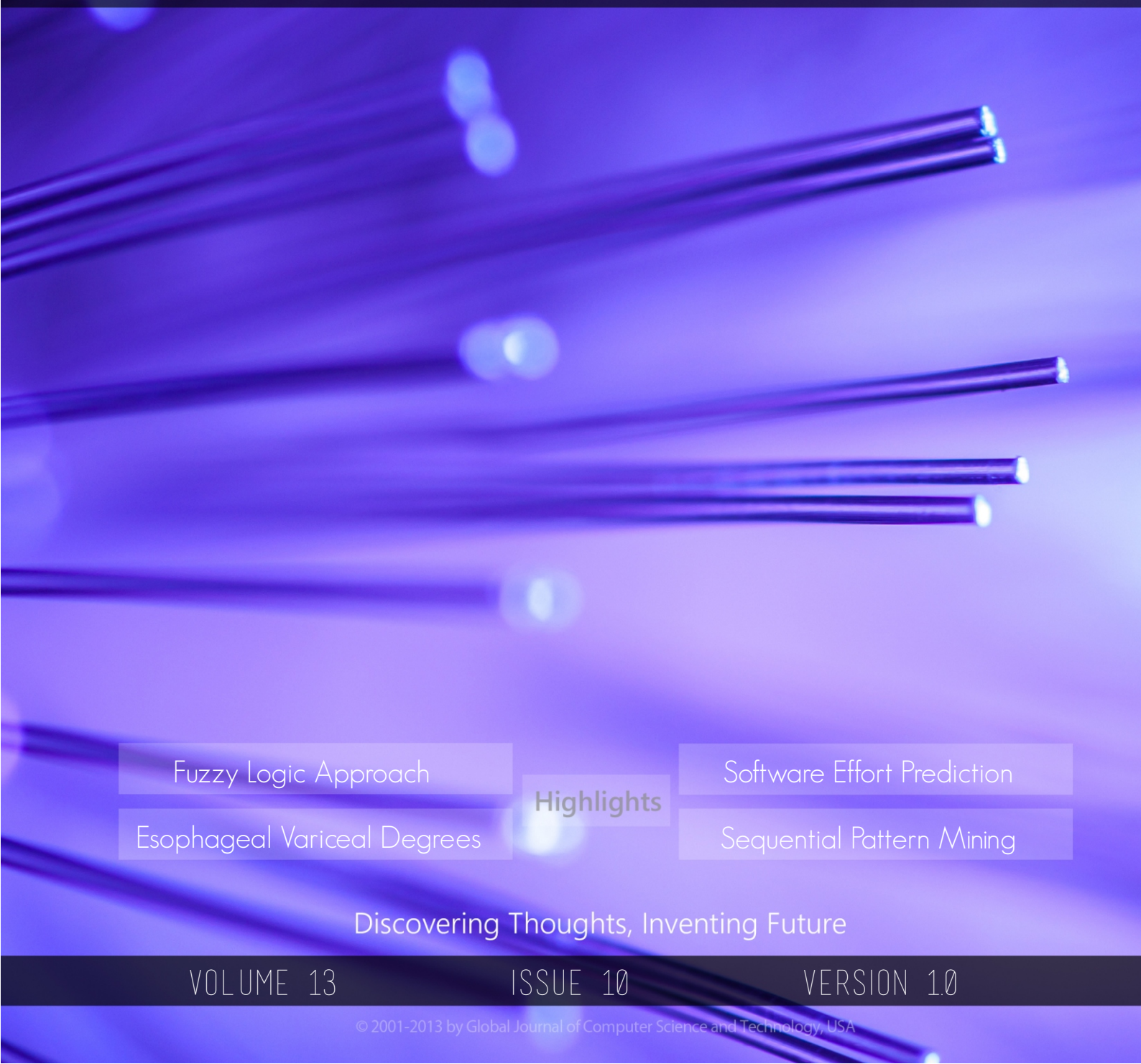
Online ISSN : 0975-4172
Print ISSN : 0975-4350

Global Journal

of Computer Science and Technology: C

Software & Data Engineering



Highlights

- Fuzzy Logic Approach
- Esophageal Variceal Degrees
- Software Effort Prediction
- Sequential Pattern Mining

Discovering Thoughts, Inventing Future

VOLUME 13 ISSUE 10 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 13 Issue 10 (Ver. 1.0)

Open Association of Research Society

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 Inc., Headquarters Corporate Office,
Cambridge Office Center, II Canal Park, Floor No. 5th, ***Cambridge (Massachusetts)***, Pin: MA 02141
United States
USA Toll Free: +001-888-839-7392
USA Toll Free Fax: +001-888-839-7392

Offset Typesetting

Global Association of Research, Marsh Road, Rainham, Essex, London RM13 8EU
United Kingdom.

Packaging & Continental Dispatching

Global Journals, 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: *investers@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)

Editorial Board Members (HON.)

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

Dr. Pina C. Sanelli
Associate Professor of Public Health
Weill Cornell Medical College
Associate Attending Radiologist
NewYork-Presbyterian Hospital
MRI, MRA, CT, and CTA
Neuroradiology and Diagnostic Radiology
M.D., State University of New York at Buffalo,School of Medicine and Biomedical Sciences

Dr. Roberto Sanchez
Associate Professor
Department of Structural and Chemical Biology
Mount Sinai School of Medicine
Ph.D., The Rockefeller University

Dr. Wen-Yih Sun
Professor of Earth and Atmospheric SciencesPurdue University Director
National Center for Typhoon and Flooding Research, Taiwan
University Chair Professor
Department of Atmospheric Sciences, National Central University, Chung-Li, TaiwanUniversity Chair Professor
Institute of Environmental Engineering, National Chiao Tung University, Hsin-chu, Taiwan.Ph.D., MS The University of Chicago, Geophysical Sciences
BS National Taiwan University, Atmospheric Sciences
Associate Professor of Radiology

Dr. Michael R. Rudnick
M.D., FACP
Associate Professor of Medicine
Chief, Renal Electrolyte and Hypertension Division (PMC)
Penn Medicine, University of Pennsylvania
Presbyterian Medical Center, Philadelphia
Nephrology and Internal Medicine
Certified by the American Board of Internal Medicine

Dr. Bassey Benjamin Esu
B.Sc. Marketing; MBA Marketing; Ph.D Marketing
Lecturer, Department of Marketing, University of Calabar
Tourism Consultant, Cross River State Tourism Development Department
Co-ordinator , Sustainable Tourism Initiative, Calabar, Nigeria

Dr. Aziz M. Barbar, Ph.D.
IEEE Senior Member
Chairperson, Department of Computer Science
AUST - American University of Science & Technology
Alfred Naccash Avenue – Ashrafieh

Contents of the Volume

Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 13 Issue 10 Version 1.0 Year 2013
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Prediction Analysis of Esophageal Variceal Degrees using Data Mining: Is Validated in Clinical Medicine?

By Abd Elrazek Mohammad Aly Abd Elrazek & Hamdy Mahfouz
Al-Azhar University, Egypt

Abstract - The objective of this study is to assess the feasibility of a data mining association analysis technique in early prediction of esophageal varices in cirrhotic patients and prediction of risky groups candidates for urgent interventional procedure. A manuscript titled "Detection of Risky Esophageal varices using 2D U/S: when to perform Endoscopy", published in The American Journal of The Medical Science on 21Th of December 2012, to our knowledge it was the first prospective study to assess the degree of esophageal varices by 2D ultrasound using the data mining statistical computed analysis in 673 patients. A descriptive model was generated using a decision tree algorithm (Rapid Miner, version 4.6, Berlin, Germany), the over all accuracy was 95%. Following another 59 patients using statistical analysis to determine the association between esophageal variceal degrees detected by Ultrasound in comparable to Upper Endoscopy, was done.

Categorical data were compared using the x^2 test, where as continuous variables were compared using Student's *t* test.

The comparative results accuracy of both two studies was 97.9%.

GJCST-C Classification : H.2.8

Strictly as per the compliance and regulations of:

Prediction Analysis of Esophageal Variceal Degrees using Data Mining: Is Validated in Clinical Medicine?

Abd Elrazek Mohammad Aly Abd Elrazek [α] & Hamdy Mahfouz [σ]

Abstract - The objective of this study is to assess the feasibility of a data mining association analysis technique in early prediction of esophageal varices in cirrhotic patients and prediction of risky groups candidates for urgent interventional procedure. A manuscript titled "Detection of Risky Esophageal varices using 2D U/S: when to perform Endoscopy", published in The American Journal of The Medical Science on 21Th of December 2012, to our knowledge it was the first prospective study to assess the degree of esophageal varices by 2D ultrasound using the data mining statistical computed analysis in 673 patients. A descriptive model was generated using a decision tree algorithm (Rapid Miner, version 4.6, Berlin, Germany), the over all accuracy was 95%. Following another 59 patients using statistical analysis to determine the association between esophageal variceal degrees detected by Ultrasound in comparable to Upper Endoscopy, was done.

Categorical data were compared using the χ^2 test, where as continuous variables were compared using Student's *t* test.

The comparative results accuracy of both two studies was 97.9%.

I. Introduction

Although the rate of mortality from GI bleeding episodes has decreased with improved endoscopic and radiologic techniques together with new pharmacologic therapies, the 13% to 20% mortality implies the clinical importance, whatever most of mortalities due vaiceal rupture or leakage [1]. Concern has been raised about the prediction of Esophageal varices for those expected to develop gastrointestinal bleeding in the future, analyzing data by Data mining programs discovered the significant important leading mortality factors; risky Esophageal varices [2]. Thus the major challenge of biomedical data mining over the next 5-10 years is to make these systems useful to biomedical researchers [3]. This paper discuss the utility usefulness of data mining as a predictor statistical tool used to predict esophageal variceal degrees in cirrhotic patients, through measuring the intra-abdominal esophageal wall thickness, by non invasive ultrasound technique.

II. Clinical View

a) Causes of Upper GI bleeding

The most common causes of upper gastrointestinal bleeding are: Gastric and/or duodenal ulcers, Esophagogastric varices with or without portal hypertensive gastropathy, Esophagitis, Erosive gastritis/duodenitis, Mallory-Weiss syndrome, Angiodysplasia, Mass lesions (polyps/cancers), Dieulafoy's lesion. [4].

Helicobacter pylori infection, Non Steroidal anti-inflammatory drugs (NSAIDs), physiologic stress, and excess gastric acid are major risk factors for bleeding peptic ulcers. Reduction or elimination of these risk factors reduces ulcer recurrence and rebleeding rates, whatever Esophagogastric varices develop as a consequence of portal hypertension Less common causes of upper gastrointestinal bleeding include Hemobilia, Hemosuccus pancreaticus, Aortoenteric fistula and Cameron lesions.[5,6].

b) Causes of Esophageal Varices

Esophageal varices are swollen blood vessels in the esophagus, which is the tube that connects the mouth to the stomach. Esophageal varices often happen in people with serious liver disease, called "cirrhosis." As a result of portal hypertension [Figure1], Cirrhosis is a medical term, usually used to describe a diseased liver that has been severely scarred injury. There are numerous causes of liver cirrhosis, the two most common causes of cirrhosis worldwide are alcoholic liver disease and viral hepatitis C, which together accounted for approximately one-half of patients on the liver transplantation wait list worldwide whatever portal hypertension can also be present in the absence of cirrhosis, a condition referred to as "noncirrhotic portal hypertension". The causes of non-cirrhotic portal hypertension can be divided into prehepatic, intrahepatic and post hepatic (presinusoidal, sinusoidal and post sinusoidal causes).

Authors α σ : Tropical Medicine Department, (Medical Informatics) Al-Azhar Faculty of Medicine-Egypt. E-mail : ahmadrazek@gmail.com

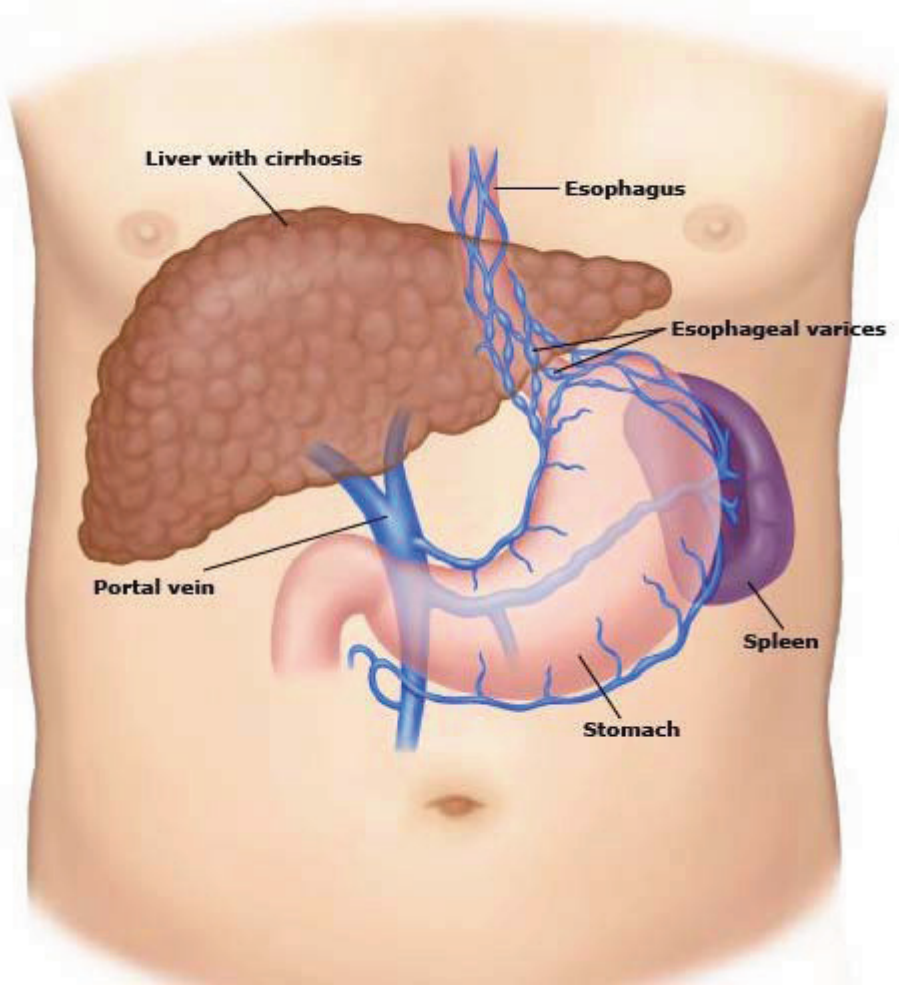


Figure 1 : Showing the esophageal varices interlacing the Esophagus and Upper part of the stomach, these varices can burst and cause internal bleeding, leading to death in many institutions.
With a permission from UpToDate®; Graphic 63611, Version 4.0

c) Diagnostic Techniques

Upper gastrointestinal endoscopy remains the gold standard for the diagnosis of esophagogastroduodenal lesions, despite its limitations. Early endoscopy (within 24 hours) is recommended for most patients with acute UGI bleeding, though whether early endoscopy affects outcomes and resource utilization is unsettled [7].

d) Non- Invasive Diagnoses

The identification of Cirrhotic patients with esophageal varices using non-invasive means has been attempted using different clinical, laboratory and radiological approaches, unfortunately many of these approaches still controversial ,recently non-invasive 2 D U/S can detect esophageal variceal degrees in cirrhotic patients with high accuracy, through measuring the intra-abdominal wall thickness of the esophagus, there were proportional relations between esophageal wall thicknesses detected by ultrasound and the esophageal variceal degrees diagnosed by Upper Endoscopy.

III. Analysis by Data Mining

The basic classification is based on supervised algorithm, whatever algorithms are applicable for the input data, the accuracy of each algorithm could be changed according to data nature. In clinical medicine, there are different laboratory, clinical and radiological factors determine the progress of each disease. Identification of important and risky factors is of great importance to predict out come in each disease stage. Analyzing of our data by data mining computed program shed light on the significant important factors for each disease condition. Thus the major challenge of biomedical data mining is to make these systems useful to biomedical researchers. The decision tree analysis was performed using Intelligent Miner software (Rapid Miner, Berlin, ver.4.6, Germany), which can automatically search a data set to find the optimal classification variables leading to the building of a decision tree algorithm . Briefly, all items derived from the patients were evaluated to determine which variables and cutoff points might produce the most

significant division into two subgroups; group with risky esophageal varices when esophageal wall thickness >6.5 mm and not risky group when esophageal wall thickness ≤ 6.5 mm.

IV. Validation Accuracy

In our previous study the algorithm was selected by evaluating each supervised machine learning algorithms by using supervised learning assessment (10-folds-cross validation) on the training set, we chose the best test applicable to our clinical data, accordingly we used Naïve-base test, and the overall accuracy we obtained was 95%.

We followed 59 patients presented with portal hypertension as a result of End stage-liver cirrhosis; the thicknesses of esophageal walls were measured using 2D U/S. All patients underwent diagnostic Esophag-ogastrodudonoscopy to estimate corresponding degrees of varices. According to the decision tree algorithm we obtained from the previous study, the esophageal wall thickness > 4.2 mm with inner wall irregularities, should be Esophageal varices with variable degrees, esophageal wall thickness <4.2mm measured by U/S without inner wall irregularities, indicating no varices, if the esophageal wall thickness > 6.5mm it should be risky esophageal varices (Grade III or IV varices), accordingly upper endoscopy should be performed early as a prophylactic intervention before further bleeding.

V. Results

(out of 59 patients presented with Liver cirrhosis and portal hypertension , only 53 patients had varices when esophageal wall thicknesses were > 4.2 and 6 patients had not varices when esophageal wall thicknesses were <4.2mm, all measured by U/S and diagnosed by upper Endoscopy; (100 % Comparative validation result for detection of Esophageal varices), whatever 48 patients had esophageal wall thicknesses > 6.5mm; risky esophageal varices, only 45 patients had risky esophageal varices diagnosed with Upper endoscopy; (93.75% Comparative validation result for detection of risky Esophageal varices).

Categorical data were compared using the χ^2 test, where as continuous variables were compared using Student's *t* test.

The overall comparative validation accuracy was 97.9 %. Table [1].

Esophageal wall Thickness	Cases	Result	Comparative
< 4.2 mm ; No varices	6	6	100%
>4.2 mm*; Varices	53	53	100%
>6.5 mm; Risky varices	48	45	93.75%
Over All Accuracy			97.9%

Table 1 : Describing the Comparative results between Analyses obtained by Data mining Computed Software; Rapid Miner ver.4.6 Berlin, Germany and results obtained from the current study

* Inner esophageal wall irregularies are confirmatory leading signs for esophageal varices with variable degrees.

VI. Conclusion

Based on the available evidence, prediction identification of esophageal variceal degrees in cirrhotic patients presented with manifestations of portal hypertension using non-invasive 2D ultrasound could be a very helpful tool saving time and money. Data mining shed light on the most significant predictors-related esophageal wall thicknesses in 673 patients with very high accuracy; identification high risky groups needed urgent interventional Endoscopy; less complication, better out come and decrease mortality.

Data mining would be the coming statistical evolution in clinical medical data statistical analysis with reasonable limitation.

VII. Limitation of the Study

Given the small sample size of 59 patients (8.7%) in comparable to 673 reported in the pervious study, the validation accuracy might be changed if we apply more patients in the future, whatever our clinical experience played a major role in assessing the information mentioned above, in our point of view our results should be confirmed with more evidence-based criteria using independent laboratory and clinical factors all together.

References Références Referencias

1. Crooks C, Card T, West J. Reductions in 28-day mortality following hospital admission for upper gastrointestinal hemorrhage. Gastroenterology 2011; 141:62.
2. Abd El Razek et al., Detection of Risky Esophageal varices by 2D U/S: when to perform Endoscopy. Am J Med Sci. 2012 Dec 21. [Epub ahead of print]. PMID: 23267234.
3. Cohen AM, Hersh WR, A survey of current work in biomedical text mining, Briefings in Bioinformatics, 2005, 6: 57-712.
4. Lanas A, Perez-Aisa MA, Feu F, et al. A nationwide study of mortality associated with hospital admission due to severe gastrointestinal events and those associated with nonsteroidal antiinflammatory drug use. Am J Gastroenterol 2005; 100:1685.

5. De Berardis G, Lucisano G, D'Ettorre A, et al. Association of aspirin use with major bleeding in patients with and without diabetes. JAMA 2012; 307:2286.
6. Wilcox CM, Alexander LN, Cotsonis G. A prospective characterization of upper gastrointestinal hemorrhage presenting with hematochezia. Am J Gastroenterol 1997; 92:231.
7. Balderas V, Bhore R, Lara LF, et al. The hematocrit level in upper gastrointestinal hemorrhage: safety of endoscopy and outcomes. Am J Med 2011; 124:970.
8. Kohonen T. Self-organizing maps. Berlin: Springer; 2001.
9. Talbi ML, Charef A. PVC discrimination using the QRS power spectrum and self-organizing maps. Comput Methods Programs Biomed. 2009; 94:223–231. doi: 10.1016/j.cmpb.2008.12.009.
10. Basara HG, Yuan M. Community health assessment using self-organizing maps and geographic information systems. Int. J Health Geogr. 2008; 30(7):67. doi: 10.1186/1476-072X-7-67.
11. Nobuaki Nakayama, Makoto Oketani, Yoshihiro Kawamura, Mie Inao, Sumiko Nagoshi, Kenji Fujiwara, Hirohito Tsubouchi, and Satoshi Mochida. Algorithm to determine the outcome of patients with acute liver failure: a data-mining analysis using decision trees. J Gastroenterol. 2012 June; 47(6): 664–677.
12. Balderas V, Bhore R, Lara LF, et al. The hematocrit level in upper gastrointestinal hemorrhage: safety of endoscopy and outcomes. Am J Med 2011; 124:970.
13. de Souza AR, La Mura V, Reverter E, et al. Patients whose first episode of bleeding occurs while taking a β-blocker have high long-term risks of rebleeding and death. Clin Gastroenterol Hepatol 2012; 10:670.
14. Asrani SK, Larson JJ, Yawn B, et al. Underestimation of liver-related mortality in the United States. Gastroenterology 2013; 145:375.
15. Groszmann RJ, Garcia-Tsao G, Bosch J, et al. Beta-blockers to prevent gastroesophageal varices in patients with cirrhosis. N Engl J Med 2005; 353:2254.
16. Jensen DM, Machicado GA. Diagnosis and treatment of severe hematochezia. The role of urgent colonoscopy after purge. Gastroenterology 1988; 95:1569.
17. Zuckerman GR, Trellis DR, Sherman TM, Clouse RE. An objective measure of stool color for differentiating upper from lower gastrointestinal bleeding. Dig Dis Sci 1995; 40:1614.
18. Parsonnet J, Hansen S, Rodriguez L, et al. Helicobacter pylori infection and gastric lymphoma. N Engl J Med 1994; 330:1267.
19. Pajares JM. H. pylori infection: its role in chronic gastritis, carcinoma and peptic ulcer. Hepatogastroenterology 1995; 42:827.
20. Shibata T, Imoto I, Ohuchi Y, et al. Helicobacter pylori infection in patients with gastric carcinoma in biopsy and surgical resection specimens. Cancer 1996; 77:1044.
21. Yang, Q., Xindong Wu, 10 Challenging Problems in Data Mining Research, International Journal of Information Technology & Decision Making, Vol. 5, No. 4 (2006) 597–604.
22. Lu S, Jin B, Cowart LA, Lu X. From Data towards Knowledge: Revealing the Architecture of Signaling Systems by Unifying Knowledge Mining and Data Mining of Systematic Perturbation Data. Plos One. 2013 Apr 23; 8(4):e61134.
23. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. Int J Med Inform. 2008; 77:81–97. doi: 10.1016/j.ijmedinf.2006.11.006
24. Udell JA, Wang CS, Tinmouth J, et al. Does this patient with liver disease have cirrhosis? JAMA 2012; 307:832.
25. Ghany MG, Nelson DR, Strader DB, et al. An update on treatment of genotype 1 chronic hepatitis C virus infection: 2011 practice guideline by the American Association for the Study of Liver Diseases. Hepatology 2011; 54:1433.
26. US Burden of Disease Collaborators. The state of US health, 1990-2010: burden of diseases, injuries, and risk factors. JAMA 2013; 310:591.
27. Cheung K, Lee SS, Raman M. Prevalence and mechanisms of malnutrition in patients with advanced liver disease, and nutrition management strategies. Clin Gastroenterol Hepatol 2012; 10:117.
28. Abd Elrazek Mohammad Ali Abdelrazek, Nakamura Yoko, Moriguchi Hiroki, Mohammad Afify,Badr Ismael, Mohammad Asar, Magdy Salah. Endoscopic management of Dieulafoy's lesion using Isoamyl-2-cyanoacrylate. 2013 Aug 16; 5(8):417-9. doi: 10.4253/ wjge.v5.i8.417.World Journal of Gastrointestinal Endoscopy. PMID: 23951399 [PubMed].

Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 13 Issue 10 Version 1.0 Year 2013
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

A Novel Approach for Scalability - Two Way Sequential Pattern Mining using UDDAG

By Dr. P. Raguraman, Mr. S. Hariharan & Dr. J. Jaya A Celin
Noorul Islam University, India

Abstract - Traditional pattern growth-based approaches for sequential pattern mining derive length- (k + 1) patterns based on the projected databases of length-k patterns recursively. At each level of recursion, they unidirectionally grow the length of detected patterns by one along the suffix of detected patterns, which needs k levels of recursion to find a length-k pattern. In this paper, a novel data structure, UpDown Directed Acyclic Graph (UDDAG), is invented for efficient sequential pattern mining. UDDAG allows bidirectional pattern growth along both ends of detected patterns. Thus, a length-k pattern can be detected in $\lfloor \log_2 k + 1 \rfloor$ levels of recursion at best, which results in fewer levels of recursion and faster pattern growth. When minSup is large such that the average pattern length is close to 1, UDDAG and PrefixSpan have similar performance because the problem degrades into frequent item counting problem. However, UDDAG scales up much better. It often outperforms PrefixSpan by almost one order of magnitude in scalability tests. UDDAG is also considerably faster than Spade and LapinSpam. Except for extreme cases, UDDAG uses comparable memory to that of PrefixSpan and less memory than Spade and LapinSpam. Additionally, the special feature of UDDAG enables its extension toward applications involving searching in large spaces.

Indexterms : *data mining algorithm, directed acyclic graph, performance analysis, sequential pattern, transaction database.*

GJCST-C Classification : *H.2.8*

Strictly as per the compliance and regulations of:

A Novel Approach for Scalability – Two Way Sequential Pattern Mining using UDDAG

Dr. P. Raguraman [α], Mr. S. Hariharan [σ] & Dr. J. Jaya A Celin [ρ]

Abstract - **Traditional pattern growth-based approaches for sequential pattern mining derive length- (k + 1) patterns based on the projected databases of length-k patterns recursively. At each level of recursion, they unidirectionally grow the length of detected patterns by one along the suffix of detected patterns, which needs k levels of recursion to find a length-k pattern. In this paper, a novel data structure, UpDown Directed Acyclic Graph (UDDAG), is invented for efficient sequential pattern mining. UDDAG allows bidirectional pattern growth along both ends of detected patterns. Thus, a length-k pattern can be detected in $\lceil \log_2 k + 1 \rceil$ levels of recursion at best, which results in fewer levels of recursion and faster pattern growth. When minSup is large such that the average pattern length is close to 1, UDDAG and PrefixSpan have similar performance because the problem degrades into frequent item counting problem. However, UDDAG scales up much better. It often outperforms PrefixSpan by almost one order of magnitude in scalability tests. UDDAG is also considerably faster than Spade and LapinSpam. Except for extreme cases, UDDAG uses comparable memory to that of PrefixSpan and less memory than Spade and LapinSpam. Additionally, the special feature of UDDAG enables its extension toward applications involving searching in large spaces.**

Indexterms : data mining algorithm, directed acyclic graph, performance analysis, sequential pattern, transaction database.

I. Introduction

SEQUENTIAL pattern mining is an important data mining problem, which detects frequent subsequences in a sequence database. A major technique for sequential pattern mining is pattern growth. Traditional pattern growth-based approaches (e.g., PrefixSpan) derive length-(k + 1) patterns based on the projected databases of a length-k pattern recursively. At each level of recursion, the length of detected patterns is grown by 1, and patterns are grown unidirectionally along the suffix direction. Consequently, we need k levels of recursion to mine a length-k pattern, which is expensive due to the large number of recursive database projections. In this paper, a new approach based on UpDown Directed Acyclic Graph (UDDAG) is proposed for fast pattern growth. UDDAG is a novel data structure, which supports bidirectional pattern growth from both ends of detected patterns. With UDDAG, at level i recursion, wemay grow the length of patterns by 2i_1 at most. Thus, a length-k pattern can be detected in $\lceil \log_2 k + 1 \rceil$ levels of recursion at minimum, which results in better scale-up property for UDDAG compared to PrefixSpan. Our extensive experiments clearly demonstrated the strength of UDDAG with its bidirectional pattern growth strategy. When minSup is very large such that the average length of patterns is very small (close to 1), UDDAG and PrefixSpan have similar performance because in this case, the problem degrades into a basic frequent item counting problem. However, UDDAG scales up much better compared to PrefixSpan. It often outperforms PrefixSpan by one order of magnitude in our scalability tests. UDDAG is also considerably faster than two other representative algorithms, Spade and LapinSpam. Except for some extreme cases, the memory usage of UDDAG is comparable to that of PrefixSpan. UDDAG generally uses less memory than Spade and LapinSpam. UDDAG may be extended to other areas where efficient searching in large searching spaces is necessary.

II. Related Work

The problem of sequential pattern mining was introduced by Agrawal and Srikant [1]. Among the many algorithms proposed to solve the problem, GSP [17] and PrefixSpan[13], [14] represent two major types of approaches: a prioribased and pattern growth-based. A priori principle states that any supersequence of a nonfrequent sequence must not be frequent. A priori based approaches can be considered as breadth-first traversal algorithms because they construct all length-k patterns before constructing length-(k+1) patterns. The AprioriAll algorithm [1] is one of the earliest a prioribased approaches. It first finds all frequent item sets, transforms the database so that each transaction is replaced by all frequent item sets it contains, and then finds patterns. The GSP algorithm [16] is an improvement over AprioriAll. To reduce candidates, GSP only creates a new length-k candidate when there are two frequent length-(k _ 1) sequences with the prefix of one equal to the suffix of the other. To test whether a candidate is a frequent length-k pattern, the support of each length-k candidate is counted by examining all the sequences. The PSP algorithm [12] is similar to GSP except that the placement of candidates is improved through a prefix tree arrangement to speed up pattern

Authors α σ : Assistant Professor, Department of Computer Science, Sri Sankara Arts and Science College, Enathur, Kanchipuram, TN, India. E-mail : contactragu84@gmail.com
Author ρ : Associate Professor, Department of IT Noorul Islam University, India.

discovery. The SPIRIT algorithm [9] uses regular expressions as constraints and developed a family of algorithms for pattern mining under constraints based on a priori rule. The SPaRSe algorithm [3] improves GSP by using both candidate generation and projected databases to achieve higher efficiency for high pattern density conditions.

III. Problem Definition

a) Updown Directed Acyclic Graph-Based Sequential Pattern Mining

UDDAG-based pattern mining approach, which first transforms a database based on frequent item sets, then partitions the problem, and finally, detects each subset using UDDAG. The absolute support for an item set in a sequence database is the number of tuples whose sequences contain the item set. An item set with a support larger than minSup is called a frequent item (FI) set. Based on frequent item sets, we transform each sequence in a database D into an alternative representation.

i. *Transformed Database Definition*

Let D be a database and P be the complete set of sequential patterns in D, D' be its transformed database, substituting the ids of each item pattern contained in D' with the corresponding item sets, and denoting the resulted pattern set by P', we have P = P'.

Based on frequent item sets, we transform each sequence in a database D into an alternative representation. Steps involved in Database Transformation:

1. Find the set of frequent items in D.
2. Assign a unique id to each FI in D and then replace each item set in each sequence with the ids of all the FIs contained in the item set.

Seq. Id	Sequence
1	<1 (1,2,3) (1,3) 4 (3,6)>
2	<(1,4) 3 (2,3) (1,5)>
3	<(5,6) (1,2) (4,6) 3 2>
4	<5 7 (1,6) 3 2 3>

Table 1 : Sequence Database

Seq. Id	Sequence
1	<1 (1,2,3,4,5) (1,5) 6 (5,8)>
2	<(1,6) 5 (3,4,5) (1,7)>
3	<(7,8) (1,2,3) (6,8) 5 3>
4	<7 (1,8) 5 3 5>

Table 2 : Transformed Database

For the database in Table 1, **the FIs are: (1),(2), (3), (4), (5), (6), (1,2), (2,3).**

By assigning a unique id to each FI, e.g., (1)-1, (1,2)-2, (2)-3, (2,3)-4, (3)-5, (4)-6, (5)-7, (6)-8, we can transform the database as shown in Table 2 (infrequent items are eliminated).

ii. *Problem Partitioning Definition*

Let {x1, x2, . . . , xt} be the frequent item sets in a database D, x1 < x2 <. . . . < xt, the complete set of patterns (P) in D can be divided into t disjoint subsets. The ith subset (denoted by $P_{xi,}$ 1 < = i <= t) is the set of patterns that contains xi and FIs smaller than xi.

iii. *Definition (Projected database)*

The collection of all the tuples whose sequences contain an item set x in a database D is called x-projected database, denoted by ^{x}D.

The total number of different Frequent Items FIs in the Sequential Database are found by Database Transformation module. For a database with n different frequent items, its patterns can be divided into n disjoint subsets. The ith subset (1 < i < n) is the set of patterns that contain i (the root item of the subset) and items smaller than i.Each subset i of the problem is mapped into a projected database denoted by (^{i}D).

P is partitioned into eight subsets as there are 8 FIs in table-2, the one contains 1 (P1), the one contains 2 and smaller ids (P2), . . . ,and the one contains 8 and smaller ids (P8).

Given the following database (P8) alone is found as given by ^{8}D:

1. <9 4 5 8 3 6>
2. <3 9 4 5 8 3 1 5>
3. <3 8 2 4 6 3 9>
4. <2 8 4 3 6>
5. <9 6 3>

^{8}D is,

1. <4 5 8 3 6>
2. <3 4 5 8 3 1 5>
3. <3 8 2 4 6 3>
4. <2 8 4 3 6>

b) UpDown Directed Acyclic Graph

i. *Definition*

Directed acyclic graph (UDDAG) is a graphical approach that represents patterns as vertices and contain relationships as directed edges in between vertices.

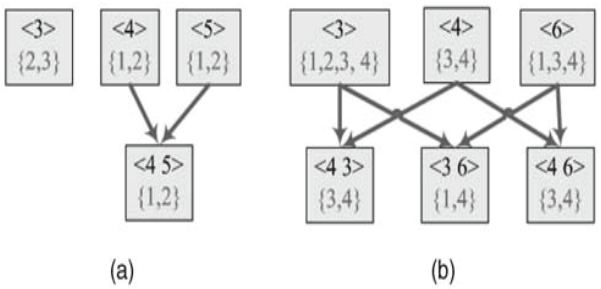
Given an FI x and ^{x}D, an UpDown Directed Acyclic Graph based on P_x, denoted by x-UDDAG, is constructed as follows:

1. Each pattern in P_x corresponds to a vertex in x-UDDAG. <x> corresponds to the root vertex, denoted by V_x

2. Let P_U be the set of length-2 patterns ending with x in P_x . Add a directed edge from V_x to U_v called up root child of V_x. This represents the common prefix set to root X.
3. Let P_D be the set of length-2 patterns starting with x in P_x, add a directed edge from V_x to V_D called a down root child of V_x. This represents the common suffix set to root X.

Such a DAG can be recursively constructed in an efficient way to derive the contain relationship of patterns

Example of UDDAG Construction



By concatenating the patterns {<3>,<4>,<5>,<4 5>} of Pre(^{8}D) with 8, we derive patterns {<3 8>, <4 8>, <5 8>, <4 5 8>} in figure-a.

By concatenating patterns {<3>, <4>, <6>, <3 6>, <4 3>, <4 6>} of Suf(^{8}D) with 8,we derive patterns {<8 3>, <8 4>, <8 6>, <8 3 6>, <8 4 3>, <8 4 6>} in figure-b.

iii. *UDDAG based Pattern Mining*

In the ith subset, each pattern in the projected database (^{x}D) can be divided into two parts, prefix and suffix of i.

The collection of all the prefix/suffix tuples of a frequent item set X in ^{x}D is called the prefix/suffix-projected database of x, denoted by Pre(^{x}D) / Suf(^{x}D).

To detect the sequential pattern in projected database (^{x}D) P_x,

- Detect patterns in Pre(^{x}D) called pattern prefix (PP).
- Detect pattern in Suf(^{x}D) called pattern suffix (PS)
- The above steps are repeated recursively until no frequent items are found in the pre(^{x} D) / suf(^{x} D).
- Combine the patterns of all the iterations to derive P_x.

The complete set of patterns is the union of patterns of the all subsets or projected database (^{x}D) detected above.

The Apriori property is used to reduce the number of candidate sets to be considered

Example for Pattern Mining

Assuming ^{8}D is,

1. <4 5 8 3 6>
2. <3 4 5 8 3 1 5>
3. <3 8 2 4 6 3>
4. <2 8 4 3 6>.

The prefix subsequences of 8 in ^{8}D, or Pre(^{8}D) is :**{<3 8>, <4 8>, <5 8>, <4 5 8>}** the patterns with 8 at the end.

The suffix subsequences of 8 in ^{8}D, or Suf(^{8}D) is : **{<8 3>, <8 4>,<8 6>, <8 3 6>, <8 4 3>, <8 4 6>}** the patterns with 8 at the beginning.

The patterns with 8 in between the beginning and end of each pattern is: **{<3 8 3>,<4 8 3>, <5 8 3>, <4 5 8 3>}**

Example: UDDAG based Pattern Mining

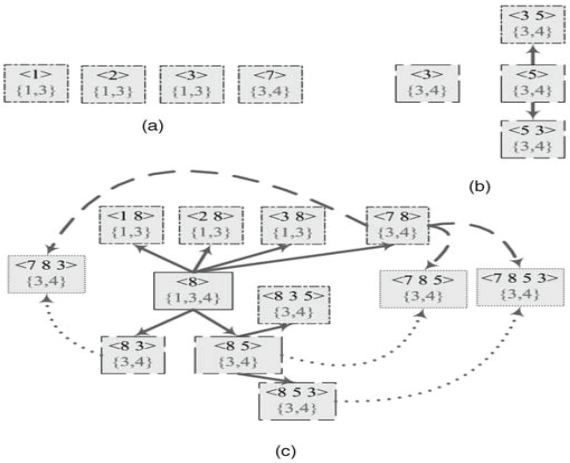


Figure : (a) UP DAG, (b) DOWN DAG & (c) UPDOWN DAG

Algorithm 1 : UDDAG based pattern Mining.
Input : A database D and the minimum support
Output : P, the complete set of patterns in D
Method : findP (D, minSup){
P = Φ
FISet=D:getAllFI(minSup);
D.transform( );
for each FI x in FISet{
UDVertexrootVT = newUDVertex(x)
findP(D.getPreD(x), rootVT, up, minSup)
findP(D.getSurD(x),rootVT,down,minSup)
findPUDDAG(rootVT)
P = P U rootVT.getAllPatterns( )
}
}

Subroutine
findP(PD,rootVT,type, minSup){
FISet=PD.getAllFI(minSup);
for each FI x in FISet{
UDVertexcurVT=new DVertex(x, rootVT)
if(type==up) rooVT.addUpChild(curVT)
else rootVT.addDownChild(curVT)
findP(PD. getPreD(x), curVT, up, minSup)
findP(PD.getSufD(x),curVT,down,minSup)
findPUDDAG(curVT)
}
}

Subroutine

```
findPUDDAG(rootVT){
upQueue.enQueue (rootVT.upChildren)
while(!upQueue.isEmpty()){
UDVertex upVT=upQueue.deQueue()
if(upVT.upParent==rootVT)
downQueue.enQueue(rootVT.downChildren)
else if (upVT.downParent==null)
downQueue.enQueue(upVT.upParent.VDVS)
else downQueue.enQueue(upVT.upParent.VDVS n
upVT.downParent.VDVS)
while(!downQueue.isEmpty()){
UDVertexdownVT=downQueue.deQueue()
if(isValid(upVT, downVT){
UDVertexcurVT=new UDVertex (upVT, downVT)
upVT.addVDVS(downVT)
if(upVT.upParent==rootVT)
downQueue.enQueue(downVT.children)
}
}
if(upVT.VDVS.size>0)upQueue.enQueue(upVT.children)
}
}
```

IV. Performance Evaluation

We conducted an extensive set of experiments to compare our approach with other representative algorithms. All the experiments were performed on a windows Server 2003 with 3.0 GHz Quad Core Intel Xeon Server and 16 GB memory. The algorithms we compared are PrefixSpan, Spade, and LapinSpam, which were all implemented in C++ by their authors (Minor changes have been made to adapt Spade to Windows). Two versions of UDDAG were tested. UDDAG-bv uses bit vector to verify candidates and UDDAG-co uses co-occurrences to verify candidates whenever possible. We perform two studies using the same data generator as in [14]: 1) Comparative study, which uses similar data sets as that in [14]; 2) Scalability study. The data sets were generated by maintaining all except one of the parameters as shown in Table 4 fixed, and exploring different values for the remaining ones.

V. Conclusion

In this paper, a novel data structure UDDAG is invented for efficient pattern mining. The new approach grows patterns from both ends (prefixes and suffixes) of detected patterns, which results in faster pattern growth because of less levels of database projection compared to traditional approaches. Extensive experiments on both comparative and scalability studies have been performed to evaluate the proposed algorithm.

One major feature of UDDAG is that it supports efficient pruning of invalid candidates. This represents a promising approach for applications involving searching in large spaces. Thus, it has great potential to related areas of data mining and artificial intelligence. In the future, we expect to further improve UDDAG-based pattern mining algorithm as follows: 1) Currently, FI detection is independent from pattern mining. Practically, the knowledge gained from FI detection may be useful for pattern mining. In the future, we will integrate the solutions of the two so that they can benefit from each other. 2) Different candidate verification strategies may have different impacts to the efficiency of the algorithm. In the future, we will study more efficient verification strategy. 3) UDDAG has big impact to the memory usage when the number of patterns in a subset is extremely large. In the future, we will find an efficient way to store UDDAG.

References Références Referencias

1. R. Agrawal and R. Srikant, "Mining Sequential Patterns," Proc. Int'l Conf. Data Eng. (ICDE '95), pp. 3-14, 1995. CHEN: AN UPDOWN DIRECTED ACYCLIC GRAPH APPROACH FOR EQUENTIAL PATTERN MINING 927
2. R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," Proc. 20th Int'l Conf. Very Large Data Bases (VLDB), pp. 487-499, 1994.
3. C. Antunes and A.L. Oliveira, "Generalization of Pattern-Growth Methods for Sequential Pattern Mining with Gap Constraints," Proc. Int'l Conf. Machine Learning and Data Mining 2003, pp. 239-251, 2003.
4. J. Ayres, J. Gehrke, T. Yu, and J. Flannick, "Sequential Pattern Mining Using a Bitmap Representation," Proc. Int'l Conf. Knowledge Discovery and Data Mining 2002, pp. 429-435, 2002.
5. S. Berkovich, G. Lapir, and M. Mack, "A Bit-Counting Algorithm Using the Frequency Division Principle," Software: Practice and Experience, vol. 30, no. 14, pp. 1531-1540, 2000.
6. J. Chen and T. Cook, "Mining Contiguous Sequential Patterns from Web Logs," Proc. World Wide Web Conf. (WWW '07) Poster Session, May 2007.
7. J. Chen and K. Xiao, "BISC: A Binary Itemset Support Counting Approach Towards Efficient Frequent Itemset Mining," to be published in ACM Trans. Knowledge Discovery in Data.
8. G. Grahne and J. Zhu, "Efficiently Using Prefix-Trees in Mining Frequent Itemsets," Proc. Workshop Frequent Itemset Mining Implementations (FIMI '03), 2003.
9. M. Garofalakis, R. Rastogi, and K. Shim, "SPIRIT: Sequential Pattern Mining with Regular Expression Constraints," Proc. Int'l Conf. Very Large Data Bases (VLDB '99), pp. 223-234, 1999.
10. J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M.C. Hsu, "FreeSpan: Frequent Pattern-Projected Sequential Pattern

Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 13 Issue 10 Version 1.0 Year 2013
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Software Effort Prediction - A Fuzzy Logic Approach

By Sanjay Kumar, Jaya Pal & Vandana Bhattacherjee
Birla Institute of Technology, India

Abstract - Accuracy in the estimation of software Effort/Cost is one of the desirable criteria for any software cost estimation model. The estimation of effort or cost before the actual development of any software is the most crucial task of the present day software development project managers. Software project attributes are often measured in terms of linguistic values such as very low, low, Average, high and very high. The imprecise nature of such attributes constitutes uncertainty and vagueness in their subsequent interpretation. In this paper we propose a Fuzzy logic based model for software effort prediction. We feel that fuzzy Software cost estimation Model should be able to deal with imprecision and uncertainty associated with various parameter values. Fuzzy analogy model has been developed and validated upon student data.

Keywords : *software cost estimation, effort prediction, fuzzy logic, linear regression.*

GJCST-C Classification : *D.2.8*

Strictly as per the compliance and regulations of:

Software Effort Prediction - A Fuzzy Logic Approach

Sanjay Kumar [α], Jaya Pal [σ] & Vandana Bhattacherjee [ρ]

Abstract - **Accuracy in the estimation of software Effort/Cost is one of the desirable criteria for any software cost estimation model. The estimation of effort or cost before the actual development of any software is the most crucial task of the present day software development project managers. Software project attributes are often measured in terms of linguistic values such as very low, low, Average, high and very high. The imprecise nature of such attributes constitutes uncertainty and vagueness in their subsequent interpretation. In this paper we propose a Fuzzy logic based model for software effort prediction. We feel that fuzzy Software cost estimation Model should be able to deal with imprecision and uncertainty associated with various parameter values. Fuzzy analogy model has been developed and validated upon student data.**

Keywords : software cost estimation, effort prediction, fuzzy logic, linear regression.

I. Introduction

Accurate and timely prediction of the development effort and schedule required to develop a software system is one of the most critical activities in managing software projects. In addition software estimation has been identified as one of the three great challenges for half-century-old computer science. [19] In the last 30 years many different studies have been done in the area of Software Cost Estimation to improve the estimation accuracy and so many models are introduced. The rest of the paper contains the following sections as follows: section II represents Research Method, section III represents Experimental Results, and section IV represents Conclusion and Future Scope.

a) Fuzzy Logic

Intelligent Systems provide alternative paradigms aimed at facilitating the representation and manipulation of uncertain, incomplete, imprecise or noisy data. Fuzzy logic is a form of many-valued logic or probabilistic logic; it deals with reasoning that is approximate rather than fixed and exact.

The traditional approach to building system controllers requires a prior model of the system. The quality of the model, that is, loss of precision from linearization and/or uncertainties in the system's parameters negatively influences the quality of the resulting control. It is well known that the fuzzy theory not only provides natural tool for describing quantitative data but also generally produces good performance in many applications. In addition, fuzzy rules allow us to effectively classify data having non-axis-parallel decision boundaries, which is difficult for the conventional attribute-based methods. However, one of the difficulties with fuzzy decision trees is determining an appropriate set of membership functions representing fuzzy linguistic terms. Usually membership func-tions are given manually, however, it is difficult for even an expert to determine an appropriate set of membership functions when the volume and dimensionality of data are large.

At the same time, methods of soft computing such as fuzzy logic possess non-linear mapping capabilities, do not require an analytical model and can deal with uncertainties in the system's parameters. Although fuzzy logic deals with imprecise information, the information is processed in sound mathematical theory [40]. Based on the nature of fuzzy human thinking, Lofti Zadeh originated the "fuzzy logic" or "fuzzy set theory", in 1965. Fuzzy logic deals with the problems that have fuzziness or vagueness. In fuzzy set theory based on fuzzy logic a particular object has a degree of membership in a given set that may be anywhere in the range of 0 (completely not in the set) to 1 (completely in the set) [41].

For this reason fuzzy logic is often defined as multi-valued logic (0 to 1), compared to bi-valued Boolean logic [42]. Specifically, Fuzzy Logic offers a particularly convenient way to generate a keen mapping between input and output spaces thanks to fuzzy rules' natural expression [2]. Fuzzy logic has been used in[36][37][38][39]. Fuzzy set theory and fuzzy logic are a highly suitable and applicable basis for developing knowledge-based systems This paper presents a fuzzy rule based system having two fuzzy inputs, namely Line of code (LOC) and Adjusted difficulty level (Adj.diff.level) and one output Development time (Devtime) as shown in Figure 1.

