always use encyclopedias, Wikipedia to get pictures so that you can go into the depth.

10. Bookmarks are useful: When you read any book or magazine, you generally use bookmarks, right! It is a good habit, which helps to not to lose your continuity. You should always use bookmarks while searching on Internet also, which will make your search easier.

11. Revise what you wrote: When you write anything, always read it, summarize it and then finalize it.

12. Make all efforts: Make all efforts to mention what you are going to write in your paper. That means always have a good start. Try to mention everything in introduction, that what is the need of a particular research paper. Polish your work by good skill of writing and always give an evaluator, what he wants.

13. Have backups: When you are going to do any important thing like making research paper, you should always have backup copies of it either in your computer or in paper. This will help you to not to lose any of your important.

14. Produce good diagrams of your own: Always try to include good charts or diagrams in your paper to improve quality. Using several and unnecessary diagrams will degrade the quality of your paper by creating "hotchpotch." So always, try to make and include those diagrams, which are made by your own to improve readability and understandability of your paper.

15. Use of direct quotes: When you do research relevant to literature, history or current affairs then use of quotes become essential but if study is relevant to science then use of quotes is not preferable.

16. Use proper verb tense: Use proper verb tenses in your paper. Use past tense, to present those events that happened. Use present tense to indicate events that are going on. Use future tense to indicate future happening events. Use of improper and wrong tenses will confuse the evaluator. Avoid the sentences that are incomplete.

17. Never use online paper: If you are getting any paper on Internet, then never use it as your research paper because it might be possible that evaluator has already seen it or maybe it is outdated version.

18. Pick a good study spot: To do your research studies always try to pick a spot, which is quiet. Every spot is not for studies. Spot that suits you choose it and proceed further.

19. Know what you know: Always try to know, what you know by making objectives. Else, you will be confused and cannot achieve your target.

20. Use good quality grammar: Always use a good quality grammar and use words that will throw positive impact on evaluator. Use of good quality grammar does not mean to use tough words, that for each word the evaluator has to go through dictionary. Do not start sentence with a conjunction. Do not fragment sentences. Eliminate one-word sentences. Ignore passive voice. Do not ever use a big word when a diminutive one would suffice. Verbs have to be in agreement with their subjects. Prepositions are not expressions to finish sentences with. It is incorrect to ever divide an infinitive. Avoid clichés like the disease. Also, always shun irritating alliteration. Use language that is simple and straight forward. put together a neat summary.

21. Arrangement of information: Each section of the main body should start with an opening sentence and there should be a changeover at the end of the section. Give only valid and powerful arguments to your topic. You may also maintain your arguments with records.

22. Never start in last minute: Always start at right time and give enough time to research work. Leaving everything to the last minute will degrade your paper and spoil your work.

23. Multitasking in research is not good: Doing several things at the same time proves bad habit in case of research activity. Research is an area, where everything has a particular time slot. Divide your research work in parts and do particular part in particular time slot.

24. Never copy others' work: Never copy others' work and give it your name because if evaluator has seen it anywhere you will be in trouble.

25. Take proper rest and food: No matter how many hours you spend for your research activity, if you are not taking care of your health then all your efforts will be in vain. For a quality research, study is must, and this can be done by taking proper rest and food.

26. Go for seminars: Attend seminars if the topic is relevant to your research area. Utilize all your resources.

27. Refresh your mind after intervals: Try to give rest to your mind by listening to soft music or by sleeping in intervals. This will also improve your memory.

28. Make colleagues: Always try to make colleagues. No matter how sharper or intelligent you are, if you make colleagues you can have several ideas, which will be helpful for your research.

29. Think technically: Always think technically. If anything happens, then search its reasons, its benefits, and demerits.

30. Think and then print: When you will go to print your paper, notice that tables are not be split, headings are not detached from their descriptions, and page sequence is maintained.

31. Adding unnecessary information: Do not add unnecessary information, like, I have used MS Excel to draw graph. Do not add irrelevant and inappropriate material. These all will create superfluous. Foreign terminology and phrases are not apropos. One should NEVER take a broad view. Analogy in script is like feathers on a snake. Not at all use a large word when a very small one would be sufficient. Use words properly, regardless of how others use them. Remove quotations. Puns are for kids, not grunt readers. Amplification is a billion times of inferior quality than sarcasm.

32. Never oversimplify everything: To add material in your research paper, never go for oversimplification. This will definitely irritate the evaluator. Be more or less specific. Also too, by no means, ever use rhythmic redundancies. Contractions aren't essential and shouldn't be there used. Comparisons are as terrible as clichés. Give up ampersands and abbreviations, and so on. Remove commas, that are, not necessary. Parenthetical words however should be together with this in commas. Understatement is all the time the complete best way to put onward earth-shaking thoughts. Give a detailed literary review.