Authors α σ ρ : Department of CSE, Birla Institute of Technology Ranchi, India. E-mails : k_sanjay71@yahoo.co.in, jayapal@bitmesra.ac.in, vbhattacharya@bitmesra.ac.in

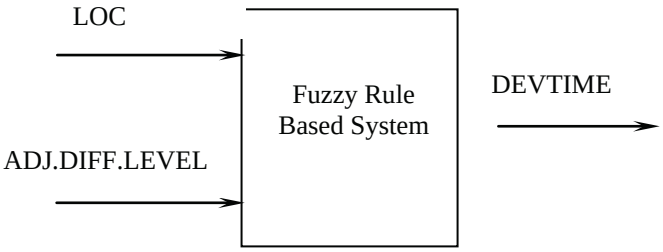
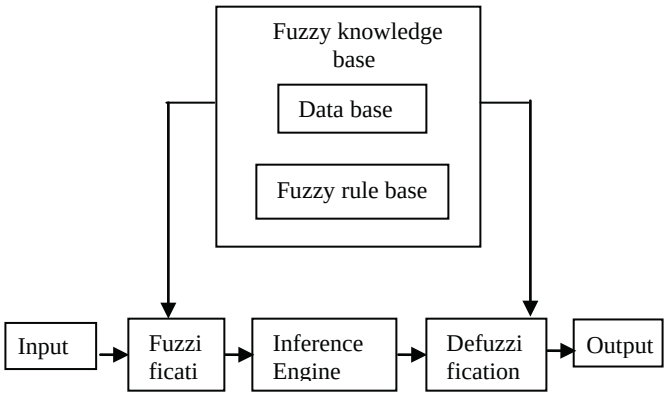


Figure 1 : Fuzzy Rule Based System

b) Fuzzy Rule Based System

A typical fuzzy logic system consists of four main components as shown in Figure.



i. *Fuzzification*

That contains predefined set of linguistic values. It converts non-fuzzy (Deterministic) inputs of fuzzy system into fuzzy inputs for inference mechanism.

ii. *Knowledge Base*

That consists of two parts: database that defines linguistic variables conditional statements (Fuzzy sets, and rule base that represents the mapping of fuzzy input set into a fuzzy output set. Rules are fuzzy implications). Fuzzy sets, and rule base that represents the mapping of fuzzy input set into a fuzzy output set. Rules are fuzzy conditional statements (implications).

iii. *Decision Logic*

That simulates human decision making based on fuzzy concepts. Conclusion of certain condition is derived by decision making logic.

iv. *Defuzzification*

That converts rule base fuzzy outputs into non-fuzzy (.numerical) values.

Central mechanism of knowledge base and decision making logic considers the fuzzy extension of conventional rule inferencing concept to fuzzy rules inferencing. Premises and conclusions of rules now contain fuzzy values. These facts by definition describe practically continual input set of characteristics. In this manner, one rule can replace more conventional rules. Fuzzy inferencing rules generally connect m conditional variables X1,..., Xm to n consequent variables Y1,..., Yn in form of:

IF (X1 is A1 and ……..Xm is Am) THEN (Y1 is B1 and ... Yn is Bn).

Where *A1,..., Am* and *B1,..., Bn* are linguistic terms of linguistic variables *X1,...,Xm* and *Y1,..., Yn,* respectively.

The IF part is called the "antecedent" and the THEN part is called the "consequent". To make a decision based on a set of rules, a rules-based system follows these steps:

1. All the rules that apply are invoked, using the membership functions and truth values obtained from the inputs (by a process called fuzzification), to determine the result of the antecedent.
2. This result in turn will be mapped into a membership function and truth value controlling the output variable. This process is known as implication. Two of the more common implication functions are: clipping (the fuzzy set is clipped to a value given by the level of activation of the input variables) and scaling (the fuzzy set is multiplied by a value given by the level of activation of the input variables).
3. These results are combined by a process called aggregation. One common approach for the aggregation involves using the "maximum" of the implicated sets.
4. Finally, a process known as defuzzification is used to compute a single value that is representative of the aggregated fuzzy set.

c) Multiple Regressions

A linear equation with three independent variables (multiple regressions) may be expressed as:

$$y = b_0 + b_1x_1 + b_2x_2 \quad (1)$$

Where b_0, b_1, and b_2 are constants; x_1, and x_2 are the independent variables, and y is the dependent variable. The values of b_0, b_1, and b_2 of the multiple regression. Equation may be obtained solving the following system of linear equations

$$\Sigma y = nb_0 + b_1(\Sigma x_1) + b_2(\Sigma x_2) \quad (2)$$

$$\Sigma x1y = b_0(\Sigma x_1) + b_1(\Sigma_{x1}{}^2) + b_2(\Sigma x_1x_2) \quad (3)$$

$$\Sigma x_2y = b_0(\Sigma x_2) + b_1(\Sigma x_1x_2) + b_2(\Sigma x_2{}^2) \quad (4)$$

d) Evaluation Criteria

A common criterion for the evaluation of software effort models is the Magnitude of Relative Error (MRE) which is defined as follows:

$$MRE = \left| \frac{\text{Actual devtime–Predicted devtime}}{\text{Actual devtime}} \right|$$

The MRE value is calculated for each observation whose devtime is predicted. The

aggregation of MRE over multiple observations (N) can be achieved through the Mean MRE (MMRE) as follows:

$$MMRE = \frac{1}{N} \sum_{i}^{N} MRE$$

A complementary criterion is the prediction at level *l*, Pred *(l)* = k/N, where *k* is the number of observations where MRE is less than or equal to *l,* and *N* is the total number of observations. Thus, Pred (25) gives the percentage of development time of software which were predicted with a MRE less or equal than 0.25.

II. Research Method

a) Metrics Used

The following metrics have been used Line of Code (LOC), and Adjusted Difficulty Level (ADJ.DIFF.LEVEL) which is served as input to the Fuzzy Logic System. And one output Development Time (DEVTIME).

Description of Metrics

1. *Line of Code (LOC):* Loc is the total number of lines of code used to develop the software excluding the comment lines. This metric was measured on the scale of 0-60.
2. *Adjusted Difficulty Level (ADJ.DIFF.LEVEL):* This is the difficulty level of the programmers to develop the software which is further adjusted with the help of expert judgments. This metric was measured on the scale of 0-6.
3. *Development Time (Devtime):* It is the time taken to develop the software. This metric was measured on the scale of 0-24.

b) Data Gathered

The proposed model was validated by a data set collected from the BIT, students of MCA. This data set consists of 10 project data. The data set is applied to the proposed fuzzy model is shown in the Table 1.

Table 1

PROG_ID	LOC	ADJ.DFF.LEVEL	DEVTIME(ACT)
1	16	1.5	3
2	32	1.8	4
3	23	2.3	10
4	18	2.8	15
5	16	1.5	3
6	32	1.3	4
7	16	1.3	4
8	18	1.4	4
9	19	1.8	7
10	22	1.9	5

c) Fuzzy Rules

The term fuzzy identification usually refers to the techniques and algorithms for constructing fuzzy models from data. The expert knowledge in a verbal form is translated into a set of if–then rules. A certain model structure can be created, and parameters of this structure, such as membership functions and weights of rules, can be tuned using input and output data.

This paper is based on five fuzzy rules as follows:

1. If ADJ.DIFF.LEVEL is (Very Low) then (DEVTIME (Very Low). (1).
2. If (loc is Low) and (ADJ.DIFF.LEVEL is (Average)) then (DEVTIME is (Low)) (1).
3. If (loc is Average) and (ADJ.DIFF.LEVEL is (Average)) then (DEVTIME is (Average)) (1).
4. If (loc is High) and (ADJ.DIFF.LEVEL is (Average)) then (DEVTIME is (High)) (1).
5. If (loc is Very High) and (ADJ.DIFF.LEVEL is (High)) then (DEVTIME is (Very High)) (1).

The weight of all the rule is 1.

Input and output Membership Functions (MF) are depicted in Table 2. All are triangular and their scalar parameters (a, b, c) are defined as follows:

Table 2 : Membership Function Characteristics

Variable Name	Range	MF	Parameters		
			a	b	c
LOC	0-60	VL	0	10	20
		L	10	20	30
		AV	20	30	40
		H	30	40	50
		VH	40	50	60
ADJ.DIFF.LEVEL	0-6	VL	0	1	2
		L	1	2	3
		AV	2	3	4
		H	3	4	5
		VH	4	5	6

Output

Variable Name	Range	MF	Parameters		
			a	b	c
DEVTIME	0-24	VL	0	4	8
		L	4	8	12
		AV	8	12	16
		H	12	16	20
		VH	16	20	24

Where MF is membership function.

The membership function plots corresponding to Table 2 are shown in figures2 (a), 2(b) and 2(c).

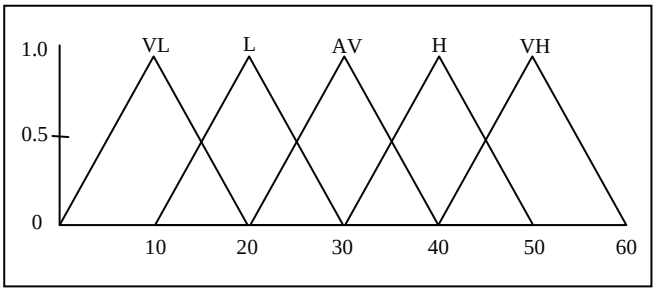


Figure 2(a) : LOC *(Input)*

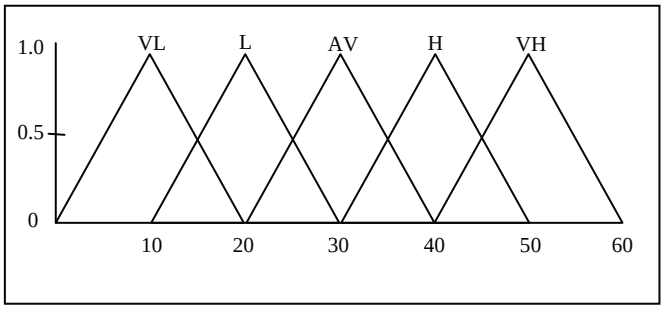


Figure 2(b) : ADJ.DIFF.LEVEL *(input)*

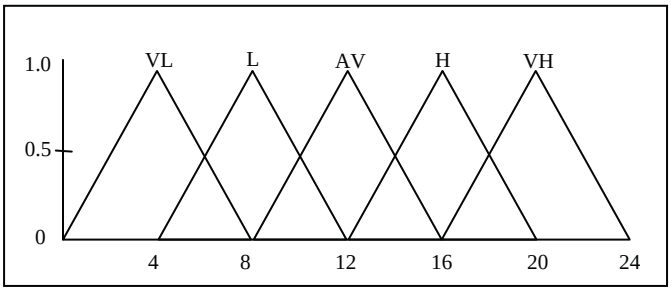


Figure 2(c) : DEVTIME *(output)*

d) Multiple Regressions

The same dataset has been used in Multiple Regression model to estimate the development time, which is to be used to develop a software. The estimated value of development time is compared with the actual value of development time, and with the help of this, using the evaluation criterion the MRE, MMRE and the Pred(25) value is also calculated.

e) Evaluation criteria

For this model the same evaluation criterion is used. The criterion which is used to evaluate the fuzzy model.

III. Experimental Results

The results show more accuracy in case of effort estimation by the proposed fuzzy model. The result is shown in the Table 3, Table 4 and Table 5.

Table 3

PROG_ID	DEVTIME (ACT)	DEVTIME (PRED) CALCULATED USING FUZZY LOGIG	MRE
1	3	4	0.3333
2	4	4	0.0000
3	10	10	0.0000
4	15	8	0.4667
5	3	4	0.3333
6	4	4	0.0000
7	4	4	0.0000
8	4	4	0.0000
9	7	4	0.4285
10	4	4	0.0000

Table 4

PROG_ID	DEVTIME (ACT)	DEVTIME (PRED) CALCULATED USING LINEAR REGRESSION	MRE
1	3	4.2048	0.4016
2	4	5.7181	0.4295
3	10	9.8191	0.1808
4	15	13.7431	0.0838
5	3	4.2048	0.4016
6	4	2.0155	0.4961
7	4	2.7237	0.3191
8	4	3.3757	1.5606
9	7	6.2935	1.0092
10	4	6.9013	0.4753

Table 5 : Prediction Results

	Multiple Regression	Fuzzy Logic
Min(MRE)	0.0838	0.0000
Max(MRE)	1.5606	0.4667
MMRE	0.5358	0.1762
Pred(25)	0.3	0.6

We have compared the actual development time with the predicted development time given by the model for each data set and found that difference between the actual devtime and predicted devtime. Then we calculated the MRE of each project and MMRE =0.1762 and pred(0.25%) which is 0.6. The same dataset has been tested using multiple regression model and the calculated MMR was used to further find out the MMRE=0.5358 and Pred(0.25%)=0.3. after going through the results we conclude that the proposed fuzzy model gives the better accuracy.

IV. Conclusion and Future Scope

The main benefit of this model is its good interpretability by using fuzzy rules and another great advantage of this research is that it can put together expert knowledge (Fuzzy rules) and project data into one general framework that may have a wide range of applicability in software estimation.

Further the comparison with multiple regression model to fuzzy logic model, and the results support the fuzzy logic model.

In our future work we will test this model upon different real-time datasets. Datasets have been collected from the software engineering data repository [34]. These datasets have been used by various researchers in their work [4] [34] [35].

V. Acknowledgment

This research work has been partially funded by UGC [F.No.:33-61/2007(SR)] under financial grants for Major Research Project.

References Références Referencias

1. Ali Idri, Alain Abran, Laila Kjiri(2000), "COCOMO Cost Model Using Fuzzy Logic" in Proc. 7th Intl. Conf. On Fuzzy theory & Technology, 2000.
2. Ali Idri, Alain Abran, T.M. Khoshgoftaar (2002), "Fuzzy Case-Based Reasoning Models for Software Cost Estimation", to appear in the book Soft Computing in Software Engineering: Theory and Aplications Published by Springer- Verlag, 2002.
3. Imran Attarzadeh, Siew Hock Ow (2010), "Soft Computing Approach for Software Cost Estimation", In international Journal of Software Engineering, IJSE Vol.3 No.1 January 2010.
4. M. Shepperd and C. Schofield. (1997), "Estimating Software Project Effort Using Analogies", IEEE Transactions on Software Engineering, vol. 23, no. 12, pp. 736-743, November 1997.
5. Sarah Jane delany, Padraig Cunningham and Wolfgang Wilke (1998), "The Limits of CBR in Software Projects Estimation". Presented at 1998 German Workshop on Case-Based Reasoning.
6. Magne Jargensen (2005), "Practical Guidelines for Expert-Judgment-Based Software Effort Estimation", In Proceedings of the 7th ACIS International Conference on Software Engineering Artificial Intelligence, Networking, and Parallel/Distributed Computing (SNPD'06), pp. 57-64, IEEE Computer Society.
7. A. Aamodt and E. Plaza (1994), "Case-Based Reasoning: Foundational Issues, Methodological Variations and System Approaches", AI Communications, IOS Press, vol. 7:1, pp. 39-59.
8. Hareton Leung, Zhang Fan, "Software Cost Estimation".
9. Krishnamoorthy Srinivasan and Douglas Fisher (1995), "Machine Learning approaches to Estimating Software Development Effort", In IEEE transactions on Software Engineering, Vol.21, No.2, February 1995.
10. Lionel C. Briand, "On the many ways Software Engineering can benefit from Knowledge Engineering".
11. Martin Auer, Adam Trendowicz, Bernhard Graser, Ernst Haunschmid, and Stefan Biffl (2006), "Optimal Project Feature weights in Analogy – Based Cost Estimation: Improvement and Limitations", In IEEE transactions on Software Engineering, Vol.32, No.2, February 2006.
12. M. Wasif Nisar, Yong-Ji WANG, Manzoor Elahi (2008), "Software Development Effort Estimation Using Fuzzy Logic – A Survey", In 5th International Conference on Fuzzy Systems and Knowledge Discovery, IEEE Computer Society.
13. Martin Auer, Stefan Biffl (2004), "Increasing the Accuracy and Reliability of Analogy – Based Cost Estimation with Extensive Project feature Dimention Weighting", In International Symposium on Empirical Software Engineering (ISESE'04), IEEE Computer Society.
14. Christopher Rush and Rajkumar Roy, "Expert Judgement in cost estimating: Modelling reasoning process".
15. G. Kadoda, M. Cartwright, L. Chen, and M. Shepperd. (2000), "Experiences Using Case-Based Reasoning to Predict Software Project Effort", In Proceedings of EASE, p.23-28, Keele, UK.
16. Marcio Rodrigo Braz, and Silvia Regina Vergilio (2006), "Software Effort Estimation Based on Use Cases", In Proceedings of the 30th Annual International Computer Software and Applications Conferences (COMPSAC'06), IEEE Computer Society.
17. Cuauhtemoc Lopez Martin, Jerome Leboeuf Pasqueir, Cornelio Yanez M., Agustin Gutierrez T., "Software Development Effort Estimation using Fuzzy Logic: A Case Study", In Proceedings of the 6th International conference on Computer Science (ENC'05, IEEE Computer Society).
18. Marcio Rodrigo Braz, and Silvia Regina Vergilio (2004), "Using Fuzzy Theory for Effort Estimation of Object – Oriented Software", In Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2004), IEEE Computer Society.
19. Cuauhtemoc Lopez Martin, Cornelio Yanez-Marquez, Agustin Gutierrez-Tornes(2006), "A Fuzzy Logic Model Based Upon Reused and New & Cahnged Code for Software Development Effort Estimation at Personal Level", In Proceedings of the 15th International conference on Computing (CIC'06), IEEE Computer Society.
20. A. Idri and A. Abran. (2001), "A Fuzzy Logic Based Measures for Software Project Similarity: Validation and Possible Improvements", In Proceedings of the 7th International Symposium on Software Metrics, pp. 85-96, England, UK, IEEE Computer Society.
21. Ali. Idri, Alain Abran, and Taghi M. Khoshgoftaar, "Computational Intelligence in Empirical Software Engineering".
22. V. Bhattacherjee and S. Kumar,(2004),"Software Cost Estimation and its Relevance in The Indian Software Industry", In Proceedings of the International Conference on Emerging Technologies IT Industry, PCTE, Ludhiana, India.
23. S. Kumar and V.Bhattacharjee, (2005), "Fuzzy Logic Based Model for Software Cost Estimation", In Proceedings of the international Conference on information Technology, Nov'05, PCTE, Ludhiana, India.
24. V. Bhattacherjee and S .Kumar, (2006), "An Expert-Case Based Frame work for Software Cost

Estimation", In Proceedings of the National Conference on Soft Computing Techniques for Engineering Application (SCT-2006), NIT Rourkela.
25. S. Kumar and V. Bhattacharjee, (2007), "Analogy and Expert Judgment: A Hybrid Approach to Software Cost Estimation", In Proceedings of the National Conference on information Technology: Present practice and Challenge, Sep'07, New-Delhi, India.
26. V. Bhattacherjee, S. Kumar and Ekbal Rashid (2008), "Estimation of Software Development Effort in University Setting: A Case Study", Presented at National Conference on Architecturing Future IT Systems (NCAFIS '08), Indore, October 2008.
27. V. Bhattacherjee, S. Kumar and E. Rashid, (2008), "A Case Study on Estimation of Software Development Effort.", In Proceedings of the & Technology, Hydrabad.(A National Conference ICACT-2008, Gokaraju Rangaraju Institute of Engineering.P.), India.
28. V. Bhattacherjee, S. Kumar and Ekbal Rashid (2009), "Case Based Estimation Model using Project Feature Weights", ISM, Dhanbad. Jharkhand.
29. V. Bhattacherjee, P. K. Mahanti and S. Kumar, (2009), "Complexity Metric For Analogy Based Effort Estimation", Journal of Theoretical and Applied Information Technology, Vol. 6, No. 1, 31st july 2009.
30. B.W. Boehm.(1981), "Software Engineering Economic", Prentice –Hall.
31. Internal Reports, Birla Institute of Technology Ranchi.
32. R. Fairley, (1997) "Software Engineering Concepts", T M H.
33. S. Rajasekaran and G. A. Vijayalakshmi Pai (2003), "Newral Networks Fuzzy logic, and Genetic Algorithms Synthesis and Applications", PHI.
34. Software Engineering Data Repository. http://promisedata.org/repository/data/desharnais/desharnais.arff.
35. J. M. Desharnais, "Analyse statistique de la productivitie des projets informatique a partie de la technique des point des foncti\ on," University of Montreal, Masters Thesis, 1989.
36. V. Bhattacherjee, and J. Pal, "Fuzzy Rule Based System for Fine Prediction ction in Accidents" in proceedings of Conference on information Science, Technology and Management, CISTM 2009, MDI Gurgaon, India, July 13-15, 2009.
37. J. Pal and V. Bhattacherjee, "A Fuzzy Logic System for Software Quality Estimation", in proceedings ICIT 2009, pp 183-187, 2009.
38. J. Pal and V. Bhattacherjee., "Movie Analyzer using Fuzzy Rule based System", in proceedings ICCS 2010, pp 16-21, November 19-20, 2010.
39. Bhattacherjee, V. and Pal, J. "Applying Fuzzy Clustering To Predict Software Usability", International Journal of Applied Research on Information Technology and Computing (IJARITAC), Volume: 1, Issue: 3, Sept-Dec, 2010, pp.279-289.
40. Bose B. K., "Expert systems, fuzzy logic, and neural network application in power electronics and motion control", Proceeding of the IEEE, vol. 82, Aug. 1994.
41. L. H. Tsoukalas and R. E. Using, "Fuzzy and Neural Approaches in Engineering", John Wiley, NY, 1997.
42. Math Works, Fuzzy Logic Toolbox User's Guide, Jan., 1998.

Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 13 Issue 10 Version 1.0 Year 2013
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

An On-Chip Delay Measurement Technique for Small-Delay Defect Detection using Signature Registers

By Rajeshwari Soma, Zulekha Tabassum & S. Prathap
JNTUH University, India

Abstract - This paper presents a delay measurement technique using signature analysis, and a scan design for the proposed delay measurement technique to detect small-delay defects. The proposed measurement technique measures the delay of the explicitly sensitized paths with the resolution of the on-chip variable clock Generator. The proposed scan design realizes complete on-chip delay measurement in short measurement time using the proposed delay measurement technique and extra latches for storing the test vectors. The evaluation with Rohm 0.18- m process shows that the measurement time is 67.8% reduced compared with that of the delay measurement with standard scan design on average. The area overhead is 23.4% larger than that of the delay measurement architecture using standard scan design, and the difference of the area overhead between enhanced scan design and the proposed method is 7.4% on average. The data volume is 2.2 times of that of test set for normal testing on average.

Keywords : *delay estimation, design for testability (DFT), integrated circuit measurements, semiconductor device reliability, signature register.*

GJCST-C Classification : *D.2.5*

Strictly as per the compliance and regulations of:

An On-Chip Delay Measurement Technique for Small-Delay Defect Detection using Signature Registers

Rajeshwari Soma [α], Zulekha Tabassum [σ] & S. Prathap [ρ]

Abstract - This paper presents a delay measurement technique using signature analysis, and a scan design for the proposed delay measurement technique to detect small-delay defects. The proposed measurement technique measures the delay of the explicitly sensitized paths with the resolution of the on-chip variable clock Generator. The proposed scan design realizes complete on-chip delay measurement in short measurement time using the proposed delay measurement technique and extra latches for storing the test vectors. The evaluation with Rohm 0.18- m process shows that the measurement time is 67.8% reduced compared with that of the delay measurement with standard scan design on average. The area overhead is 23.4% larger than that of the delay measurement architecture using standard scan design, and the difference of the area overhead between enhanced scan design and the proposed method is 7.4% on average. The data volume is 2.2 times of that of test set for normal testing on average.

Indexterms : delay estimation, design for testability (DFT), integrated circuit measurements, semiconductor device reliability, signature register.

I. Introduction

With the scaling of semiconductor process technology, performance of modern VLSI chips will improve significantly. However, as the scaling increases, small-delay defects which are caused by resistive-short, resistive-open, or resistive-via become serious problems. If small-delay defects cannot be detected in LSI screening, the chips will behave abnormally under particular operations in certain applications, and their lifetime may become very short due to the vulnerability to the transistor aging. Therefore to keep the reliability after shipping, Enhancing the quality of the testing for the smalldelay defect detection is an urgent need. The delay measurement of paths inside the circuits is useful for detection and debugging of small-delay defects.

However, it is impossible to measure the small circuit path delays using an external tester, even if the resolution is high. Therefore development of the embedded delay measurement technique is required. Some embedded delay measurement techniques have been proposed. The scan-based delay measurement technique with the variable clock generator is one of these on-chip delay measurements techniques. In this technique, the delay of a path is measured by continuous sensitization of the path under measurement with the test clock width reduced gradually by the resolution. The main good point of the scan-based delay measurement technique is its high accuracy. The reason of the high accuracy is that the technique measures just the period between the time when a transition is launched to the measured path and the time when the transition is captured by the flip flop connected to the path, directly. The variation of the measured value just depends on the variation of the clock frequency of the clock generator. Therefore, if the clock generator is compensated the influence of the process variation, the measured value does not depend on the process variation. However, it has a drawback. The measurement time of the technique depends on the time for the scan operation. These days, the gap between the functional clock and scan clock frequency increases. Therefore the measurement time becomes too long to make it practical. Noguchi *et al.* proposed the self testing scan-FF. The flip flop reduces the required number of scan operations, which makes the measurement time practical. They also proposed the area reduction technique of the self testing scan-FF. However, the area overhead of these methods is still expensive compared with the conventional scan designs. This paper presents a scan-based delay measurement technique using signature registers for small-delay defect detection.

The proposed method does not require the expected test vector because the test responses are analyzed by the signature registers. The overall area cost is of the order of conventional scan designs for design for test (DFT). The measurement time of the proposed technique is smaller than conventional scan-based delay measurement. The extra signature registers can be reused for testing, diagnosis, and silicon debugging. The rest of this paper is organized as follows. In section II, preliminaries are discussed. Section III explains the proposed method. Section IV

Author α : M. TECH (VLSI), JITS, Narsampet, India.
E-mail : rajeshwari.473@gmail.com
Author σ : ASST.PROF. (ECE), JITS, Narsampet, India.
E-mail : tabbu_2620@yahoo.com
Author ρ : M.TECH (VLSI), GEC, Hanamkonda, India.
E-mail : prathap.soma@gmail.com

shows the experimental results. Finally, Section V concludes this paper.

II. Preliminaries

a) Related Works

These days, various methods for small-delay defect detection have been proposed. The path delay fault testing with a normal clock width is the most popular and is widely used. In this method, we choose the longer paths to detect the smaller cumulative delay due to the small delay distributed on the paths. The larger the cumulative delays, the higher the probability of the detection of the distributed small delay. However, the coverage of the small delay defect detection largely depends on the normal clock width, which is a problem of this method. On the other hand, to solve the problem, methods with delay fault testing using a variable clock generator have been proposed. The delay fault testing with a smaller test clock reduces the slack of the paths. Therefore the smaller delay defects which cannot be captured with the normal clock width can be captured with the appropriate smaller test clock width. Lieu*et al.* proposed a small-delay defect detection method consisting of two test phases using both a tighter test clock and a normal clock. In this method, the transition delay fault testing with a tighter test clock width which is calculated basedon the characterized delay distribution is applied in the first phase. After that, path or transition delay fault testing is applied with a normal clock width in the second phase. The quality of the method is higher than the conventional one phase test with normal clock width. Yan. proposed a delay testing scheme that identifies small-delay defects in the slack interval by comparing switching delays of a neighboring die on a wafer. In this method, a fault site is sensitized multiple times continuously with reducing the test clock width by the slack interval. The abnormal switching delays are detected by comparison with the test results in the neighboring die. Another work detects small-delay defects by analyzing the failing frequency, which is the minimum frequency that the delay fault testing fails when the path is sensitized multiple times continuously with increasing the test clock frequency. Generally, the variable clock testing requires a variable on-chip clock generator. Various variable on-chip clock generators have been proposed. In variable clock testing, the test clock frequency should be optimized to each test vector. To improve the test quality, various optimization methods of the test clock and test set have been proposed. These days, small-delay defect detection methods using on-chip delay measurement techniques have been proposed. The direct measurement of the real delay of each path of each chip screens outlier chips robustly even in the presence of process variation or the gap between real and simulation environment. It realizes higher fault coverage of small-delay defects than the simulation based ones. In addition, it can be used not only for the detection of small delay defects, but also for the debugging. Because modern chips are too huge and complex, LSI CAD tools cannot optimize the design enough. Hence, the manufactured first silicon chip usually does not meet the specification in spite of the tighter release to production (RTP) schedule.

Therefore silicon debugging and design for debugging (DFD) become much more important in modern chips. Various silicon debugging technologies and DFD methods have been proposed. On-chip delay measurement provides accurate information of the delay of inside paths for the debugging of small-delay defects. Most of the conventional works of on-chip delay measurement are classified to either a proposal of an embedded delay Fig.1.On-chip variable clock generator measurement circuit or that of a scan architecture for scan-basedon-chip delay measurement with a variable clock generator. Some works proposed embedded delay measurement circuits of modified vernier delay line (VDL). Datta *et al.* proposed the embedded delay measurement circuit with high resolution. It is the first work of an embedded measurement circuit of modified VDL to the best of our knowledge. Tsai *et al.* proposed the area efficient and noise-insensitive modified VDL with coarse and fine parts namely BIDM. Pei *et al.* also proposed the area efficient modified VDL. The feature of this method is delay range of each stage of VDL .The delay ranges increase bya factor of two gradually, which reduces the required stages. Therefore the area is smaller than Data's work. The modified VDLs achieve high resolution. However they require redundant lines to feed the input and output signals of the measured paths, which needs the compensation of the delay effect of the redundant lines. Tanabe *et al.* solved the problem by removing the delay of the redundant lines from the measured delay using some additional embedded circuits. The proposed technique is categorized to the scan-based one.

b) Variable Clock Generator

In the proposed method, the clock width should be reduced continuously by a constant interval as explained later. It is difficult for an external tester to control this clock operation. Therefore an on-chip variable clock generator is indispensable for the proposed method. In this paper, we use the on-chip variable clock generator proposed by Noguchi. Fig. 1 illustrates the circuit.

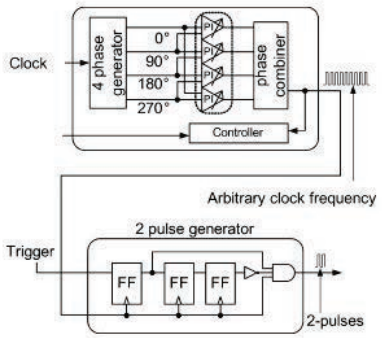


Figure 1 : On-chip Variable Clock Generator [2]

The circuit consists of the phase interpolator-based clock generator and the 2-pulse generator. The phase interpolator-based clock generator generates an arbitrary clock width. The 2-pulse generator generates 2-pulsetest clocks with arbitrary timing in response to a trigger signal.

Some of the specification and the evaluation results are shown in Table I.

Table 1 : Specification and Measurement Result

Process	90nm CMOS
Occupied area	300 [μm] × 128 [μm]
Input clock	1.5GHz 4phase
Output clock	1GHz to 2GHz 4phase
Functions	· Frequency control · Jitter generation · Duty ratio control · Phase control
Timing (Phase) Step resolution	5.2ps (2.8°)
Step error (5 chip measured)	±1.3ps ~ ±2.4ps (±0.7° ~ ±1.3°)
Cumulative timing error (5 chip measured)	Best chip Worst chip ±5.4ps ~ ±11.2ps (±2.9° ~ ±6.3°)

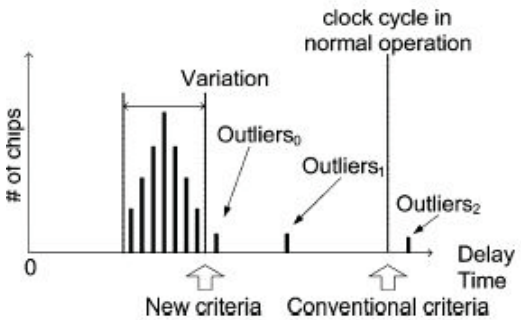


Figure 2 : Decision of test clock width based on path delay distribution obtained by chip measurement

c) Small-Delay Defect Detection with Delay Measurement of Chips

The proposed method uses the Noguchi's small-delay defect detection technique. In this technique, the test clock width for delay fault testing of each path is determined with the normal path delay distribution of each path. This strategy has already been applied to various small-delay detection techniques. But its originality is to obtain the path delay distribution with the delay measurement of the paths of the fabricated chips. Fig. 2 shows the path delay distribution of a path obtained by the delay measurement of the fabricated sample chips. The horizontal axis is measured delay. The vertical axis is the number of chips. The chips which have delay inside the range Variation are normal chips. The chips which have delay outside the range Variation, namely, are abnormal chips. The delay is the outside of the clock cycle in normal operation. Therefore it will be detected by conventional delay fault testing with the clock cycle in normal clock operation which is Conventional criteria. The delay and are within the clock cycle in normal operation. The conventional delay fault testing regards them as good chips. However because the delay is outside Variation, it will cause improper operations under particular operation in certain applications and may cause improper operations after shipping due to the effect of aging. In Noguchi's technique, the test clock cycle is set to the upper limit of the distribution of normal chips, which are new criteria. Then all the outlier chips are detected by the delay fault testing. In small technology, the path-delay distribution calculated by simulation is different from that of the fabricated chips. Therefore the quality of its strategy is higher than that of simulation based ones. Because the Noguchi's technique requires the measurement of the explicit paths, the paths should be single-path sensitizable. The aim of the technique is to screen the chips.

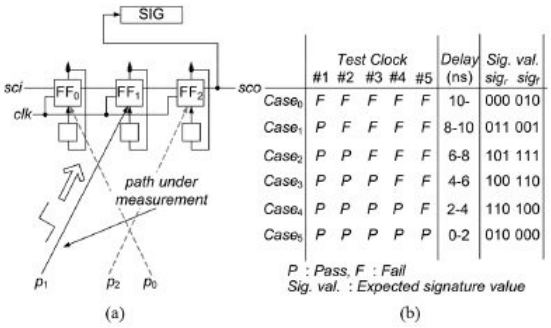


Figure 3 : Concept of proposed delay measurement. (a) Basics of proposed measurement, (b) Signature Table

Which have abnormal delay in gates or wires Therefore the test Set for the measurement should detect all the transition faults which are sensitized through single-path sensitizable paths. The proposed method is a new delay measurement technique for the small-delay defect detection technique.

III. Delay Measurement Technique using Signature Registers

This section explains the proposed measurement method. Section III-A presents the

concepts of the proposed method. Section III-B explains the implementation of the proposed method. Section III-C describes the measurement sequence. The data volume and area overhead should be realistic compared with the conventional scan designs for DFT. Section III-D explains the reduction method of the tester channel. Section III-E describes the scheme for the decision of the number of the required extra latches to keep the cost realistic. To apply the proposed method and realize short measurement time, some constraints should be put on ATPG. Section III-F explains the ATPG constraints. Section III-G describes the measurement time and data volume. Finally, Section III-H describes the test response tracing mode for finding lowest failing frequency or diagnosis with transition fault test vectors.

a) Basics

This section explains the concept of the proposed delay measurement. The target paths of the proposed method are single path sensitizable. Basically, the proposed method is scan-based delay measurement. The difference from the basic one is the usage of the signature registers and the additional latches for the acceleration of the delay measurement. Fig. 3(a) shows the basics of the proposed method. This example has three flip flops, and each flip flop has the input line (bottom), the output line (upper), and the clock line. Each flip flop is connected to an extra latch. At first, we assume that each flip flop has its own extra latch. The value of each flip flop is stored in the correspondent latch, and the value of each latch can be loaded to the correspondent flip flops in arbitrary timing. In the proposed measurement, the test vector is stored in these latches after scan-in operation. Once the test vector is stored in the latches, the test vector can be loaded from these latches in a clock without scan-in operation. It reduces the time for multiple sensitization of a path drastically. The horizontal line through these flip flops represents the scan path. The symbols and represent the scan input and output, respectively. The rectangle SIG represents the signature register using the linear feedback shift register as its basic component. The input of SIG is connected to the output of the last flip flop. More detail structures of the flip fops and the signature register are shown in Figs. 4 and 5, respectively.

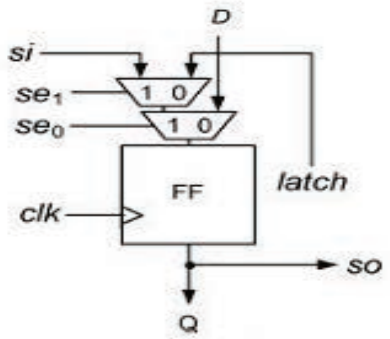


Figure 4 : Scan Flip Flops for Proposed Measurement

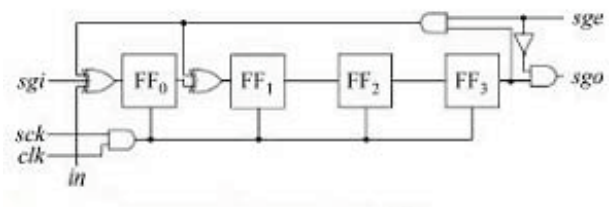


Figure 5 : Four Bit Reconfigurable Signature Register

Here, we measure the delay of. In this example, we assume that the clock width of normal operation is 10 ns, and the resolution of the delay measurement is 2 ns. First, SIG is initialized with reset operation. Second, the target path is tested continuously 5 times with the test clock reduced gradually by the resolution. The multiple clock width testing is realized by the variable clock generator explained in Section II-B. The test clock of the 1st testing (#1) is 10 ns. After the test, the test response is sent to SIG through the scan path with two clock shift out operation. The test clock of the second testing (#2) is 8 ns. Similarly, the test clock width of the third, fourth, and fifth testing's (#3, #4, #5) are the difference between 2 ns and the previous test clock width. Each test response is sent to SIG with two clocks. After the above 5 times of delay fault testing's, the signature value of SIG is retrieved. To estimate the delay, the retrieved signature value is compared with the expected signature values of the signature table. Fig. 3(b) shows the signature table in this example. The table has four columns. The first column is the cases of the measurement. The second column is the sequences of the test responses of #1–#5. The third column is the path delay value. The fourth column is the signature values of each case.

Here, and are the signature values for rising and falling transition testing's, respectively. The delay of each path is decided as more than 10, 8–10, 6–8, 4–6, 2–4, or 0–2ns, with 2ns resolution. The sequences of the test responses of the 5 times measurement are shown in Fig. 3(b). The symbols indicate the cases with path delays, more than 10, 8–10, 6–8, 4–6, 2–4, 0–2 ns, respectively. The symbols and represent the pass and fail of a testing, respectively. In case of rising transition testing, and, and in case of falling transition testing, and. The retrieved signature value is compared with the expected signature values on the table, and decides the delay value. When the number of flip flops is, clock width is, the measurement resolution is, and the continuous testing time is, the delay measurement sequence of a target path is as follows. Here, we assume that the test vector is already stored in the latches. The end point of the measured path is.

Step 1: Initializing SIG.
Step 2: Test vector is loaded from the latches.
Step 3: Test clock width is set to normal clock width.
Step 4: Test clock is applied.
Step 5: The test response is sent to SIG which is connected to the output of with clocks.
Step 6: If testing time is equal to, go to Step 7 after the signature value of SIG is retrieved, otherwise go back to Step 2 after the test clock width is updated to
Step 7: The delay value is estimated by comparing the retrieved signature value and the signature table.

b) Implementation

In this subsection, we explain the implementation of the proposed measurement system. First we explain the important components to understand the whole system. After that, the whole system is presented.

i. *Scan Flip Flop for Measurement*

Fig. 4 is the gate level description of the scan flip flop for the proposed measurement. The lines, and are the input, output, and clock lines, respectively. The line is connected to an extra latch which provides the test bit to the flip flop. The lines and are the input and output for constructing the scan path. The input is connected to of an adjacent scan flip flop or the scan input. The output is connected to of an adjacent scan flip flop or the scan output. The flip flop has two multiplexers. The lines and are the inputs of the upper multiplexer controlled by the output of the upper multiplexer and are the inputs of the bottom multiplexer controlled by When, the flip flop is in normal operation mode. When and, the flip flop is in scan operation mode. When the flip flop loads the value stored in the latch connected to the latch line.

ii. *Reconfigurable Signature Register*

The signature register for the proposed measurement requires the following functions to meet the demand of the proposed measurement.

- Capturing the test response in arbitrary timing.
- Shifting out the signature data in arbitrary timing.

Fig. 5 shows the architecture of the signature register for the proposed measurement. The length of the signature register in this example is four bit. Therefore it has four flip flops. The signature register can be configured to a shift register. The line controls the configuration. When, it works as a signature register. When, it works as a shift register.

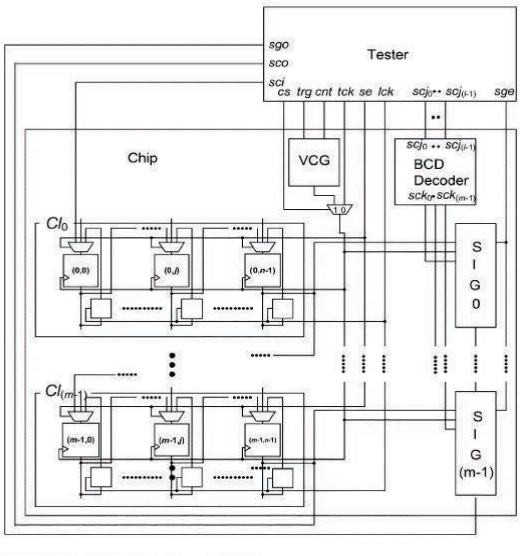


Figure 6 : Proposed Measurement System

The line is the input of the signature register. During measurement, test responses are sent to. The line is clock line. The clock line is controlled by. When, the signature register does not capture the input value. When, the signature register captures the input value synchronizing with the positive edge of. By controlling, the signature registers capture only the target test response. When, this circuit is configuring to the shift register. The input is. The output is. As explained later, the measurement system requires multiple signature registers generally. The input and output are connected to the output and the input of adjacent signature registers to construct a long shift register for sending all the signature values to the external tester.

iii. *Whole System*

Fig. 6 shows the proposed measurement system. The proposed system consists of the low cost tester and the chip with the variable clock generator (VCG) explained in Section II-B and a BCD decoder. The chip is assumed to have single functional clock in the proposed method, and the chip has two reset lines for initializing the flip flops and the signature registers independently. The reset operations are controlled by the tester. The low cost tester controls the whole measurement sequence. The clock frequency is slower than the functional clock. The line retrieves the signature data from the signature registers to estimate the measured delay. The line sends the test vectors to the scan input of the chip. The line gets the data of the flip flops from the scan output of the chip. In the proposed measurement sequence, is not used. However, it is used to check the flip flops or the additional latches before the measurement. The line is the clock control line. The proposed measurement uses both the slow tester clock and the fast double pulse generated by on-chip VCG. The lines elect the slow and fast clock. If is 1,

the fast clock is sent to the clock line of the components. Otherwise the slow tester clock is sent. The lines and are the input lines for VCG. The fast double pulse is launched synchronizing with the positive edge of. The line controls the width of the double pulse. The line controls the scan flip flops. The line controls the latches for storing test vectors. The lines are the inputs for the encoded data to control the capture operation of the signature registers. The BCD decoder decodes the encoded input data to the control data of the signature registers. As explained later, the decoder is used to reduce the input lines for the control data of the signature registers. The is the enable signal for the signature registers. The flip flops in the chip are classified to the clusters. Here, we assume that each cluster has flip flops, and thus the number of the flip flops is. In general, the number of the flip flops of the last cluster is, where is the number of flip flops, or. The coordinate written in the flip flops indicates the location. The number is the cluster id. The number is the order in the cluster. The output of the flip flop of which is the tail flip flop of each cluster is connected to the flip flop of which is the head flip flop of. These lines construct the scan chain. The output of the tail flip flop of each cluster is connected to the input of the corresponding signature register. The path whose test response is captured by the flip flops included in is measured by. The control lines of the signature registers are connected to the BCD decoder.

c) Measurement Sequence

Here, we explain the measurement sequence. First, the measurement sequence of the paths simultaneously sensitized in a test vector is explained in Section III-C1. After that, the whole measurement sequence is explained in Section III-C2.

i. *Measurement Sequence per a Test Vector*

When the measurement system has signature registers, Paths can be measured in parallel maximally. To reduce the measurement time, we measure multiple paths simultaneously. We explain the measurement strategy using the example depicted in Fig. 7. In this example, the proposed method is applied to the circuit with six flip flops. These flip flops are classified to the two clusters and. The cluster includes, and includes. The cluster has the signature register. The cluster has the signature register. The clock line controls these flip flops. The control lines controls the capture operation of, and controls the capture operation of. The paths are sensitized simultaneously by the test vector. The test response of is captured by The expected test response is The paths and are measured by. The paths and are measured by. The combination of the two paths, one of which is selected from and, the other of which is selected from and, can be measured simultaneously.

IV. Experimental Result

In this section, we present the experimental results. The proposed method is compared with the conventional methods. The clock frequencies are the same as that of Noguchi's methods that is, the normal clock frequency is 100 MHz, and the scan clock frequency is 10 MHz the length of the signature register is 8 bit. The test set consists of test vectors which detects all single-path sensitizable transition faults. The paths sensitized by these test vectors are measured. In this evaluation, the average delay of signal propagation of the measured paths is assumed to be the half of the clock width. Usually the length of the sensitized paths for transition delay defect detection is relatively short because ATPG seeks the paths whose length is as short as possible for the cost of test generation. From this point of view, we believe that this assumption is valid. In this evaluation, the test set for the measurement is generated by podem-based ATPG algorithm, which is implemented by C++. The back track limit is 200. For the evaluation, relatively larger ISCAS'89 benchmark circuits are used. The measurement time using the proposed scan design and standard scan design is calculated by the (1) and (3), respectively. The results of area are obtained by synthesis with Synopsys design compiler using Rohm 0.18 m process. The data volume of these methods is calculated by the (4) and (5), respectively. In this evaluation, area overhead, area overhead including the area of VCG, and the routing overhead are defined as follows. , where the area of the circuits is implemented the evaluated method except VCG, and is the area of the non-scan circuit except clock generator. Where the area of the circuits is implemented the evaluated method including VCG. The area of VCG is obtained from the paper. Note that, and include the area of the wires. , where and are the routing area of the evaluated circuit and non-scan circuit, respectively. The reduction ratio of measurement time is defined as, where and is the measurement time of the evaluated method and that of standard scan, respectively. The increase ratio of the data volume is defined as, where and is the data volume of the evaluated method and that of standard scan, respectively.

a) Comparison with Conventional Scan-Based Measurement

Tsig=time required for {whole scan-in + double pulse + SIG data

Table 1 : Measurement Time

Circuit	EMB	T_{sig}	R_1
S5378	635.4	181.2	71.5
S9234	703.1	135.5	80.7
S38584	4025	1011.5	72.6
Ave			74

Table 2 : Comparison with Delay Measurement using Embedded Delay Measurement Circuit

Circuit	EMB	SIG	R
S5378	46.7	181.2	3.9
S9234	164.5	135.5	0.8
Ave			2.3

The result of the measurement time is shown in Table III. Here, we show the result when and, respectively. The sub columns (ms) and (ms) are the measurement time of standard scan design and the proposed method, respectively. According to the result, when and are 67.8% and 71.9%, respectively. The reduction ratio when is 4.1% higher than when. It is because the impact of the time for scan-in operation of each test vector on the measurement time decreases as increases.

V. Conclusion

The proposals of this paper are as follows:

- The proposal of the delay measurement method using signature analysis and variable clock generator.
- The proposal of a scan design for the delay measurement of internal paths of SoC.

The first proposal can be applied not only SoC but also field programmable gate array (FPGA). Because the process of FPGA is getting extraordinary smaller these days, the small delay defect becomes serious problem in FPGA, too. In this meaning, the application of the proposed method to FPGA is also useful. A future work is the low cost application of the proposed measurement to FPGA. When we measure short paths the measurement error can increase for the IR drop induced by higher test clock frequency. It can reduce the test quality. Another future work is the reduction and the avoidance of the measurement error caused by the IR drop.

References Références Referencias

1. N. Ahmed and M. Tehranipoor, *Nanometer Technology Designs: High Quality Delay Tests*. New York: Springer, 2007.
2. K. Noguchi, K. Nose, T. Ono, and M. Mizuno, "A small-delay defect detection technique for dependable LSIs," in *Proc. IEEE Symp. VLSI Circuits*, 2008, pp.64–65.
3. B. I. Dervisoglu and G. E. Stong, "Design for testability: Using scan path techniques for path-delay test and measurement," in *Proc. IEEE Int. Test Conf. (ITC)*, 1991, pp. 365–374.
4. K. Noguchi, K. Nose, T. Ono, and M. Mizuno, "An area reduction technique of self-testing FFs for small delay defects detection," The Inst. Electron., Inf. Commun. Eng. (IEICE), Tech. Rep. DC2009-16.
5. A. Krstic and K.-T. Chen, *Delay Fault Testing for VLSI Circuits*. Norwell, MA: Kluwer, 1998.
6. J.-J. Liou, L.-C. Wang, and K.-T. Cheng, "Enhancing test efficiency for delay fault testing using multiple clocked schemes," in *Proc. Des. Autom. Conf. (DAC)*, 2002, pp. 371–374.
7. H. Yan and A. D. Singh, "Evaluating the effectiveness of detecting delay defects in the slack interval: A simulation study," in *Proc. IEEE Int. Test Conf. (ITC)*, 2004, pp. 242–251.
8. J. Lee and E. J. McCluskey, "Failing frequency signature analysis," in *Proc. Int. Test Conf. (ITC)*, 2008, pp. 1–8.
9. R. Tayade and J. A. Abraham, "On-chip programmable capture for accurate path delay test and characterization," in *Proc. IEEE Int. Test Conf. (ITC)*, 2008, pp. 148–159.
10. M. Beck, O. Barondeau, X. Lin, and R. Press, "Logic design for on-chip test clock generation-implementation details and impact on delay test quality," in *Proc. IEEE Des., Autom. Test Eur. Conf. Exhib. (DATE)*, 2005, pp. 56–61.
11. S. Kaeriyama, M. Kajita, and M. Mizuno, "A 1-to-2GHz 4-phaseon-chip clock generator with timing margin test capability," in *Dig. Tech. Paper Int. Solid-State Circuits Conf. (ISSCC)*, 2007, pp. 174–594.
12. T.-Y. Li, S.-Y. Huang, H.-J. Hsu, C.-W. Tzeng, C.-T. Huang, J.-J. Liou, H.-P. Ma, P.-C. Huang, J.-C. Bor, and C.-W. Wu, "Af test: Adaptive-frequency scan test methodology for small-delay defects," in *Proc. IEEE Int. Symp. Defect Fault Toler. VLSI Syst.(DFT)*, 2010,
13. N. Ahmed and M. Tehranipoor, "A novel faster-than-at speed transition-delay test method considering IR-drop effects," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 28, no. 10, pp.1573–1582, Oct. 2009.
14. M. Yilmaz, K. Chakrabarty, and M. Tehranipoor, "Test-pattern grading and pattern selection for small

delay defects," in *Proc. IEEE VLSI Test Symp. (VTS)*, 2008, pp. 233–239.
15. B. Kruseman, A. K. Majhi, G. Gronthoud, and S. Eichenberger, "On hazard-free patterns for fine-delay fault testing," in *Proc. Int. Test Conf.(ITC)*, 2004, pp. 213–222.
16. M. Sharma and J. H. Patel, "Bounding circuit delay by testing a very small subset of paths," in *Proc. IEEE VLSI Test Symp. (VTS)*, 2000, pp. 333–341.

Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 13 Issue 10 Version 1.0 Year 2013
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

On the $M/M/c/N$ Call Center Queue Modeling and Analysis

By Nwonye Chukwunoso
University of Port Harcort, Nigeria

Abstract – The $M/M/c/c$ model is the most widely applied queueing model in the mathematical analysis of call centers. The $M/M/c/c$ model is also referred to as the Erlang Loss System. The Erlang loss model does not take into consideration system attributes such as blocking and busy signals, balking and reneging, retrials and returns. Although, the Erlang loss model is analytically tractable, it is not easy to obtain insight from its results.

The need to develop a more accurate call center model has necessitated the modification of the Erlang loss model. In this research, we model and analyze a call center using $M/M/c/N$ the model. The goal of this paper is to extend existing results and prove new results with regards to the monotonicity and limiting behaviour of the $M/M/c/N$ model with respect to the system capacity N.

GJCST-C Classification : D.2.0

Strictly as per the compliance and regulations of:

On the $M/M/c/N$ Call Center Queue Modeling and Analysis

Nwonye Chukwunoso

Abstract - The $M/M/c/c$ model is the most widely applied queueing model in the mathematical analysis of call centers. The $M/M/c/c$ model is also referred to as the Erlang Loss System. The Erlang loss model does not take into consideration system attributes such as blocking and busy signals, balking and reneging, retrials and returns. Although, the Erlang loss model is analytically tractable, it is not easy to obtain insight from its results.

The need to develop a more accurate call center model has necessitated the modification of the Erlang loss model. In this research, we model and analyze a call center using the $M/M/c/N$ model. The goal of this paper is to extend existing results and prove new results with regards to the monotonicity and limiting behaviour of the $M/M/c/N$ model with respect to the system capacity N.

I. Introduction

The call center industry has grown explosively in the recent past and that has aroused the interest of researchers from different disciplines. Mandelbaum [11] have provided a comprehensive research bibliography with abstracts in diverse disciplines such as Operations research, Statistics, Engineering, and so on. Call Center research has been reviewed in the tutorial and survey paper by Gans et al. [6]. In this paper, our focus is on the computational rigor of the call center performance metrics using the $M/M/c/N$ model.

a) *Description of a Call Center*

A call center is a department of an establishment that attends to customers via telephone conversation often for the purpose of sales and product support, or that makes outgoing telephone calls to customers usually for the purpose of advertisement or telemarketing. Suppose the department also attends to e-mails, faxes, letters, and other similar written correspondence, then, it is called a contact center.

Inbound call center only handle incoming telephone calls initiated by customers while out bound call centers only make outgoing telephone calls to customers. There are call centers that deal with both types of calls. In majority of the call centers, inbound calls form the bulk of contacts with customers. In addition, inbound calls are more time consuming compared to other types of contacting options (e.g. e-mails, faxes, or letters) in terms of waiting times in the telequeue or sojourn times. Hence, we will only focus on inbound call centers. In an inbound call center, there is a group of agents (Customer Sales Representatives, CSRs) who provide the needed service through talking to customers on phones. In this paper, we shall use the terms "agents" and "CSRs" interchangeably. Agents are equipped with equipment, such as a Private Automatic Branch Exchange (PABX or PBX), an Interactive Voice Response Unit (IVRU or VRU), an Automatic Call Distributor (ACD), and computers [16]. See Figure 1.1 for details on the operational process and components of an inbound call center.

Author : *Department of Mathematics and Statistics, University of Port Harcort, Nigeria. E-mail : nwonye@gmail.com*

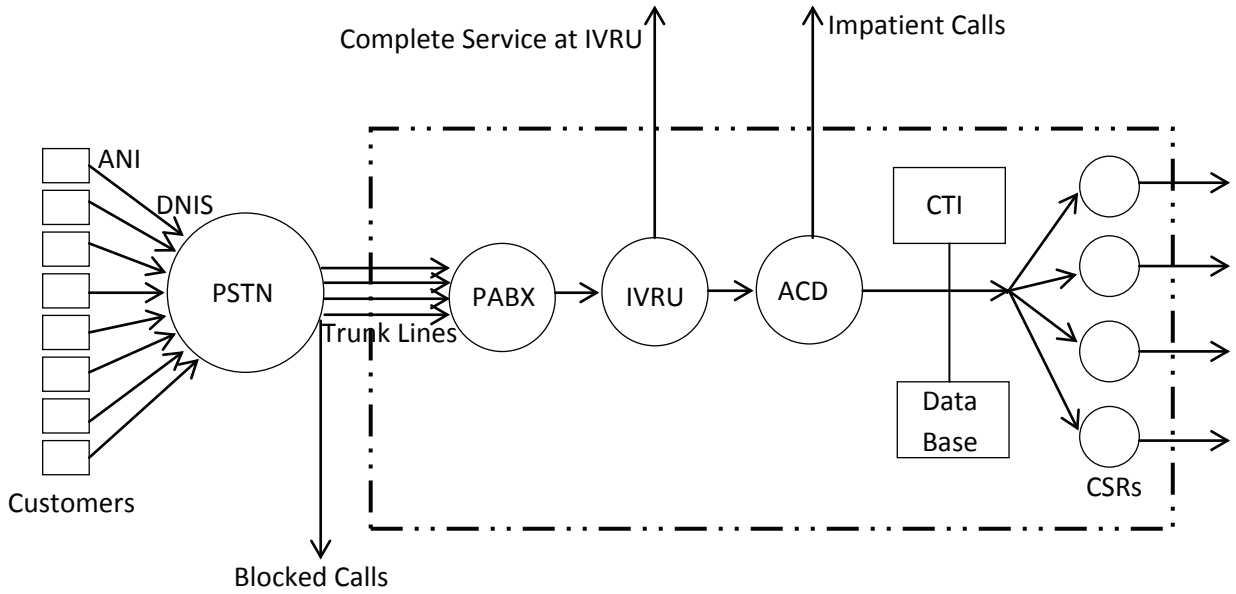


Figure 1.1 : Operational process of an inbound call center

b) The Operational Process of an Inbound Call Center

At some point in our lives, we have all called a call center. We will describe the operational process and components of an inbound call center in line with the description in [6, 17]. The process is depicted in Figure 1.1. Customers wanting to receive service from a call center, dial a special number provided by the call center. The Public Service Telephone Network (PSTN) company then uses the Automatic Number Identification (ANI) number (the phone number from which the customer dials) and the customer's Dialed Number Identification Service (DNIS) number (the special number being dialed) to connect the customer to the PABX privately-owned by the call center. The telephone lines (usually called trunk lines) connect the PABX to PSTN. If a trunk line is available, the customer seizes it; else the customer receives a busy signal and will be rejected. Hence, this customer is said to be blocked. Once the call is accepted, the customer will be connected through the PABX to the IVRU. The IVRU provides some automatic service for customers as well as several options for customers to choose from. Upon service completion at the IVRU, some customers leave the system and release the trunk lines. If the customer requires the service of an agent, the call will be passed from the IVRU to the ACD. The ACD is a sophisticated instrument designed to route calls to agents based on the specific needs of calls. If no appropriate CSRs are available, the customer is informed to wait and join a queue at the ACD. The customer is said to be delayed. The ACD decides the next customer to get service according to some preprogrammed queueing discipline (usually First Come First Served, FCFS). Delayed customers may decide to hang up and abandon (or renege) before they are served if they perceive that the service is not worth the wait. Such customers are said to be impatient. Patient customers (who do not abandon service) will eventually be connected to an agent. In serving a customer, the CSR works with a PC furnished with Computer-Telephony Integration (CTI), which is the technology that allows interactions on a telephone and a computer to be integrated. CTI will help ACD to route the call, help the CSR to get the caller's information from the database and hence facilitate the service process. At the completion of service and exit of the customer, the CSR still needs some wrap-up time to finish the whole service process and then may be available for the next customer. The service time is the sum of talk time and wrap-up time. Customers who abandoned and were blocked may try to call again after some random amount of time and these calls are referred to as retrials. Customers who finished talking with anagent may also need further assistance and therefore call back. Hence they become return customers or feedback customers. Notice that these two types of customers are not shown in Figure 1.1.

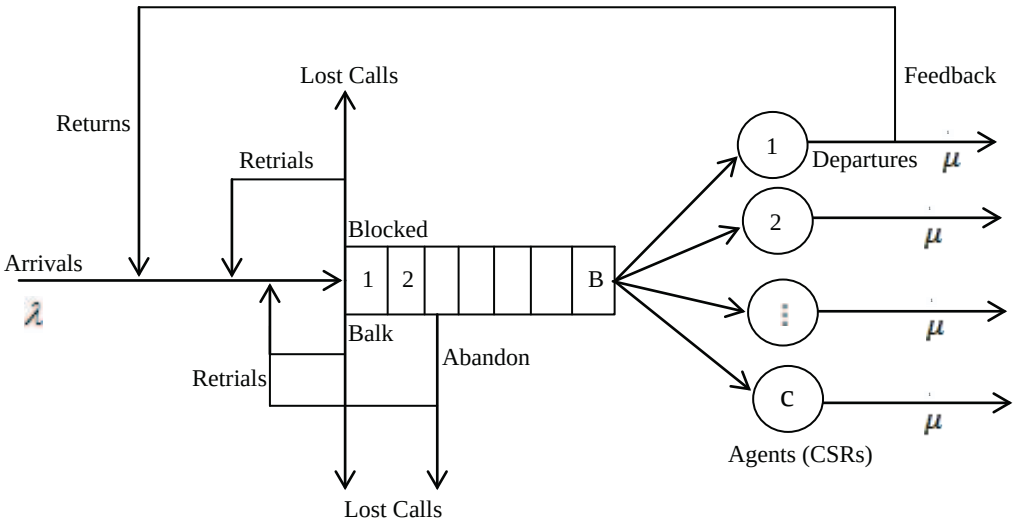


Figure 1.2 : Call Center as a Queueing System

c) The Call Center as a Queueing System

Figure 1.2 depicts the call center as a queueing system [7]. The number of agents (CSRs) and waiting spaces are denoted respectively by c and B. Hence there are $N = B + c$ trunk lines at the call center with B waiting spaces. If an arriving call finds all N trunk lines occupied, it gets a busy signal and as such is blocked and cannot access the system. If there is an available truck line, the call is either connected to the system and seizes one of the free trunk lines or it balks. Suppose there is an available trunk line and at least a free agent, then the call is immediately serviced. Otherwise the call experiences delay and has to wait in a queue at the ACD for a CSR to become available. Calls at the ACD may become impatient and abandon (renege) the system before being served and thus release the trunk line. The ACD usually implements the FCFS queueing discipline. Upon service completion by a CSR, the call leaves the system and then releases both the trunk line and the CSR and these resources become available to other arriving calls. Return (or feedback) calls are calls that return after been served by an agent. Some of those calls who do not get served (blocked, abandon or balk) may call again and they become retrials. The remaining calls become lost calls.

Suppose that the call arrivals follow a Poisson process with mean rate λ and that the service times of the calls are independent and identically distributed $(i.i.d)$ exponential random variables with mean $1/\mu$. Then we can model the system as a $M/M/c/N$ queueing system with features such as balking, abandonment, retrial, and feedback.

d) Performance Evaluation of the Call Center Queueing Model

In this paper, we will ignore features such as balking, abandonment, retrial, and feedback. Following the above assumptions, we will apply the $M/M/c/N$ model in analyzing the call center performance. The $M/M/c/N$ queueing system has a closed-form solution for the system state (number of calls in the system), the queue length (number of calls in the queue) distribution and waiting time distribution. Then we can obtain system performance metrics such as average waiting time, average queue length, and probability of blocking. We will apply the performance analysis of the $M/M/c/N$ queueing system to call center modeling and in turn show new results. The call center performance measures (metrics or indicators) provide useful information in the design and management of call centers. Performance measures are used in determining the service levels (or quality of service) in call centers.

Not all queueing models can be analyzed exactly to obtain performance measures as $M/M/c/N$ model. For instance, if we include additional features such as Non-Poisson time varying arrival process, balking, abandonment, retrial, feedback, and non-exponential service times, the model may become insolvable using traditional queueing techniques and other techniques have to be used to analyze the model such as simulation modeling.

II. Modeling Call Centers as Single-Node Exponential Queueing Models

In this section, we provide a detailed review of relevant single-node multiserver Markovian queueing models of call centers. Table 2.1 provides a list some main Markovian queueing models and their

performance indicators. Our emphasis is on the computational rigor of the exact performance measures of these well-known models as well deriving new results.

Table 2.1 : Some Multiserver Markovian Queueing Models

Notation	Models	Performance Indicators
$M/M/c$	Delay Model (Erlang C)	$P(delay) = P(W_q > 0)$ (Erlang C formula) *TSF ; ASA*
$M/M/c/c$	Blocking/Loss Model (Erlang B)	$P(blocking) = p_c$ (Erlang B formula)
$M/M/c/N$	Blocking and Delay Model	$P(delay) = P(W_q > 0); P(blocking) = p_N$ *TSF ; ASA*

TSF : Telephone Service Factor ; ASA: Average Speed to Answer; AWT: Acceptable Waiting Time

a) Model Assumptions

The $M/M/c/N$ is a generalization of the $M/M/c$ and $M/M/c/c$ models. In order to analyze a call center using the $M/M/c/N$ Markovian queueing model, we assume that the inter arrival and service times are exponentially distributed random variables.

Calls arriving at the call center are of a single type following a homogeneous Poisson process with rate λ. Callers are assumed to be patient and there is no form of impatience (balking or reneging). All agents (CSRs) are assumed to be statistically identical (i.e., equally skilled and provide service at the same rate). The service times are assumed to follow the exponential distribution with mean $1/\mu$. Services are rendered according to the First-Come-First-Serve queueing discipline. There are $N - c$ waiting spaces in the $M/M/c/N$ queueing system.

Let $L(t)$ be the number of calls in the system and $L_q(t)$ be the number of calls waiting in queue at time t. Let $W(t)$and $W_q(t)$ be the steady-steady sojourn time and waiting time in queue respectively. Since the models are Markovian, $L(t), L_q(t), W(t), and\ W_q(t)$ can be obtained using the birth-death processes. Our focus is on steady-state distribution of $L(t), L_q(t), W(t),$ and $W_q(t)$ with corresponding variables $L, L_q, W, and\ W_q$, respectively.

Let $p_n = P(L = n),\ n = 0,1,2, \ldots$ denote the steady-state probability (if it exits) of the system being in state n (i.e. having n calls in the system). Applying the modeling techniques of the birth-death processes, we can obtain some interesting system performance measures such as $E(L), E(L_q), E(W), and\ E(W_q)$.

Due to the PASTA property, we have for the $M/M/c$ model, $P(waiting) = \sum_{n=c}^{\infty} p_n$ and in the cases of the $M/M/c/N$ and $M/M/c/c$, we have $P(blocking) = p_N$ and $P(blocking) = p_c$ respectively.

b) Review of the M/M/c/c Model and the Erlang B Formula

In this section of the paper we will review the $M/M/c/c$ Erlang B model paying attention to the aspects that are relevant to call center modeling. The $M/M/c/c$ queue models a single-node system with c truck lines and no waiting spaces. Figure 2.1 depicts the $M/M/c/c$ queue and figure 3.2, its state transition.

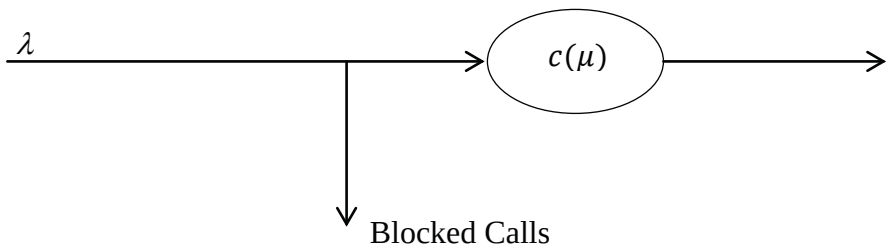


Figure 2.1 : Description of the $M/M/c/c$ Model and its parameters

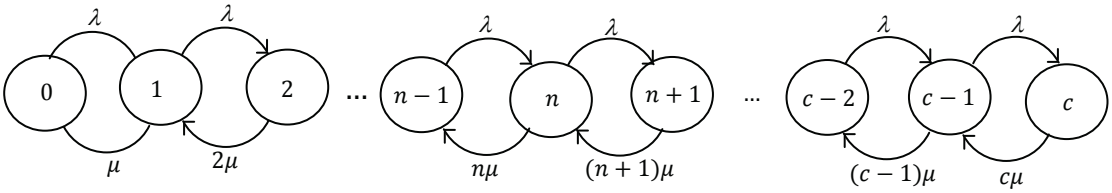


Figure 2.2 : $M/M/c/c$ Flow Rate [Multiple-Server Case($c \geq 2$)]

Considering figures 2.1 and 2.2, it is obvious that $L(t)$ is a finite birth-death process with birth rate

$$\lambda_n = \begin{cases} \lambda, & if\ n < c \\ 0, & if\ n \geq c \end{cases}$$

and state-dependent death rate $\mu_n = n\mu, n = 0,1,2, \ldots, c$.

By the application of the fundamental equation in queueing theory, the steady-state solution of the $M/M/c/c$ model using birth-death process is given by

$$p_n = p_0 \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i}$$

where p_0 is computed from $\sum_{n=0}^{N} p_n = 1$. The solution is given by

$$p_n = \frac{(\lambda/\mu)^n/n!}{\sum_{i=0}^{c}(\lambda/\mu)^i/i!} = \frac{a^n/n!}{\sum_{i=0}^{c} a^i/i!}, \qquad 0 \leq n \leq c$$

where $a := \lambda/\mu$ is called the offered load which describes the demand made on the system.

i. *PASTA: Poisson Arrivals See Time Averages*

An important feature of the Markovian queueing models is that the arrival process follows a Poisson process. Considering the Poisson arrival process, the distribution of customers seen by an arrival to a queueing facility is, stochastically the same as the limiting distribution of customers at that facility. In other words, once the queueing system has reached steady state, each arrival from a Poisson process finds the system at equilibrium. If p_n is the probability that the system contains n customers at equilibrium and a_n denotes the probability that an arriving customer finds $n$customers already present, then PASTA states that $a_n = p_n$. This implies that the Poisson process sees the same distribution as a random observer, i.e., at equilibrium, Poisson arrivals take a random look at the system. This result is a direct consequence of the memoryless property of the interarrival time distribution of customers to a queueing system fed by a Poisson process. In particular, it does not depend on the service time distribution. To prove the PASTA property, we proceed as follows.

Let

$L(t)$ = Number of customers in the system at time t

$p_n(t) = P(System\ is\ in\ state\ n\ at\ time\ t) = P(L(t) = n)$

$a_n(t) = P(Arrival\ at\ time\ t\ finds\ system\ in\ state\ n)$

$A(t, t + \delta t]$ = The event of an arrival in $(t, t + \delta t]$

Then

$$a_n(t) = \lim_{\delta t \to 0} P(L(t) = n | A(t, t + \delta t])$$

$$= \lim_{\delta t \to 0} \frac{P(L(t) = n\ and\ A(t, t + \delta t])}{P(A(t, t + \delta t])}$$

$$= \lim_{\delta t \to 0} \frac{P(A(t, t + \delta t] | L(t) = n) P(L(t) = n)}{P(A(t, t + \delta t])}$$

$$= \lim_{\delta t \to 0} \frac{P(A(t, t + \delta t]) P(L(t) = n)}{P(A(t, t + \delta t])}$$

$$= P(L(t) = n) = p_n(t)$$

The crucial step in the above argument is

$$P(A(t,t+\delta t]|L(t)=n)=P(A(t,t+\delta t])$$

This results from the fact that, since interarrival times possess the memoryless property, $P(A(t,t+\delta t])$ is independent of the past history of the arrival process and hence independent of the current state of the queueing system. With the Poisson arrival process having a constant rate λ, the probability of having an arrival in $(t,t+\delta t]$ is equal to $\lambda\delta t+o(\delta t)$ which does not depend on $L(t)$. Note that the PASTA property only holds for Poisson arrival processes.

The formula for p_c is called "Erlang Loss Formula" and is the fraction of time that all c servers are busy. It denotes the probability that an arrival call finds all the truck line busy, (i.e. the blocking probability, p_c). It is written as $B(c,\lambda/\mu)$s and is called "Erlang B formula":

$$p_c=a_c=P(blocking)=B(c,\lambda/\mu)=B(c,a)=\frac{(\lambda/\mu)^n/n!}{\sum_{n=0}^{c}(\lambda/\mu)^n/n!}=\frac{a^c/c!}{\sum_{n=0}^{c}a^n/n!} \quad (2.1)$$

Notice that the probability that an arrival is lost is equal to the probability that all channels are busy. Erlang loss formula is also valid for the $M/G/c/c$ queue. In other words, the steady-state probabilities are a function only of the mean service time, and not of the complete underlying cumulative distribution function. An efficient recursive algorithm for computing $B(c,a)$ is given by

$$B(0,a)=1,\quad B(c,a)=\frac{aB(c-1,a)}{c+aB(c-1,a)} \quad (2.2)$$

Recall that $a=\lambda/\mu$ is the offered load, we define $a'\coloneqq E(L_b)=a[1-B(c,a)]$ as the carried load, where we obtain the last equality by Little's law applying to number of busy servers and L_b is a random variable representing the number of busy servers in steady-state.

The utilization

$$v\coloneqq\frac{a'}{c}=\frac{a[1-B(c,a)]}{c}=\rho[1-B(c,a)]<1$$

is the fraction of time that a server is busy, where $\rho\coloneqq\frac{a}{c}$ is called the traffic intensity.