33. Report concluded results: Use concluded results. From raw data, filter the results and then conclude your studies based on measurements and observations taken. Significant figures and appropriate number of decimal places should be used. Parenthetical remarks are prohibitive. Proofread carefully at final stage. In the end give outline to your arguments. Spot out perspectives of further study of this subject. Justify your conclusion by at the bottom of them with sufficient justifications and examples.

34. After conclusion: Once you have concluded your research, the next most important step is to present your findings. Presentation is extremely important as it is the definite medium though which your research is going to be in print to the rest of the crowd. Care should be taken to categorize your thoughts well and present them in a logical and neat manner. A good quality research paper format is essential because it serves to highlight your research paper and bring to light all necessary aspects in your research.

Informal Guidelines of Research Paper Writing

Key points to remember:

- Submit all work in its final form.
- Write your paper in the form, which is presented in the guidelines using the template.
- Please note the criterion for grading the final paper by peer-reviewers.

Final Points:

A purpose of organizing a research paper is to let people to interpret your effort selectively. The journal requires the following sections, submitted in the order listed, each section to start on a new page.

The introduction will be compiled from reference matter and will reflect the design processes or outline of basis that direct you to make study. As you will carry out the process of study, the method and process section will be constructed as like that. The result segment will show related statistics in nearly sequential order and will direct the reviewers next to the similar intellectual paths throughout the data that you took to carry out your study. The discussion section will provide understanding of the data and projections as to the implication of the results. The use of good quality references all through the paper will give the effort trustworthiness by representing an alertness of prior workings.

Writing a research paper is not an easy job no matter how trouble-free the actual research or concept. Practice, excellent preparation, and controlled record keeping are the only means to make straightforward the progression.

General style:

Specific editorial column necessities for compliance of a manuscript will always take over from directions in these general guidelines.

To make a paper clear

· Adhere to recommended page limits

Mistakes to evade

- Insertion a title at the foot of a page with the subsequent text on the next page
- Separating a table/chart or figure - impound each figure/table to a single page
- Submitting a manuscript with pages out of sequence

In every sections of your document

· Use standard writing style including articles ("a", "the," etc.)

· Keep on paying attention on the research topic of the paper

· Use paragraphs to split each significant point (excluding for the abstract)

· Align the primary line of each section

· Present your points in sound order

· Use present tense to report well accepted

· Use past tense to describe specific results

· Shun familiar wording, don't address the reviewer directly, and don't use slang, slang language, or superlatives

· Shun use of extra pictures - include only those figures essential to presenting results

Title Page:

Choose a revealing title. It should be short. It should not have non-standard acronyms or abbreviations. It should not exceed two printed lines. It should include the name(s) and address (es) of all authors.

Abstract:

The summary should be two hundred words or less. It should briefly and clearly explain the key findings reported in the manuscript--must have precise statistics. It should not have abnormal acronyms or abbreviations. It should be logical in itself. Shun citing references at this point.

An abstract is a brief distinct paragraph summary of finished work or work in development. In a minute or less a reviewer can be taught the foundation behind the study, common approach to the problem, relevant results, and significant conclusions or new questions.

Write your summary when your paper is completed because how can you write the summary of anything which is not yet written? Wealth of terminology is very essential in abstract. Yet, use comprehensive sentences and do not let go readability for briefness. You can maintain it succinct by phrasing sentences so that they provide more than lone rationale. The author can at this moment go straight to shortening the outcome. Sum up the study, with the subsequent elements in any summary. Try to maintain the initial two items to no more than one ruling each.

- Reason of the study - theory, overall issue, purpose
- Fundamental goal
- To the point depiction of the research
- Consequences, including definite statistics - if the consequences are quantitative in nature, account quantitative data; results of any numerical analysis should be reported
- Significant conclusions or questions that track from the research(es)

Approach:

- Single section, and succinct
- As a outline of job done, it is always written in past tense
- A conceptual should situate on its own, and not submit to any other part of the paper such as a form or table
- Center on shortening results - bound background information to a verdict or two, if completely necessary
- What you account in an conceptual must be regular with what you reported in the manuscript
- Exact spelling, clearness of sentences and phrases, and appropriate reporting of quantities (proper units, important statistics) are just as significant in an abstract as they are anywhere else

Introduction:

The **Introduction** should "introduce" the manuscript. The reviewer should be presented with sufficient background information to be capable to comprehend and calculate the purpose of your study without having to submit to other works. The basis for the study should be offered. Give most important references but shun difficult to make a comprehensive appraisal of the topic. In the introduction, describe the problem visibly. If the problem is not acknowledged in a logical, reasonable way, the reviewer will have no attention in your result. Speak in common terms about techniques used to explain the problem, if needed, but do not present any particulars about the protocols here. Following approach can create a valuable beginning:

- Explain the value (significance) of the study
- Shield the model - why did you employ this particular system or method? What is its compensation? You strength remark on its appropriateness from a abstract point of vision as well as point out sensible reasons for using it.
- Present a justification. Status your particular theory (es) or aim(s), and describe the logic that led you to choose them.
- Very for a short time explain the tentative propose and how it skilled the declared objectives.