Hence we have that

$$1-\frac{1}{\rho}<B(c,a) \quad (2.3)$$

which defines a lower bound for $B(c,a)$

Next, we show the monotonicity property of the $B(c,a)$ with respect to c.

$$B(c,a)=\frac{aB(c-1,a)}{c+aB(c-1,a)}<\frac{aB(c-1,a)}{a[1-B(c-1,a)]+aB(c-1,a)}=B(c-1,a) \quad (2.4)$$

since

$$a[1-B(c-1,a)]<c-1<c$$

We do not consider performance measures relating to waiting time and queue length since there is no waiting space in the $M/M/c/c$ model.

c) Review of the $M/M/c$ Model and the Erlang C Formula

The $M/M/c$ queue can be used to model multiprocessor systems or devices that have several identical servers (or agents) and all jobs (or calls) waiting for these servers are kept in one queue. It is assumed that there are c agents each with a service rate of μ jobs per unit time. The arrival rate is λ calls per unit time. If any of the c agents are idle, the arriving call is serviced immediately. If all c agents are busy, the arriving calls wait in a queue. The state of the system is represented by the number of calls n in the system. The state transition diagram is shown in figure 2.4. It is easy to see that the number of jobs in the system is a birth-death process with the following correspondence:

$$\lambda_n = \lambda, \qquad n \geq 0$$

$$\mu_n = \begin{cases} n\mu, & if\ 1 \leq n \leq c-1 \\ c\mu, & if\ n \geq c \end{cases}$$

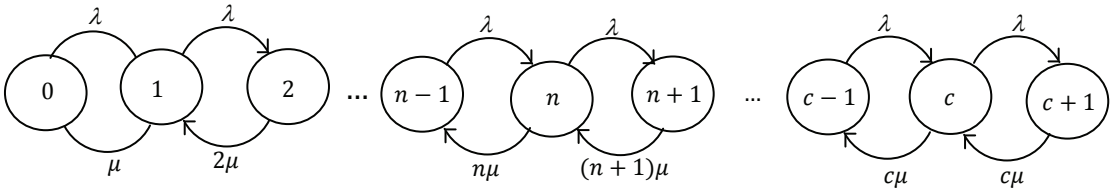


Figure 2.4 : $M/M/c$ Flow Rate [Multiple-ServerCase($c > 1$)]

Owning to the fact that the system is of infinite capacity, the carried load is equal to the offered load, i.e., $a' = a$ so that the utilization $v = \rho$ and as such, we require the stability condition

$$\rho = \frac{\lambda}{c\mu} = \frac{a}{c} < 1$$

Given that the system is stable, the solution to the balance equations obtained from figure 2.4 is

$$p_n = \begin{cases} \frac{a^n}{n!} p_0, & if\ 0 \leq n \leq c \\ \frac{a^n}{c!\, c^{n-c}} p_0, & if\ n \geq c \end{cases} \tag{2.5}$$

with

$$p_0 = \left(\sum_{n=0}^{c-1} \frac{a^n}{n!} + \frac{a^c}{c!} \frac{1}{1-\rho} \right)^{-1} \tag{2.6}$$

Since the arrival process follows the Poisson process, the PASTA property holds so that $a_n = p_n$ for $n \geq 0$. The steady-state probability of waiting of an arriving call is given by the Erlang C formula:

$$P(waiting) = \sum_{n=c}^{\infty} a_n = \sum_{n=c}^{\infty} p_n = C(c,a) = \frac{\frac{a^c}{c!(1-\rho)}}{\sum_{n=0}^{c-1} \frac{a^n}{n!} + \frac{a^c}{c!(1-\rho)}} = \frac{a^c}{c!\,(1-\rho)} p_0 \tag{2.7}$$

So that

$$p_0 = \frac{C(c,a)c!\,(1-\rho)}{a^c} \tag{2.8}$$

The relationship between $C(c,a)$ and $B(c,a)$ is given by

$$C(c,a) = \frac{B(c,a)}{B(c,a)\rho + 1 - \rho}$$

Further simplification using (2.2) yields

$$C(c,a) = \frac{\rho}{\rho + (1-\rho)/B(c-1,a)}$$

From equation 2.5, we it is obvious that $p_c = \frac{a^c}{c!} p_0$, hence by the application of equation 2.8, we obtain $p_c = C(c,a)(1-\rho)$. Using $p_c = \frac{a^c}{c!} p_0$ in equation 2.5 we obtain a new expression for p_n in terms of p_c as follows:

$$p_n = \begin{cases} \frac{c!}{n!\, a^{c-n}} p_c, & if\ 0 \leq n \leq c \\ \rho^{n-c} p_c, & if\ n \geq c \end{cases} \tag{2.9}$$

The new expression for p_c and the above formula for p_n is computationally more efficient, especially for calculating $p_n, n \geq c$ because it does not involve factorials.

i. *The Waiting Time Distribution of the M/M/c Model*

To compute the Telephone Service Factor $TSF = P(W_q \leq AWT)$, we need to compute $P(W_q \leq t)$ which is the steady-state probability of waiting time in queue less than or equal to $t \geq 0$.

Using the concept of total probability, we have that

$$P(W_q > t) = \sum_{n=c}^{\infty} P(W_q > t \mid system\ in\ state\ n\ upon\ arrival) P(system\ in\ state\ n\ upon\ arrival)$$

$$= \sum_{n=c}^{\infty} P(service\ completion\ time\ of\ n-c+1\ calls > t) a_n$$

Using the PASTA property, we can write $P(system\ in\ state\ n\ upon\ arrival) = p_n = a_n$ which is the steady-state probability of an arriving call meeting n calls in the system. Since the service times are exponentially distributed and $n \geq c$, the completion time of $n-c+1$ calls (denoted by) Y_i has an Erlang distribution $Er(n-c+1, c\mu)$ with survival function given by

$$R(t) = P(Y_i > t) = P(service\ completion\ time\ of\ n-c+1\ calls > t) = \sum_{i=0}^{n-c} \frac{(c\mu t)^i e^{-c\mu t}}{i!}$$

Then we have that

$$P(W_q > t) = \sum_{n=c}^{\infty} \sum_{i=0}^{n-c} \frac{(c\mu t)^i e^{-c\mu t}}{i!} p_n = \frac{a^c}{c!\,(1-\rho)} p_0\, e^{-(c\mu-\lambda)t} \tag{2.10}$$

$$E(W_q) = \int_0^{\infty} P(W_q > t)dt = \frac{C(c,a)}{c\mu - \lambda} \tag{2.11}$$

Note that $P(W_q = 0) = 1 - C(c,a)$. By the application of Little's law, we have

$$E(L_q) = \lambda E(W_q) = \frac{\rho C(c,a)}{1-\rho}$$

Because of the closed-form solutions of most the performance indicators of the $M/M/c$ model, it is commonly used in performance modeling and analysis of call centers. In the application of $M/M/c$ model in call center analysis, it is usually assumed that the arrival and service rate are piece-wise constant and time-independent. Using the parameters of each interval, the $M/M/c$ is applied to each time interval. The $M/M/c$ model is not a realistic tool for modeling call centers due to the following reasons:

- It assumes there is no blocking since it has infinite buffer capacity.
- It does not consider the impatience (balking and reneging) attributes of customers.

d) Review of the M/M/c/N Model

When the waiting room in a queueing system has a capacity limit we get a finite queue. In most situations, a finite queue occurs more naturally than a queue with a waiting room of infinite size. However, as the capacity limit gets larger, the behavior of the system approximates that of an infinite-capacity system, and in such cases we are justified in ignoring the size limit. A call center with a finite buffer and several agents is a good example of a finite queueing system. In this section we will review the $M/M/c/N$ model and prove new monotonicity properties of performance measures with respect to N.

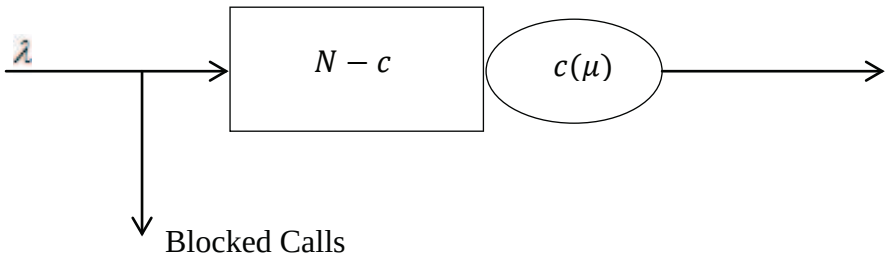


Figure 2.5 : Description of the $M/M/c/N$ Model and its parameters

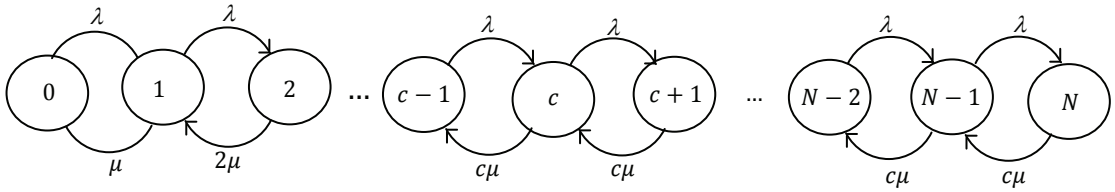


Figure 2.6 : $M/M/c/N$ Flow Rate [Multiple-ServerCase($c > 1$)]

The $M/M/c/N$ queue is similar to the $M/M/c$ queue except that the number of buffers is finite. After $B = N - c$ buffers are full, all arrivals are lost. We assume that B is greater than or equal to c; otherwise, some servers will never be able to operate due to a lack of buffers and the system will effectively operate as a $M/M/B/B$ queue.

The state transition diagram for a $M/M/c/N$ queue is shown in Figure 2.6. The system can be modeled as a birth-death process using the following respective arrival and service rates:

$$\lambda_n = \lambda, \qquad if\ 0 \le n \le N-1$$

$$\mu_n = \begin{cases} n\mu, if\ 0 \le n \le c \\ c\mu, if\ c \le n \le N \end{cases}$$

Solving the balance equations derived from the state diagram, we obtain the following state probabilities.

$$p_n = \begin{cases} \dfrac{a^n}{n!} p_0, & if\ 0 \le n \le c \\ \dfrac{a^n}{c!\, c^{n-c}} p_0, & if\ c \le n \le N \end{cases} \tag{2.12}$$

with

$$p_0 = \begin{cases} \left(\sum_{n=0}^{c-1} \dfrac{a^n}{n!} + \dfrac{a^c}{c!} \dfrac{1-\rho^{N-c+1}}{1-\rho} \right)^{-1}, & if\ \rho \ne 1 \\ \left(\sum_{n=0}^{c-1} \dfrac{a^n}{n!} + \dfrac{a^c}{c!} (N-c+1) \right)^{-1}, & if \rho = 1 \end{cases}$$

and

$$P(waiting) = \sum_{n=c}^{N-1} p_n \,;\ P(blocking) = p_N;\ and\ \ P(no-waiting) = 1 - P(waiting) - P(blocking)$$

i. *The M/M/c/N Waiting Time Distribution*

In this section, we shall provide a mathematical derivation of the waiting time distribution of the $M/M/c/N$ model. Due to the finiteness of the capacity of the $M/M/c/N$ system, deriving the waiting time distribution of the $M/M/c/N$ model is complicated because it results to finite series and also the arrival process is truncated by the system size N. The arrival process no longer follows the Poisson process and has necessitated the need to derive the arrival point probabilities, q_n since $p_n \ne q_n$. In this derivation of q_n, we shall apply the well-known Bayes' theorem.

$$q_n = P(system\ is\ in\ state\ n\ |an\ arrival\ in\ (t, t+\delta t])$$

$$= P(L(t) = n\ |A(t, t+\delta t]) = \frac{P(L(t) = n;\ A(t, t+\delta t])}{P(A(t, t+\delta t])}$$

$$= \frac{P(A(t, t+\delta t]\ |L(t) = n) P(L(t) = n)}{\sum_{n=0}^{N} P(A(t, t+\delta t]|L(t) = n) P(L(t) = n)} = \frac{P(A(t, t+\delta t]\ |L(t) = n) p_n}{\sum_{n=0}^{N} P(A(t, t+\delta t]|L(t) = n) p_n}$$

Taking limits of both sides and using the fact that the probability of an arrival in $(t, t+\delta t]$ is $\lambda\delta t + o(\delta t)$ we have that

$$\lim_{\delta t\to 0} q_n = q_n = \frac{P(A(t,t+\delta t])p_n}{\sum_{n=0}^{N} P(A(t,t+\delta t])p_n} = \lim_{\delta t\to 0}\left(\frac{(\lambda\delta t + o(\delta t))p_n}{\sum_{n=0}^{N-1}(\lambda\Delta t + o(\delta t))\,p_n}\right)$$

$$= \lim_{\delta t\to 0}\left(\frac{\left(\lambda + \frac{o(\delta t)}{\delta t}\right)p_n}{\sum_{n=0}^{N-1}\left(\lambda + \frac{o(\delta t)}{\delta t}\right)p_n}\right) = \frac{\lambda p_n}{1-p_N}, \qquad for\ 0 \le n \le N-1$$

which defines the probability of a call meeting n calls in the system upon arrival given that it is not blocked. Here we have used the fact that

$$\lim_{\delta t\to 0}\frac{o(\delta t)}{\delta t} = 0$$

Using equation (2.12), we can write $p_c = \frac{a^c}{c!}p_0$ which implies that $p_0 = \frac{c!}{a^c}p_c$. Then we can express p_n in terms of p_c as follows:

$$p_n = \begin{cases} \frac{c!}{n!\,a^{c-n}}p_c, & if\ 0 \le n \le c \\ p_c\rho^{n-c}, & if\ c \le n \le N \end{cases}$$

Then for $\rho \neq 1$,

$$1 = \sum_{n=0}^{c}\frac{c!}{n!\,a^{c-n}}p_c + \sum_{n=c+1}^{N} p_c\rho^{n-c}$$

$$p_c = \left(\sum_{n=0}^{c}\frac{c!}{n!\,a^{c-n}} + \sum_{n=c+1}^{N}\rho^{n-c}\right)^{-1} = \left(\frac{1}{B(c,a)} + \frac{\rho(1-\rho^{N-c})}{1-\rho}\right)^{-1}$$

$$p_c = \frac{(1-\rho)B(c,a)}{1-\rho+\rho B(c,a)(1-\rho^{N-c})}$$

For $\rho \neq 1$,

$$P(blocking) = p_N = p_c\rho^{N-c} = \frac{(1-\rho)B(c,a)\rho^{N-c}}{1-\rho+\rho B(c,a)(1-\rho^{N-c})}$$

In same way, for $\rho \neq 1$,

$$P(waiting) = \sum_{n=c}^{N-1} p_n = \sum_{n=c}^{N-1} p_c\rho^{n-c} = \frac{1-\rho^{N-c}}{1-\rho}p_c = \frac{1-\rho^{N-c}}{1-\rho}\frac{(1-\rho)B(c,a)}{1-\rho+\rho B(c,a)(1-\rho^{N-c})}$$

$$P(waiting) = \frac{(1-\rho^{N-c})B(c,a)}{1-\rho+\rho B(c,a)(1-\rho^{N-c})}$$

and

$$P(no-waiting) = 1 - P(waiting) - P(blocking) = \frac{(1-B(c,a))(1-\rho)}{1-\rho+\rho B(c,a)(1-\rho^{N-c})}$$

Now, let us consider computing $P(blocking)$, $P(waiting)$ and $P(no-waiting)$ in the case where $\rho = 1$.

For $\rho = 1$, implies that $a = c$ so that we have

$$p_c = \left(\sum_{n=0}^{c} \frac{c!}{n!\, a^{c-n}} + \sum_{n=c+1}^{N} \rho^{n-c}\right)^{-1} = \left(\frac{1}{B(c,a)} + N - c\right)^{-1} = \frac{B(c,c)}{1 + (N-c)B(c,c)}$$

$$P(blocking) = p_N = p_c \rho^{N-c} = p_c = \frac{B(c,c)}{1 + (N-c)B(c,c)}$$

$$P(waiting) = \sum_{n=c}^{N-1} p_n = \sum_{n=c}^{N-1} p_c \rho^{n-c} = \sum_{n=c}^{N-1} p_c = (N-c)p_c = \frac{(N-c)B(c,c)}{1 + (N-c)B(c,c)}$$

$$P(no - waiting) = 1 - P(waiting) - P(blocking) = \frac{1 - B(c,c)}{1 + (N-c)B(c,c)}$$

Theorem 2.1

Suppose $N = c$ then the $M/M/c/N$ model reduces to the $M/M/c/c$ model with

$$P(blocking) = B(c,a),\ P(waiting) = 0 \text{ and } P(no - waiting) = 1 - B(c,a).$$

Proof

If $N = c$,

$$P(blocking) = p_c = p_N = p_c \rho^{N-c} = \frac{B(c,a)\rho^{N-c}}{1 + (N-c)B(c,a)} = \frac{B(c,a)(1)}{1 + (0)B(c,a)} = B(c,a)$$

$$P(waiting) = \frac{(1 - \rho^{N-c})B(c,a)}{1 - \rho + \rho B(c,a)(1 - \rho^{N-c})} = \frac{(0)B(c,a)}{1 - \rho + \rho B(c,a)(0)} = 0$$

$$P(no - waiting) = 1 - P(waiting) - P(blocking) = 1 - 0 - B(c,a) = 1 - B(c,a)$$

Theorem 2.2

In the limit, as $N \to \infty$, we have the following results:

1. $$\lim_{N\to\infty} P(waiting) = \begin{cases} \frac{1}{\rho}, & if\ \rho > 1 \\ 1, & if\ \rho = 1 \\ C(c,a), & if 0 < \rho < 1 \end{cases}$$

2. $$\lim_{N\to\infty} P(blocking) = \begin{cases} 1 - \frac{1}{\rho}, & if\ \rho \geq 1 \\ 0, & if 0 < \rho \leq 1 \end{cases}$$

3. $$\lim_{N\to\infty} P(no - waiting) = \begin{cases} 0, & if\ \rho \geq 1 \\ 1 - C(c,a), & if 0 < \rho \leq 1 \end{cases}$$

Proof

1. For $0 < \rho < 1$,

$$P(waiting) = \frac{(1-\rho^{N-c})B(c,a)}{1-\rho+\rho B(c,a)(1-\rho^{N-c})} \xrightarrow{as\ N\to\infty} \frac{B(c,a)}{1-\rho+\rho B(c,a)} = C(c,a)$$

For $\rho = 1$,

$$P(waiting) = \frac{(N-c)B(c,c)}{1+(N-c)B(c,c)} = \frac{B(c,c)}{\frac{1}{(N-c)}+B(c,c)} \xrightarrow{as\ N\to\infty} 1$$

For $\rho > 1$,

$$P(waiting) = \frac{(\rho^{N-c}-1)B(c,a)}{\rho+\rho B(c,a)(\rho^{N-c}-1)-1} = \frac{\frac{(\rho^{N-c}-1)B(c,a)}{\rho^{N-c}-1}}{\frac{\rho-1}{\rho^{N-c}-1}+\frac{\rho B(c,a)(\rho^{N-c}-1)}{\rho^{N-c}-1}} = \frac{1}{\rho}$$

So that

$$\lim_{N\to\infty} P(waiting) = \frac{1}{\rho}, \qquad if\ \ \rho > 1$$

2. For $0 < \rho < 1$,

$$P(blocking) = p_N = p_c\rho^{N-c} = \frac{(1-\rho)B(c,a)\rho^{N-c}}{1-\rho+\rho B(c,a)(1-\rho^{N-c})} \xrightarrow{as\ N\to\infty} 0$$

Since $\rho^{N-c} \to 0\ as\ N \to \infty, for\ \ 0 < \rho < 1.$

For $\rho = 1$,

$$P(blocking) = p_N = \frac{B(c,c)}{1+(N-c)B(c,c)} \xrightarrow{as\ N\to\infty} 0$$

For $\rho > 1$,

$$P(blocking) = p_N = \frac{(\rho-1)B(c,a)\rho^{N-c}}{\rho+\rho B(c,a)(\rho^{N-c}-1)-1} = \frac{\frac{(\rho-1)B(c,a)\rho^{N-c}}{\rho^{N-c}}}{\frac{\rho-1}{\rho^{N-c}}+\frac{\rho B(c,a)(\rho^{N-c}-1)}{\rho^{N-c}}}$$

$$= \frac{(\rho-1)B(c,a)}{\frac{\rho-1}{\rho^{N-c}}+\frac{\rho B(c,a)(\rho^{N-c}-1)}{\rho^{N-c}}} \xrightarrow{as\ N\to\infty} \frac{(\rho-1)B(c,a)}{\rho B(c,a)} = 1-\frac{1}{\rho}$$

3. For $\rho > 1$,

$$P(no-waiting) = \frac{\big(1-B(c,a)\big)(\rho-1)}{1-\rho+\rho B(c,a)(1-\rho^{N-c})} = \frac{\frac{\big(1-B(c,a)\big)(\rho-1)}{\rho^{N-c}}}{\frac{\rho+\rho B(c,a)(\rho^{N-c}-1)-1}{\rho^{N-c}}}$$

$$= \frac{\frac{(1-B(c,a))(\rho-1)}{\rho^{N-c}}}{\frac{\rho-1}{\rho^{N-c}}-\frac{\rho B(c,a)}{\rho^{N-c}}+\frac{\rho B(c,a)(\rho^{N-c})}{\rho^{N-c}}} \xrightarrow{as\ N\to\infty} 0$$

For $0 < \rho < 1$,

$$(no - waiting) = \frac{\big(1 - B(c,a)\big)(1-\rho)}{1 - \rho + \rho B(c,a)(1-\rho^{N-c})} \xrightarrow{as\ N\to\infty} \frac{\big(1 - B(c,a)\big)(1-\rho)}{1 - \rho + \rho B(c,a)(1-\rho^{N-c})}$$

$$= 1 - C(c,a)$$

For $\rho = 1$,

$$P(no - waiting) = \frac{1 - B(c,c)}{1 + (N-c)B(c,c)} \xrightarrow{as\ N\to\infty} 0$$

Before we proceed to derive the formula for computing an important performance measure $E(W_q)$, we shall prove some new results that will be useful in the course of our derivations and computations.

For $\rho \neq 1$,

$$P(waiting|no - blocking) = P(W_q > 0) = \sum_{n=c}^{N-1} q_n = \sum_{n=c}^{N-1} \frac{p_n}{1-p_N} = \frac{1}{1-p_N}\sum_{n=c}^{N-1} p_n$$

$$= \frac{1}{1-p_N} P(waiting) = \frac{P(waiting)}{1 - P(blocking)}$$

$$= \frac{1 - \rho + \rho B(c,a)(1-\rho^{N-c})}{1 - \rho + \rho B(c,a)(1-\rho^{N-c-1})} \frac{B(c,a)(1-\rho^{N-c})}{1 - \rho + \rho B(c,a)(1-\rho^{N-c})}$$

$$P(waiting|no - blocking) = P(W_q > 0) = \frac{B(c,a)(1-\rho^{N-c})}{1 - \rho + \rho B(c,a)(1-\rho^{N-c-1})}$$

But

$$P(no - waiting|no - blocking) = P(W_q = 0) = \sum_{n=0}^{c-1} q_n = 1 - P(waiting|no - blocking)$$

$$= \frac{(1-\rho)[1 - B(c,a)]}{1 - \rho + \rho B(c,a)(1-\rho^{N-c-1})}$$

For $\rho = 1$,

$$P(waiting|no - blocking) = \frac{P(waiting)}{1 - P(blocking)} = \frac{\frac{(N-c)B(c,c)}{1+(N-c)B(c,c)}}{1 - \frac{B(c,c)}{1+(N-c)B(c,c)}}$$

$$= \frac{(N-c)B(c,c)}{1 + B(c,c)[N-c-1]}$$

Now, using the principles of conditional probability, we can write

$$P(W_q > t | W_q > 0) = \frac{P(W_q > t;\ W_q > 0)}{P(W_q > 0)} = \frac{P(W_q > t)}{P(W_q > 0)}$$

$$P(W_q > t) = P(W_q > t | W_q > 0) P(W_q > 0)$$

$$P(W_q > t) = P(W_q > t | W_q > 0) P(waiting | no - blocking)$$

For $\rho \neq 1$,

$$P(W_q > t | W_q > 0) = P(W_q > t | c \leq Q \leq N-1)$$

$$= \sum_{n=0}^{N-c+1} P(W_q > t | L = c+n;\ c \leq L \leq N-1) P(L = c+n | c \leq L \leq N-1)$$

$$= \sum_{n=0}^{N-c-1} \left(\sum_{k=0}^{n} \frac{(c\mu t)^k e^{-c\mu t}}{k!} \right) \frac{\rho^n}{1 + \rho + \cdots + \rho^{N-c-1}}$$

$$= \sum_{k=0}^{N-c-1} \frac{(c\mu t)^k e^{-c\mu t}}{k!} \sum_{n=k}^{N-c-1} \frac{\rho^n}{1 + \rho + \cdots + \rho^{N-c-1}}$$

$$P(W_q > t | W_q > 0) = \sum_{k=0}^{N-c-1} \frac{(c\mu t)^k e^{-c\mu t}}{k!} \frac{\rho^k - \rho^{N-c}}{1 - \rho^{N-c}} = \sum_{k=0}^{N-c-1} \frac{(\lambda t)^k e^{-c\mu t}}{k!} \frac{1 - \rho^{N-c-k}}{1 - \rho^{N-c}}$$

Then for $\rho \neq 1$, we have that

$$P(W_q > t) = P(waiting | no - blocking) \sum_{k=0}^{N-c-1} \frac{(\lambda t)^k e^{-c\mu t}}{k!} \frac{1 - \rho^{N-c-k}}{1 - \rho^{N-c}}, t \geq 0 \quad (2.13)$$

Where we have used the fact that[15]

$$P(L = c+n | c \leq L \leq N-1) = \frac{\rho^n}{1 + \rho + \cdots + \rho^{N-c-1}}, \qquad for\ 0 \leq n \leq N-c-1 \quad (2.14)$$

For $\rho = 1$, we also have that

$$P(W_q > t) = P(waiting | no - blocking) \sum_{k=0}^{N-c-1} \frac{(\lambda t)^k e^{-c\mu t}}{k!} \left(1 - \frac{k}{N-c}\right) \quad (2.15)$$

In same line of reasoning, we derive the mathematical formula for computing the Average Speed to Answer (ASA) as follows:

$$ASA = E(W_q) = \int_0^{\infty} P(W_q > t)dt$$

$$= P(waiting|no - blocking) \sum_{k=0}^{N-c-1} \frac{\rho^k - \rho^{N-c}}{1-\rho^{N-c}} \int_0^{\infty} \frac{(c\mu t)^k e^{-c\mu t}}{k!} dt$$

$$= P(waiting|no - blocking) \sum_{k=0}^{N-c-1} \frac{\rho^k - \rho^{N-c}}{(1-\rho^{N-c})c\mu}$$

$$= P(waiting|no - blocking) \frac{1-\rho^{N-c}(1+(1-\rho)(N-c))}{(1-\rho)(1-\rho^{N-c})c\mu}$$

$$= \frac{B(c,a)(1-\rho^{N-c})}{1-\rho+\rho B(c,a)(1-\rho^{N-c-1})} \frac{1-\rho^{N-c}\left(1+(1-\rho)(N-c)\right)}{(1-\rho)(1-\rho^{N-c})c\mu}$$

$$E(W_q) = \frac{(1-\rho^{N-c}(1+(1-\rho)(N-c))B(c,a)}{(1-\rho+\rho B(c,a)(1-\rho^{N-c-1}))(1-\rho)c\mu}$$

By the application of Little's law, we have that

$$E(L_q) = \lambda E(W_q)(1 - P(blocking))$$

III. Limiting Behaviour of the M/M/c/N Model Performance Indicators

In this section of the paper we shall prove some limiting properties of the $M/M/c/$ model with respect to N.

Theorem 3.1

Given that c and other model parameters remain constant, $P(W_q > t)$ is an increasing function of N.

For $0 < \rho < 1$,

Proof

First, we need to show that $P(waiting|no - blocking)$ is an increasing function of N:

$$P(waiting|no - blocking) = \frac{B(c,a)(1-\rho^{N-c})}{1-\rho+\rho B(c,a)(1-\rho^{N-c-1})} = \frac{B(c,a)}{\frac{1-\rho}{(1-\rho^{N-c-1})}+\rho B(c,a)}$$

$$< \frac{B(c,a)}{1-\rho+\rho B(c,a)}, \qquad \text{Since } 1-\rho+\rho B(c,a) > \frac{1-\rho}{(1-\rho^{N-c-1})}+\rho B(c,a)$$

For $\rho = 1$,

$$P(waiting|no-blocking) = \frac{(N-c)B(c,c)}{1+B(c,c)[N-c-1]} = \frac{B(c,c)}{\frac{1}{N-c}+\frac{B(c,c)[N-c-1]}{N-c}}$$

$$\cong \frac{B(c,c)}{\frac{1}{N-c}+B(c,c)} < 1, \qquad \text{since } \frac{1}{N-c}+B(c,c) > B(c,c)$$

Now that we have established the fact that $P(waiting|no-blocking)$ is an increasing function of N, we will proceed to show that $P(W_q > t)$ is an increasing function of N, given that c and other model parameters remain constant.

Recall from equation 2.13; for $0 < \rho < 1$

$$P(W_q > t) = P(waiting|no-blocking)\sum_{k=0}^{N-c-1}\frac{(\lambda t)^k e^{-c\mu t}}{k!}\frac{1-\rho^{N-c-k}}{1-\rho^{N-c}}, t \geq 0$$

and from equation 2.15; for $\rho = 1$, we have

$$P(W_q > t) = P(waiting|no-blocking)\sum_{k=0}^{N-c-1}\frac{(\lambda t)^k e^{-c\mu t}}{k!}\left(1-\frac{k}{N-c}\right)$$

We are only left to show that

$$\gamma(N) := \sum_{k=0}^{N-c-1}\frac{(\lambda t)^k e^{-c\mu t}}{k!}\frac{1-\rho^{N-c-k}}{1-\rho^{N-c}} \quad and \quad \varphi(N) := \sum_{k=0}^{N-c-1}\frac{(\lambda t)^k e^{-c\mu t}}{k!}\left(1-\frac{k}{N-c}\right)$$

are increasing functions of N.

$$\text{For } i \in \mathbb{N}, \qquad \gamma(N+i) := \sum_{k=0}^{N+i-c-1}\frac{(\lambda t)^k e^{-c\mu t}}{k!}\frac{1-\rho^{N+i-c-k}}{1-\rho^{N+i-c}}$$

$$\gamma(N+i) = \sum_{k=0}^{N-c-1}\frac{(\lambda t)^k e^{-c\mu t}}{k!}\frac{1-\rho^{N+i-c-k}}{1-\rho^{N+i+c}} + \sum_{k=N-c}^{N+i-c-1}\frac{(\lambda t)^k e^{-c\mu t}}{k!}\frac{1-\rho^{N+i-c-k}}{1-\rho^{N+i-c}}$$

$$> \sum_{k=0}^{N-c-1}\frac{(\lambda t)^k e^{-c\mu t}}{k!}\frac{1-\rho^{N-c-k}}{1-\rho^{N-c}} = \gamma(N)$$

since

$$\sum_{k=N-c}^{N+i-c-1}\frac{(\lambda t)^k e^{-c\mu t}}{k!}\frac{1-\rho^{N+i-c-k}}{1-\rho^{N+i-c}} > 0 \quad and \quad \frac{1-\rho^{N+i-c-k}}{1-\rho^{N+i+c}} \geq \frac{1-\rho^{N-c-k}}{1-\rho^{N-c}}$$

In same way,

$$\text{For } i \in \mathbb{N}, \qquad \varphi(N+i) := \sum_{k=0}^{N+i-c-1} \frac{(\lambda t)^k e^{-c\mu t}}{k!}\left(1 - \frac{k}{N+i-c}\right)$$

$$\varphi(N+i) = \sum_{k=0}^{N-c-1} \frac{(\lambda t)^k e^{-c\mu t}}{k!}\left(1 - \frac{k}{N+i-c}\right) + \sum_{k=N-c}^{N+i-c-1} \frac{(\lambda t)^k e^{-c\mu t}}{k!}\left(1 - \frac{k}{N+i-c}\right)$$

$$> \sum_{k=0}^{N-c-1} \frac{(\lambda t)^k e^{-c\mu t}}{k!}\left(1 - \frac{k}{N-c}\right) = \varphi(N)$$

since

$$\sum_{k=N-c}^{N+i-c-1} \frac{(\lambda t)^k e^{-c\mu t}}{k!}\left(1 - \frac{k}{N+i-c}\right) > 0 \; and \; 1 - \frac{k}{N+i-c} \geq 1 - \frac{k}{N-c}, 0 \leq k \leq N-c-1$$

IV. Conclusions

In this paper, we have discussed in detail the modeling of a call center as single-node using the Markovian queueing techniques. We considered the $M/M/c/c$ Erlang B Loss model and the $M/M/c/c$ Erlang C model as well as the more general $M/M/c/N$ model. Our emphasis is on the derivation of the exact performance measures of these well-known models. Considering the $M/M/c/N$ model, we expressed the system performance measures in terms of Erlang B formula, which facilitates the computation as well as the analysis. Using the results emanating from the analysis, we showed the monotonicity properties for performance measures with respect to N and c.

References Références Referencias

1. Adan I. and Jacques R. Queueing Theory, Eindhoven, Netherlands, 2002.
2. Boucherie, R. Product-Form in Queueing Networks. Ph.D. Thesis, Free University, Amsterdam, The Netherlands, 1992.
3. Bolch, G., Greiner, S., de Meer, H., and Trivedi, K. S. Queueing networks and Markov chains: modeling and performance evaluation with computer science applications. Wiley Interscience, New York, NY, USA, 1998.
4. Borst, S., Mandelbaum, A., and Reiman, M. Dimensioning large call centers. Operations Research 52, 17-34, 2004.
5. Chen, H., and Yao, D. Fundamentals of queueing networks: performance, asymptotics and optimization. Springer-Verlag, New York, 2001.
6. Cooper, R. Introduction to Queueing Theory. North-Holland, New York, 1981.
7. Gans, N., Koole, G., and Mandelbaum, A. Telephone call centers: Tutorial, review, and research prospects. Manufacturing and Service Operations Management 5, 79-141, 2003.
8. Gnedenko, B., and Kovalenko, I. Introduction to Queueing Theory. Birkhauser Boston Inc., Cambridge, MA, 1968.
9. Donald Gross, John F. Shortle, James M. Thompson, and Carl M. Harris. Fundamentals of Queueing Theory, third ed. Wiley, New York, NY, 1998.
10. Kleinrock, L. Queueing Systems, Vol. 1. Wiley, New York, NY, 1975.
11. Koole, G., and Mandelbaum, A. Queueing models of call centers: An introduction. Annals of Operations Research 113, 41-59, 2002.
12. Mandelbaum, A. Call centers (centres) research bibliography with abstracts. Tech. rep., Technion, Haifa, Israel, Version 7: May 4, 2006.
13. Mandelbaum, A., and Zeltyn, S. The Palm/Erlang-A queue, with applications to call centers. Tech. rep., Service Engineering Lecture Notes, 2004.
14. Ross, S. M. Stochastic Processes. Second Edition. Wiley Inc, New York, 1996.
15. Ross, S. M. Introduction to Probability Models, Tenth Edition. Academic Press, 2011.
16. Srinivasan, R., Talim, J., and Wang, J. Performance analysis of a call center with interactive voice response units. TOP: An Official Journal of the Spanish Society of Statistics and Operations Research 12, 91-110, 2004.

17. Zhidong Zhang. Call centres with balking and abandonment: from queueing to queueing network models. Ph.D. Thesis, University of Saskatchewan, Saskatoon, Saskatchewan, 2010.

Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 13 Issue 10 Version 1.0 Year 2013
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Feature Selection Algorithm for High Dimensional Data using Fuzzy Logic

By T. Jaga Priya Vathana, C. Saravanabhavan & Dr. J. Vellingiri
Anna University, India

Abstract - Feature subset selection is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy and improving results comprehensibility. This process improved by cluster based FAST Algorithm and Fuzzy Logic. FAST Algorithm can be used to Identify and removing the irrelevant data set. This algorithm process implements using two different steps that is graph theoretic clustering methods and representative feature cluster is selected. Feature subset selection research has focused on searching for relevant features. The proposed fuzzy logic has focused on minimized redundant data set and improves the feature subset accuracy.

GJCST-C Classification : 1.5

Strictly as per the compliance and regulations of:

Feature Selection Algorithm for High Dimensional Data using Fuzzy Logic

T. Jaga Priya Vathana [α], C. Saravanabhavan [σ] & Dr. J. Vellingiri [ρ]

Abstract - **Feature subset selection is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy and improving results comprehensibility. This process improved by cluster based FAST Algorithm and Fuzzy Logic. FAST Algorithm can be used to Identify and removing the irrelevant data set. This algorithm process implements using two different steps that is graph theoretic clustering methods and representative feature cluster is selected. Feature subset selection research has focused on searching for relevant features. The proposed fuzzy logic has focused on minimized redundant data set and improves the feature subset accuracy.**

I. Introduction

The performance, robustness, and usefulness of classification algorithms are improved when relatively few features are involved in the classification. Thus, selecting relevant features for the construction of classifiers has received a great deal of attention.

With the aim of choosing a subset of good features with respect to the target concepts, feature subset selection is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result compre-hensibility. Many feature subset selection methods have been proposed and studied for machine learning applications. They can be divided into four broad categories: the Embedded, Wrapper, Filter, and Hybrid approaches. The embedded methods incorporate feature selection as a part of the training process and are usually specific to given learning algorithms, and therefore may be more efficient than the other three categories. Traditional machine learning algorithms like decision trees or artificial neural networks are examples of embedded approaches. The wrapper methods use the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected sub-sets, the accuracy of the learning algorithms is usually high. However, the generality of the selected features is limited and the computational complexity is large. The filter methods are independent of learning algorithms, with good generality.

With respect to the filter feature selection methods, the application of cluster analysis has been demonstrated to be more effective than traditional feature selection algorithms. Pereira et al., Baker et al., and Dillon et al. employed the distributional clustering of words to reduce the dimensionality of text data. In cluster analysis, graph-theoretic methods have been well studied and used in many applications. Their results have, sometimes, the best agreement with human performance. The general graph-theoretic clustering is simple: Compute a neighborhood graph of in-stances, then delete any edge in the graph that is much longer/shorter (according to some criterion) than its neighbors. The result is a forest and each tree in the forest represents a cluster. In our study, we apply graph-theoretic clustering methods to features. In particular, we adopt the minimum spanning tree (MST) based clustering algorithms, because they do not assume that data points are grouped around centers or separated by a regular geometric curve and have been widely used in practice. Based on the MST method, we propose a FAST clustering-Based feature Selection algorithm (FAST).The FAST algorithm works in two steps. In the first step, features are divided into clusters by using graph-theoretic clustering methods. In the second step, the most representative feature that is strongly related to target classes is selected from each cluster to form the final subset of features. Features in different clusters are relatively independent; the clustering-based strategy of FAST has a high probability of producing a subset of useful and independent features. The proposed feature subset se-lection algorithm FAST was tested upon 35 publicly available image, microarray, and text data sets. The Experimental results show that, compared with other five different types of feature subset selection algorithms, the proposed algorithm not only reduces the number of features, but also improves the performances of the four well-known different types of classifiers.

II. Literature Review

a) Statistical Comparisons of Classifiers over Multiple Data Sets

In this method introduce some new pre- or post processing step has been proposed, and the implicit hypothesis is made that such an enhancement yields an improved performance over the existing classification

Author α : M.E-Student, Department of CSE, Kongunadu College of Engineering and Technology, Tamil Nadu, India.
E-mail : vathanatamil@gmail.com
Author σ : Research Scholars & Asst. Professor, Department of CSE, Kongunadu College of Engineering and Technology, Tamil Nadu, India. E-mail : profbhavan@gmail.com
Author ρ : Associate Professor, Department of CSE, Kongunadu College of Engineering and Technology, Tamil Nadu, India.
E-mail : vellingiri.research@gmail.com

algorithm. Alternatively, various solutions to a problem are proposed and the goal is to tell the successful from the failed. A number of test data sets is selected for testing, the algorithms are run and the quality of the resulting models is evaluated using an appropriate measure, most commonly classification accuracy. The remaining step, and the topic of this paper, is to statistically verify the hypothesis of improved performance. Various re-searchers have addressed the problem of comparing two classifiers on a single data set and proposed several solutions. The core of the paper is the study of the statistical tests that could be (or already are) used for comparing two or more classifiers on multiple data sets. Learning algorithms is used for the Classification purpose. The main disadvantage of this process is the problems with the multiple data set tests are quite different, even in a sense complementary.

b) A Features Set Measure Based on Relief

It used six real world dataset from the UCI repository have been used. Three of them have classification Problem with discrete features, the next two classifications with discrete and continuous features, and the last one is approximation problem. The learning algorithm is used to check the quality of feature selected are a classification and regression tree layer with pruning. This process and algorithms is implemented by the orange data mining System. Overall, the non-parametric tests, namely the Wilcox on and Friedman test are suitable for our problems. They are appropriate since they assume some, but limited commensurability. They are safer than parametric tests since they do not assume normal distributions or homogeneity of variance. There is an alternative opinion among statisticians that significance tests should not be per-formed at all since they are often misused, either due to misinterpretation or by putting too much stress on their results. The main disadvantage of the system is it measure to low accuracy of the search process.

c) Feature Clustering and Mutual Information for the Selection of Variables In Spectral Data

It face many problems in spectrometry require predicting a quantitative value from measured spectra. The major issue with spectrometric data is their functional nature; they are functions discredited with a high resolution. This leads to a large number of highly-correlated features; many of which are irrelevant for the prediction. The approach for the features is to describe the spectra in a functional basis whose basis functions are local in the sense that they correspond to well-defined portions of the spectra. This process has clustering algorithm that algorithm recursively merges at each step the two most similar consecutive clusters. This algorithm return the output value associated with each cluster, its representative, is chosen to be the mean of the spectra over the range of features defined by the cluster. The main disadvantage of the problem is low number of clusters identified by the method allows the interpretation of the selected variables: several of the selected clusters include the spectral variables identified on these benchmarks as meaningful in the literature.

d) On Feature Selection through Clustering

This paper introduce an algorithm for feature selection that clusters attributes using a special metric and, then uses a hierarchical clustering for feature selection. Hierarchical algorithms generate clusters that are placed in a cluster tree, which is commonly known as a dendrogram. Clustering's are obtained by extracting those clusters that are situated at a given height in this tree. It use several data sets from the UCI dataset repository and, due to space limitations we discuss only the results obtained with the votes and zoo datasets, Bayes algorithms of the WEKA package were used for constructing classifiers on data sets obtained by projecting the initial data sets on the sets of representative attributes. Approach to attribute selection is the possibility of the supervision of the process allowing the user to opt between quasi-equivalent attributes It face classification problems that involve thousands of features and relatively few examples came to the fore. We intend to apply our techniques to this type of data.

III. Fuzzy based Feature Subset Selection Algorithms

Irrelevant features, along with redundant features, severely affect the accuracy of the learning machines. Thus, feature subset selection should be able to identify and remove as much of the irrelevant and redundant information as possible. The cluster indexing and document assignments are repeated periodically to compensate churn and to maintain an up-to-date clustering solution. The k-means clustering technique and SPSS Tool to develop a real time and online system for a particular supermarket to predict sales in various annual seasonal cycles. The classification was based on nearest mean.

In order to more precisely introduce the algorithm, and because our proposed feature subset selection framework involves irrelevant feature removal and redundant feature elimination.

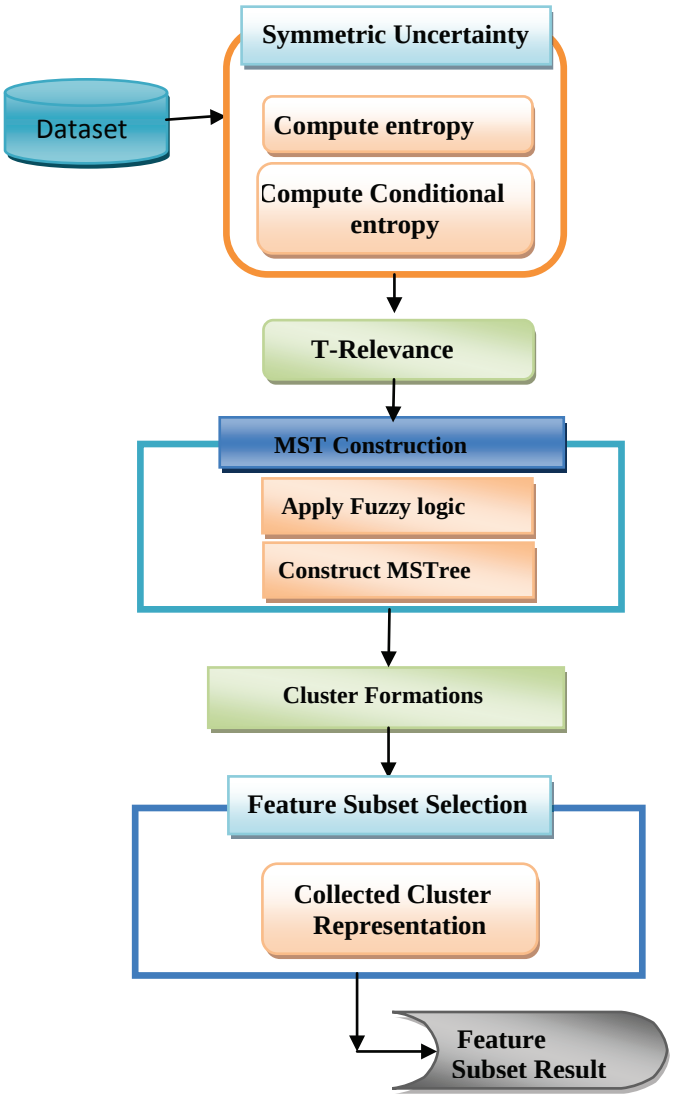


Figure 1 : Framework of the Fuzzy Based

Feature subset selection algorithm Irrelevant features, along with redundant features, severely affect the accuracy of the learning machines, Thus, feature subset selection should be able to identify and Remove as much of the irrelevant and redundant information as possible. Moreover, "good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other. Keeping these in mind, we develop a novel algorithm which can efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset. We achieve this through a new feature selection framework which composed of the two connected components of irrelevant feature removal and redundant feature elimination. The former obtains features relevant to the target concept by eliminating irrelevant ones, and the latter removes redundant features from relevant ones via choosing representatives from different feature clusters, and thus produces the final subset.

The irrelevant feature removal is straightforward once the right relevance measure is defined or selected, while the redundant feature elimination is a bit of sophisticated. In our proposed FAST algorithm, it involves (a) the construction of the minimum spanning tree (MST) from a weighted complete graph; (b) the partitioning of the MST into a forest with each tree representing a cluster; and (c) the selection of representative features from the clusters.

In order to more precisely introduce the algorithm, and because our proposed feature subset selection framework involves irrelevant feature removal and redundant feature elimination, we firstly present the traditional definitions of relevant and redundant features, then provide our definitions based on variable correlation as follows.

John et al. presented a definition of relevant features. Suppose F to be the full set of features, $i \in F$ be a feature, $Si = F - \{Fi\}$ and $S'i \subseteq Si$. Let $s'i$ be a value-assignment of all features in $S'i$, fi a value-assignment of feature Fi, and c a value-assignment of the target concept C. The definition can be formalized as follows.

Definition: (Relevant feature) Fi is relevant to the target concept C if and only if there exists some $s'i$, fi and c, such that, for probability $(S'i = s'i, Fi = fi) > 0$, $p(C = c \mid S'i = s'i, Fi = fi) \neq p(C = c \mid S'i = si)$. Otherwise, feature Fi is an irrelevant feature. Definition 1 indicates that there are two kinds of relevant features due to different $S'i$: (i) when $S'i = Si$, from the definition we can know that Fi is directly relevant to the target concept; (ii) when $S'i \subsetneq Si$, from the definition we may obtain that $p(C \mid Si, Fi) = p(C \mid Si)$. It seems that Fi is irrelevant to the target concept. However, the definition shows that feature Fi is relevant when using $S'i \cup \{Fi\}$ to describe the target concept. The reason behind is that either Fi is interactive with $S'i$ or Fi is redundant with $Si - S'i$. In this case, we say Fi is indirectly relevant to the target concept. Most of the information contained in redundant features is already present in other features. As a result, redundant features do not contribute to getting better interpreting ability to the target concept. It is formally defined by Yu and Liu based on Markov blanket. The definitions of Markov blanket and redundant feature are introduced as follows, respectively.

Let $Mi \subset F (Fi \notin Mi)$, Mi is said to be a Markov blanket for Fi if and only if $p(F - Mi - \{Fi\}, C \mid Fi, Mi) = p(F - Mi - \{Fi\}, C \mid Mi)$. Definition: (Redundant feature). Let S be a set of features, a feature in S is redundant if and only if it has a Markov Blanket within S. Relevant features have strong correlation with target concept so are always necessary for a best subset, while redundant features are not because their values are completely correlated with each other. Thus, notions of feature redundancy and feature relevance are normally in terms of feature correlation and feature-target concept correlation.

Mutual information measures how much the distribution of the feature values and target classes differ from statistical independence. This is a nonlinear estimation of correlation between feature values or feature values and target classes. The symmetric uncertainty (SU) is derived from the mutual information by normalizing it to the entropies of feature values or feature values and target classes, and has been used to evaluate the goodness of features for classification by a number of researchers (e.g., Hall], Hall and Smith, Yu and Liu,, Zhao and Liu,). Therefore, we choose symmetric uncertainty as the measure of correlation between either two features or a feature and the target concept.

The symmetric uncertainty is defined as follows $S(X, Y)=2\times Gain(X|Y)\ H(X)+H(Y)$.

Where,

1. (X)is the entropy of a discrete random variable X. Suppose (x) is the prior probabilities for all values of$X$, $H(X)$is defined by $H(X)=-\sum x\in X\ p(x)\log 2p(x)$.
2. Gain ($X|Y$) is the amount by which the entropy of Y decreases. It reflects the additional information about$Y$provided by$X$and is called the information gain which is given by $Gai(X|Y)=H(X)-H(X|Y)\ =\ H(Y)-H(Y|X)$.

Where ($X|Y$) is the conditional entropy which Quantifies the remaining entropy (i.e. uncertainty) of a random variable X given that the value of another random variable Y is known. Suppose$p(x)$ is the prior probabilities for all values of $X$and $p(x|y)$is the posterior probabilities of Xgiven the values of Y, $H(X|Y)$is defined by $H(X|Y)=-\sum y\in Y\ p(y)\sum x\in X\ p(x|y)\log 2p(x|y)$. (4) Information gain is a symmetrical measure. That is the amount of information gained about Xafter observing$Y$is equal to the amount of information gained about$Y$ after observing X. This ensures that the order of two variables (e.g., (X, Y) or (Y,X)) will not affect the value of the measure.