Approach:

- Use past tense except for when referring to recognized facts. After all, the manuscript will be submitted after the entire job is done.
- Sort out your thoughts; manufacture one key point with every section. If you make the four points listed above, you will need a least of four paragraphs.

- Present surroundings information only as desirable in order hold up a situation. The reviewer does not desire to read the whole thing you know about a topic.
- Shape the theory/purpose specifically - do not take a broad view.
- As always, give awareness to spelling, simplicity and correctness of sentences and phrases.

Procedures (Methods and Materials):

This part is supposed to be the easiest to carve if you have good skills. A sound written Procedures segment allows a capable scientist to replacement your results. Present precise information about your supplies. The suppliers and clarity of reagents can be helpful bits of information. Present methods in sequential order but linked methodologies can be grouped as a segment. Be concise when relating the protocols. Attempt for the least amount of information that would permit another capable scientist to spare your outcome but be cautious that vital information is integrated. The use of subheadings is suggested and ought to be synchronized with the results section. When a technique is used that has been well described in another object, mention the specific item describing a way but draw the basic principle while stating the situation. The purpose is to text all particular resources and broad procedures, so that another person may use some or all of the methods in one more study or referee the scientific value of your work. It is not to be a step by step report of the whole thing you did, nor is a methods section a set of orders.

Materials:

- Explain materials individually only if the study is so complex that it saves liberty this way.
- Embrace particular materials, and any tools or provisions that are not frequently found in laboratories.
- Do not take in frequently found.
- If use of a definite type of tools.
- Materials may be reported in a part section or else they may be recognized along with your measures.

Methods:

- Report the method (not particulars of each process that engaged the same methodology)
- Describe the method entirely
- To be succinct, present methods under headings dedicated to specific dealings or groups of measures
- Simplify - details how procedures were completed not how they were exclusively performed on a particular day.
- If well known procedures were used, account the procedure by name, possibly with reference, and that's all.

Approach:

- It is embarrassed or not possible to use vigorous voice when documenting methods with no using first person, which would focus the reviewer's interest on the researcher rather than the job. As a result when script up the methods most authors use third person passive voice.
- Use standard style in this and in every other part of the paper - avoid familiar lists, and use full sentences.

What to keep away from

- Resources and methods are not a set of information.
- Skip all descriptive information and surroundings - save it for the argument.
- Leave out information that is immaterial to a third party.

Results:

The principle of a results segment is to present and demonstrate your conclusion. Create this part a entirely objective details of the outcome, and save all understanding for the discussion.

The page length of this segment is set by the sum and types of data to be reported. Carry on to be to the point, by means of statistics and tables, if suitable, to present consequences most efficiently.You must obviously differentiate material that would usually be incorporated in a study editorial from any unprocessed data or additional appendix matter that would not be available. In fact, such matter should not be submitted at all except requested by the instructor.

Content

- Sum up your conclusion in text and demonstrate them, if suitable, with figures and tables.
- In manuscript, explain each of your consequences, point the reader to remarks that are most appropriate.
- Present a background, such as by describing the question that was addressed by creation an exacting study.
- Explain results of control experiments and comprise remarks that are not accessible in a prescribed figure or table, if appropriate.
- Examine your data, then prepare the analyzed (transformed) data in the form of a figure (graph), table, or in manuscript form.

What to stay away from

- Do not discuss or infer your outcome, report surroundings information, or try to explain anything.
- Not at all, take in raw data or intermediate calculations in a research manuscript.
- Do not present the similar data more than once.
- Manuscript should complement any figures or tables, not duplicate the identical information.
- Never confuse figures with tables - there is a difference.

Approach

- As forever, use past tense when you submit to your results, and put the whole thing in a reasonable order.
- Put figures and tables, appropriately numbered, in order at the end of the report
- If you desire, you may place your figures and tables properly within the text of your results part.

Figures and tables

- If you put figures and tables at the end of the details, make certain that they are visibly distinguished from any attach appendix materials, such as raw facts
- Despite of position, each figure must be numbered one after the other and complete with subtitle
- In spite of position, each table must be titled, numbered one after the other and complete with heading
- All figure and table must be adequately complete that it could situate on its own, divide from text

Discussion:

The Discussion is expected the trickiest segment to write and describe. A lot of papers submitted for journal are discarded based on problems with the Discussion. There is no head of state for how long a argument should be. Position your understanding of the outcome visibly to lead the reviewer through your conclusions, and then finish the paper with a summing up of the implication of the study. The purpose here is to offer an understanding of your results and hold up for all of your conclusions, using facts from your research and generally accepted information, if suitable. The implication of result should be visibly described. Infer your data in the conversation in suitable depth. This means that when you clarify an observable fact you must explain mechanisms that may account for the observation. If your results vary from your prospect, make clear why that may have happened. If your results agree, then explain the theory that the proof supported. It is never suitable to just state that the data approved with prospect, and let it drop at that.

- Make a decision if each premise is supported, discarded, or if you cannot make a conclusion with assurance. Do not just dismiss a study or part of a study as "uncertain."
- Research papers are not acknowledged if the work is imperfect. Draw what conclusions you can based upon the results that you have, and take care of the study as a finished work
- You may propose future guidelines, such as how the experiment might be personalized to accomplish a new idea.
- Give details all of your remarks as much as possible, focus on mechanisms.
- Make a decision if the tentative design sufficiently addressed the theory, and whether or not it was correctly restricted.
- Try to present substitute explanations if sensible alternatives be present.
- One research will not counter an overall question, so maintain the large picture in mind, where do you go next? The best studies unlock new avenues of study. What questions remain?
- Recommendations for detailed papers will offer supplementary suggestions.

Approach:

- When you refer to information, differentiate data generated by your own studies from available information
- Submit to work done by specific persons (including you) in past tense.
- Submit to generally acknowledged facts and main beliefs in present tense.

Administration Rules Listed Before
Submitting Your Research Paper to Global Journals Inc. (US)

Please carefully note down following rules and regulation before submitting your Research Paper to Global Journals Inc. (US):

Segment Draft and Final Research Paper: You have to strictly follow the template of research paper. If it is not done your paper may get rejected.

- The **major constraint** is that you must independently make all content, tables, graphs, and facts that are offered in the paper. You must write each part of the paper wholly on your own. The Peer-reviewers need to identify your own perceptive of the concepts in your own terms. NEVER extract straight from any foundation, and never rephrase someone else's analysis.

- Do not give permission to anyone else to "PROOFREAD" your manuscript.

- Methods to avoid Plagiarism is applied by us on every paper, if found guilty, you will be blacklisted by all of our collaborated research groups, your institution will be informed for this and strict legal actions will be taken immediately.)
- To guard yourself and others from possible illegal use please do not permit anyone right to use to your paper and files.

Criterion for Grading a Research Paper (Compilation) by Global Journals Inc. (US)

Please note that following table is only a Grading of "Paper Compilation" and not on "Performed/Stated Research" whose grading solely depends on Individual Assigned Peer Reviewer and Editorial Board Member. These can be available only on request and after decision of Paper. This report will be the property of Global Journals Inc. (US).

Topics	Grades		
	A-B	**C-D**	**E-F**
Abstract	Clear and concise with appropriate content, Correct format. 200 words or below	Unclear summary and no specific data, Incorrect form Above 200 words	No specific data with ambiguous information Above 250 words
Introduction	Containing all background details with clear goal and appropriate details, flow specification, no grammar and spelling mistake, well organized sentence and paragraph, reference cited	Unclear and confusing data, appropriate format, grammar and spelling errors with unorganized matter	Out of place depth and content, hazy format
Methods and Procedures	Clear and to the point with well arranged paragraph, precision and accuracy of facts and figures, well organized subheads	Difficult to comprehend with embarrassed text, too much explanation but completed	Incorrect and unorganized structure with hazy meaning
Result	Well organized, Clear and specific, Correct units with precision, correct data, well structuring of paragraph, no grammar and spelling mistake	Complete and embarrassed text, difficult to comprehend	Irregular format with wrong facts and figures
Discussion	Well organized, meaningful specification, sound conclusion, logical and concise explanation, highly structured paragraph reference cited	Wordy, unclear conclusion, spurious	Conclusion is not cited, unorganized, difficult to comprehend
References	Complete and correct format, well organized	Beside the point, Incomplete	Wrong format and structuring

Index

save our planet

Global Journal of Computer Science and Technology

Visit us on the Web at www.GlobalJournals.org | www.ComputerResearch.org
or email us at helpdesk@globaljournals.org

9 70116 58698 2 61427>

ISSN 9754350