Symmetric uncertainty treats a pair of variables sym-metrically, it compensates for information gain's bias toward variables with more values and normalizes its value to the range [0,1]. A value 1 of $S(X,\ Y)$ indicates. That knowledge of the value of either one completely predicts the value of the other and the value 0 reveals that $X$and$Y$ are independent. Although the entropy-based measure handles nominal or discrete variables, they can deal with continuous features as well, if the values are discredited properly in advance. Given $SU(X,\ Y)$ the symmetric uncertainty of variables $X$and Y, the relevance T-Relevance between a feature and the target concept C, the correlation F-Correlation between a pair of features, the feature Redundancy F-Redundancy and the representative feature R-Feature of a feature cluster can be defined as follows.

Definition: (T-Relevance) The relevance between the feature $Fi\in F$ and the target concept C is referred to as The T-Relevance of Fi and$C$, and denoted by $S(Fi,C)$. If $S(Fi,C)$ is greater than a predetermined threshold θ, we say that Fi is a strong T-Relevance feature.

Definition: (F-Correlation) The correlation between any pair of features Fi and Fj ($Fi,\ j\in F\wedge i\neq j$) is called the F-Correlation of Fi andFj, and denoted by $SU(Fi,Fj)$.

$Fk<|F|\}$ be a cluster of features. if $\exists Fj\ \in\ S$, $SU(Fj,C)\ \geq\ \ SU(Fi,C)\wedge SU(Fi,Fj)\ >\ SU(Fi,C)$ is always corrected for each $Fi\in S(i\neq j)$, then Fi are redundant features with respect to the givenFj (i.e. each Fi is a F-Redundancy).

Definition: (R-Feature) A feature $Fi\in S=\{F1,2,\ldots,Fk\}(k<|F|)$ is a representative feature of the cluster S(i.e. Fi is a R-Feature) if and only if, $Fi\ =$ argmax $Fj\in SSU(Fj,C)$. This means the feature, which has the strongest T-Relevance, can act as a R-Feature for all the features in the cluster. According to the above definitions, feature subset selection can be the process that identifies and retains the strong T-Relevance features and selects R-Features from feature clusters. The behind heuristics are that

1. Irrelevant features have no/weak correlation with Target concept;
2. Redundant features are assembled in a cluster and a representative feature can be taken out of the Cluster.

IV. Algorithm and Analysis

The proposed FAST algorithm logically consists of three steps:

1. removing irrelevant features,
2. constructing a MST from relative ones,
3. Partitioning the MST and selecting Representative features.

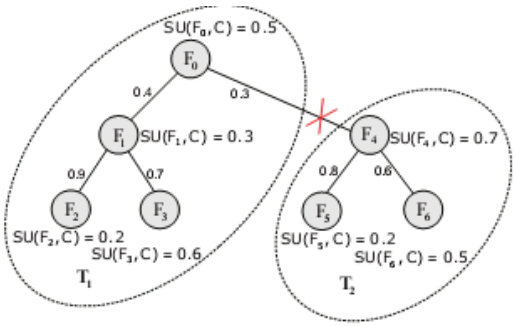


Figure 2 : Example of Clustering Step

After removing all the unnecessary edges, a forest is obtained. Each tree Forest represents a cluster hat is denoted as which is the vertex set of as well. As illustrated above, the features in each cluster are redundant.

The details of the FAST algorithm is shown in Algorithm 1.

Algorithm 1: FAST

```
inputs: D(F1, F2, ..., Fm, C) - the given data set
        θ - the T-Relevance threshold.
output: S - selected feature subset .
//==== Part 1 : Irrelevant Feature Removal ====
for i = 1 to m do
    T-Relevance = SU (Fi, C)
    if T-Relevance > θ then
        S = S ∪ {Fi};
//==== Part 2 : Minimum Spanning Tree Construction ====
G = NULL; //G is a complete graph
for each pair of features {F'i, F'j} ⊂ S do
    F-Correlation = SU (F'i, F'j)
    Add F'i and/or F'j to G with F-Correlation as the weight of
    the corresponding edge;
minSpanTree = Prim (G); //Using Prim Algorithm to generate the
    minimum spanning tree
//==== Part 3 : Tree Partition and Representative Feature Selection ====
Forest = minSpanTree
for each edge Eij ∈ Forest do
    if SU(F'i, F'j) < SU(F'i, C) ∧ SU(F'i, F'j) < SU(F'j, C) then
        Forest = Forest − Eij
S = ϕ
for each tree Ti ∈ Forest do
    F^j_R = argmax_{F'k ∈ Ti} SU(F'k, C)
    S = S ∪ {F^j_R};
return S
```

Time complexity analysis. The major amount of work for Algorithm 1 involves the computation of SU values for T-Relevance and F-Correlation, which has linear complexity in terms of the number of instances in a given data set. The first part of the algorithm has a linear time complexity (m) in terms of the number of features m. Assuming ($1 \leq k \leq m$) features are selected as relevant ones in the first part, when k =1, only one feature is selected. Thus, there is no need to continue the rest parts of the algorithm, and the complexity is(m). When $1<k\leq m$, the second part of the algorithm firstly constructs a complete graph from relevant features and the complexity is (k 2), and then generates a MST from the graph using Prim algorithm whose time complexity. The third part partitions the MST and chooses the representative features with the complexity. Thus when the complexity of the algorithm. This means when FAST has linear complexity while obtains the worst complexity when. However is heuristically set to be in the implementation of FAST. So the complexity, which is typically less than since. This can be explained as follows.

V. Data Source

For the purposes of evaluating the performance and effectiveness of our proposed FAST algorithm, verifying whether or not the method is potentially useful in practice, and allowing other researchers to confirm our results, 35 publicly available data sets were used. The numbers of features of the 35 data sets vary from 37 to 49152 with a mean of 7874. The dimensionality of the 54.3% data sets exceed 5000, of which 28.6% data sets have more than 10000 features. The 35 data sets cover a range of application domains such as text, image and bio microarray data classification.

VI. Experiment Setup

To evaluate the performance of our proposed FAST algorithm and compare it with other feature selection. Algorithms in a fair and reasonable way, we set up our experimental study as follows. 1) The proposed algorithm is compared with five different types of representative feature selection algorithms. They are (i) FCBF, (ii) Relief, (iii) CFS, (iv) Consist and (v) FOCUS SF [2], respectively. FCBF and Relief evaluate features individually. Relief searches for nearest neighbors of instances of different classes and weights features according to how well they differentiate instances of different classes. The other three feature selection algorithms are based on subset evaluation. CFS exploits best-first search based on the evaluation of a subset that contains features highly correlated with the tar-get concept, yet uncorrelated with each other. The Consist method searches for the minimal subset that separates classes as consistently as the full set can under best-first search strategy. FOCUS-SF is a variation of FOCUS [2]. FOCUS has the same evaluation strategy as Consist, but it examines all subsets of features. Considering the time efficiency, FOUCS-SF replaces exhaustive search in FOCUS with sequential forward selection.

Four different types of classification algorithms are employed to classify data sets before and after feature selection. They are (i) the probability-based Naive Bayes (NB), (ii) the tree-based C4.5, (iii) the instance-based lazy learning algorithm IB1, and (iv) the rule-based RIPPER, respectively. Naive Bayes utilizes a probabilistic method for classification by multiplying the individual probabilities of every feature-value pair. This algorithm assumes independence among the features and even then provides excellent classification results. Decision tree learning algorithm C4.5 is an extension of ID3 that accounts for unavailable values, continuous attribute value ranges, pruning of decision trees, rule derivation, and so on. The tree comprises of nodes (features) that are selected by information entropy. Instance-based learner IB1 is a single-nearest-neighbor algorithm, and it classifies entities taking the class of the closest associated vectors in the training set via

distance metrics. It is the simplest among the algorithms used in our study. Inductive rule learner RIPPER (Repeated Incremental Pruning to Produce Error Reduction) is a propositional rule learner that defines a rule based detection model and seeks to improve it iteratively by using different heuristic techniques. The constructed rule set is then used to classify new instances.

3) When evaluating the performance of the feature subset selection algorithms, four metrics, (i) the proportion of selected features (ii) the time to obtain the feature subset, (iii) the classification accuracy, and (iv) the Win/Draw/Loss record, are used. The proportion of selected features is the ratio of the number of features selected by a feature selection algorithm to the original number of features of a data set. The Win/Draw/Loss record presents three values on a given measure, i.e. the numbers of data sets for which our proposed algorithm FAST obtains better, equal, and worse performance than other five feature selection algorithms, respectively. The measure can be the proportion of selected features, the runtime to obtain a feature subset, and the classification accuracy, respectively.

VII. Results and Analysis

In this paper present the experimental results in terms of the proportion of selected features, the time to obtain the feature subset, the classification accuracy, and the Win/Draw/Loss record. For the purpose of exploring the statistical significance of the results, we performed a nonparametric Friedman test followed by Nemenyi post-hoc test, as advised by Demsar and Garcia and Herrerato to statistically compare algorithms on multiple data sets. Thus the Friedman and the Nemenyi test results are reported as well

a) Proportion of selected features

Records the proportion of selected features of the six feature selection algorithms for each data set. From it we observe that) generally all the six algorithms achieve significant reduction of dimensionality by selecting only a small portion of the original features. FAST on average obtains the best proportion of selected features of 1.82%. The Win/Draw/Loss records show FAST wins other algorithms as well. 2) For image data, the proportion of selected features of each algorithm has an increment compared with the corresponding average proportion of selected features on the given data sets except Consist has an improvement. This reveals that the five algorithms are not very suitable to choose features for image data compared with for microarray and text data. FAST ranks 3 with the proportion of selected features of 3.59% that has a tiny margin of 0.11% to the first and second best proportion of selected features 3.48% of Consist and FOCUS-SF, and a margin of 76.59% to the worst proportion of selected features 79.85% of Relief. 3) For microarray data, the proportion of selected features has been improved by each of the six algorithms compared with that on the given data sets. This indicates that the six algorithms work well with microarray data. FAST ranks 1 again with the proportion of selected features of 0.71%. Of the six algorithms, only CFS cannot choose features for two data sets whose dimensionalities are 19994 and 49152, respectively. 4) For text data, FAST ranks 1 again with a margin of 0.48% to the second best algorithm FOCUS-SF. TABLE 2: Proportion of selected features of the six feature selection algorithms.

The Friedman test can be used to compare k algorithms over Ndata sets by ranking each algorithm on each data set separately. The algorithm obtained the best performance gets the rank of 1, the second best ranks 2, and so on. In case of ties, average ranks are assigned. Then the average ranks of all algorithms on all data sets are calculated and compared. If the null hypothesis, which is all algorithms are performing equivalently, is rejected under the Friedman test statistic, post-hoc tests such as the Nemenyi test can be used to determine which algorithms perform statistically different. The Nemenyi test compares classifiers in a pairwise manner. In order to further explore whether the reduction rates are significantly different we performed a Friedman test followed by a Nemenyi post-hoc test. The null hypothesis of the Friedman test is that all the feature selection algorithms are equivalent in terms of proportion of selected features. The test result isp=0. This means that at α= 0.1, there is evidence to reject the null hypothesis and all the six feature selection algorithms are different in terms of proportion of selected features

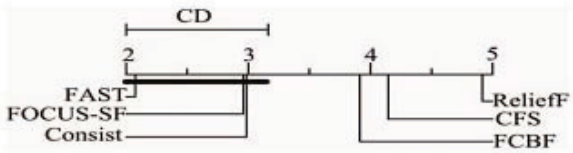


Figure 3 : Proportion of selected features

Comparison of all feature selection algorithms against each other with the Nemenyi test.

In order to further explore feature selection algorithms whose reduction rates have statistically significant differences, we performed a Nemenyi test. Fig. 3 shows the results with α= 0.1 on the 35 data sets. The results indicate that the proportion of selected features of FAST is statistically smaller than those of Relief, CFS and FCBF, and there is no consistent evidence to indicate statistical differences between FAST, Consist, and FOCUS-SF, respectively.

The 10-fold cross-validation accuracies of the four different types of classifiers on the 35 data sets before and after each feature selection algorithm is performed, respectively. The classification accuracy of Naïve Bayes. From it we observe that:

1. Compared with original data, the classification accuracy of Naive Bayes has been improved by FAST, CFS, and FCBF by 12.86%, 6.62%, and 4.32%, respectively. Unfortunately, Relief, Consist, and FOCUS-SF have decreased the classification accuracy by 0.32%, 1.35%, and 0.86%, respectively. FAST ranks 1 with a margin of 6.24% to the second best accuracy 80.60% of CFS. At the same time, the Win/Draw/Loss records show that FAST outer forms all other five algorithms.
2. For image data, the classification accuracy of Naïve Bayes has been improved by FCBF, CFS, FAST, and Relief by 6.13%, 5.39%, 4.29%, and 3.78%, respectively. However, Consist and FOCUS-SF have decreased the classification accuracy by 4.69% and 4.69%, respectively. This time FAST ranks 3 with a margin of 1.83% to the best accuracy 87.32% of FCBF.
3. For microarray data, the classification accuracy of Naive Bayes has been improved by all the six algorithms FAST, CFS, FCBF, ReliefF, Consist, and FOCUS-SF by 16.24%, 12.09%, 9.16%, 4.08%, 4.45%, and 4.45%, respectively. FAST ranks 1 with a mar-gin of 4.16% to the second best accuracy 87.22% of CFS. This indicates that FAST is more effective than others when using Naive Bayes to classify microarray data.
4. For text data, FAST and CFS have improved the classification accuracy of Naive Bayes by 13.83% and 1.33%, respectively. Other four algorithms ReliefF, Consist, FOCUS-SF, and FCBF have decreased the accuracy by 7.36%, 5.87%, 4.57%, and 1.96%, respectively. FAST ranks 1 with a margin of 12.50% to the second best accuracy 70.12% of CFS.

Selection algorithms FAST, FCBF, CFS, Relief, Consist, and FOCUS-SF by 5.31%, 4.54%, 7.20%, 0.73%, 0.60%, and 0.60%, respectively. This time FAST ranks 2 with a margin of 1.89% to the best accuracy 83.6% of CFS and a margin of 4.71% to the worst accuracy 76.99% of Consist and FOCUS-SF. 3) For microarray data, the classification accuracy of C4.5 has been improved by all the six algorithms FAST, FCBF, CFS, Relief, Consist, and FOCUS-SF by 11.42%, 7.14%, 7.51%, 2.00%, 6.34%, and 6.34%, respectively. FAST ranks 1 with a margin of 3.92% to the second best accuracy 79.85% of CFS. 4) For text data, the classification accuracy of C4.5 has been decreased by algorithms FAST, FCBF, CFS, ReliefF, Consist and FOCUS-SF by 4.46%, 2.70%, 19.68%, 13.25%, 16.75%, and 1.90% respectively. FAST ranks 3 with a margin of 2.56% to the best accuracy 83.94% of FOCUS-SF and a margin of 15.22% to the worst accuracy 66.16% of CFS.

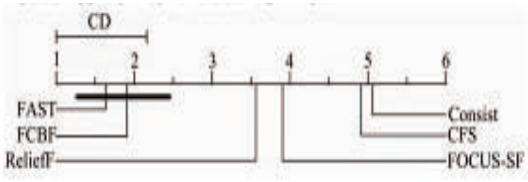


Figure 4 : Runtime comparison of all feature selection algorithms against each other with the Nemenyi test

The classification accuracy of RIPPER. From it we observe that

1. Compared with original data, the classification accuracy of RIPPER has been improved by the five feature selection algorithms FAST, FCBF, CFS, Consist, and FOCUS-SF by 7.64%, 4.51%, 4.08%, 5.48%, and 5.32%, respectively; and has been decreased by Relief by 2.04%. FAST ranks 1 with a margin of 2.16% to the second best accuracy 78.46% of Consist. The Win/Draw/Loss records show that FAST outperforms all other algorithms.
2. For image data, the classification accuracy of RIPPER has been improved by all the six feature selection algorithms FAST, FCBF, CFS, Relief, Consist, and FOCUS-SF by 12.35%, 8.23 %, 4.67%, 3.86%, 4.90%, and 4.90%, respectively. FAST ranks 1 with a margin of 4.13% to the second best accuracy 76.52% of FCBF.
3. For microarray data, the classification accuracy of RIPPER has been improved by all the six algorithms FAST, FCBF, CFS, Relief, Consist, and FOCUS-SF by 13.35%, 6.13%,

This means that at α= 0.1, there are evidences to reject the null hypotheses and the accuracies are different further differences exist in the six feature selection algorithms.

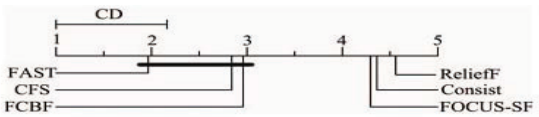


Figure 5 : Accuracy comparison of Naive Bayes with the six feature selection algorithms against each other with the Nemenyi test

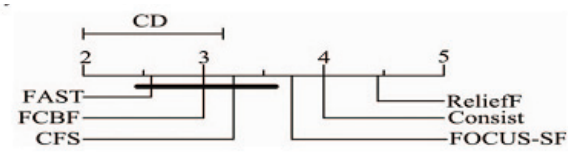


Figure 6 : Accuracy comparison of C4.5 with the six feature selection algorithms against each other with the Nemenyi test

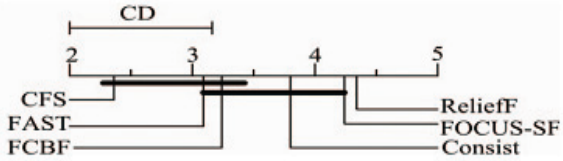


Figure 7 : Accuracy comparison of IB1 with the six feature selection algorithms against each other with the Nemenyi test

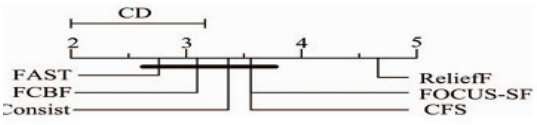


Figure 8 : Accuracy comparison of RIPPER with the six feature selection algorithms against each other with the Nemenyi test

From Fig. 5 we observe that the accuracy of Naïve Bayes with FAST is statistically better than those with Relief, Consist, and FOCUS-SF. But there is no consistent evidence to indicate statistical accuracy differences between Naive Bayes with FAST and with CFS, which also holds for Naive Bayes with FAST and with FCBF. From Fig. 6 we observe that the accuracy of C4.5 with FAST is statistically better than those with Relief, Con-sist, and FOCUS-SF. But there is no consistent evidence to indicate statistical accuracy differences between C4.5 with FAST and with FCBF, which also holds for C4.5 with FAST and with CFS. From Fig. 7 we observe that the accuracy of IB1 with FAST is statistically better than those with Relief. But there is no consistent evidence to indicate statistical accuracy differences between IB1 with FAST and with FCBF, Consist, and FOCUS-SF, respectively, which also holds for IB1 with FAST and with CFS. From Fig. 8 we observe that the accuracy of RIPPER with FAST is statistically better than those with Relief. But there is no consistent evidence to indicate statistical accuracy differences between RIPPER with FAST and with FCBF, CFS, Consist, and FOCUS-SF, respectively. For the purpose of exploring the relationship between feature selection algorithms and data types, i.e. which algorithms are more suitable for which types of data, we rank the six feature selection algorithms according to the classification accuracy of a given classifier on a specific type of data after the feature selection algorithms are performed. Then we summarize the ranks of the feature selection algorithms under the four different classifiers, and give the final ranks of the feature selection algorithms on different types of data. Table 8 shows the results. From Table 8 we observe that (i) for image data, CFS obtains the rank of 1, and FAST ranks 3; (ii) for microarray data, FAST ranks 1 and should be the undisputed first choice, and CFS is a good alternative; (iii) for text data, CFS obtains the rank of 1, and FAST and FCBF are alternatives; and (iv) for all data, FAST ranks 1 and should be the undisputed first choice, and FCBF, CFS are good alternatives.

VIII. Sensitivity Analysis

Like many other feature selection algorithms, our pro-posed FAST also requires a parameter that is the threshold of feature relevance. Different values might end with different classification results. In order to explore which parameter value results in the best classification accuracy for a specific classification problem with a given classifier, a 10 fold cross-validation strategy was employed to reveal how the classification accuracy is changing with value of the parameter.

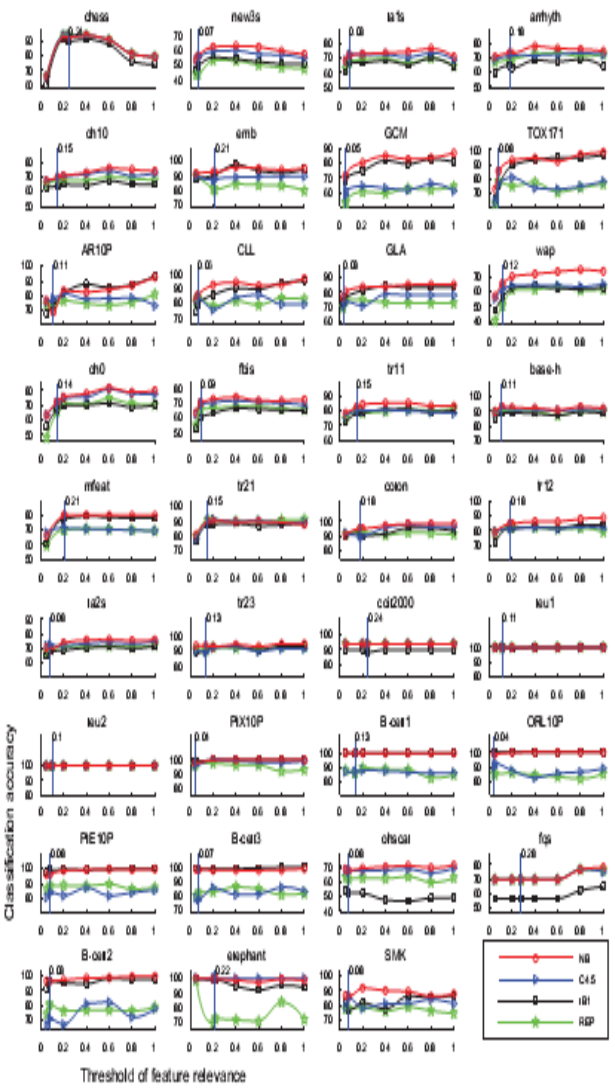


Figure 9 : Accuracies of the four classification algorithms with different θ values

The cross points of the vertical line with the horizontal axis represent the default values of the parameter θ recommended by FAST, and the cross points of the vertical line with the four curves are the

classification accuracies of the corresponding classifiers with the θ values. From it we observe that:

Classification accuracies; (ii) there is a θ value where the corresponding classification accuracy is the best; and (iii) the θ values, in which the best classification accuracies are obtained, are different for both the different data sets and the different classification algorithms. Therefore, an appropriate θ value is desired for a specific classification problem and a given classification algorithm. 2) In most cases, the default θ values recommended by FAST are not the optimal. Especially, in a few cases (e. g., data sets GCM, CLL-SUB-11, and TOX-171), the corresponding classification accuracies are very small. This means the results presented in Section 4.4.3 are not the best, and the performance could be better. 3) For each of the four classification algorithms, al-though the θ values where the best classification accuracies are obtained are different for different data sets.The value of 0.2 is commonly accepted because the corresponding classification accuracies are among the best or nearly the best ones. When determining the value of θ, besides classification accuracy.

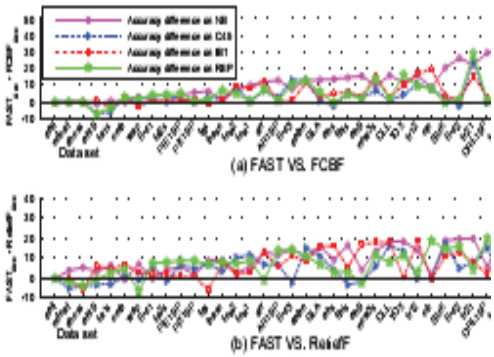


Figure 10 : Accuracy differences between FAST and the comparing algorithms

Just like the default θ values used for FAST in the experiments are often not the optimal in terms of classification accuracy, the default threshold values used for FCBF and Relief (CFS, Consist, and FOCUS-SF do not require any input parameter) could be so. In order to explore whether or not FAST still outperforms when optimal threshold values are used for the comparing algorithms, 10-fold cross-validation methods were firstly used to determine the optimal threshold values and then were employed to conduct classification for each of the four classification methods with the different feature subset selection algorithms upon the 35 data sets. The results reveal that FAST still outperforms both FCBF and Relief for all the four classification methods, Fig. 10 shows the full details. signed ranks tests with $\alpha = 0.05$ were performed to confirm the results as advised by Demsar. All the p values are smaller than 0.05, this indicates that the FAST is significantly better than both FCBF and Relief.

IX. Conclusion

In this paper, we have presented a novel clustering-based feature subset selection algorithm for high dimensional data. The algorithm involves (i) removing irrelevant features, (ii) constructing a minimum spanning tree from relative ones, and (iii) partitioning the MST and selecting representative features. In the proposed algorithm, a cluster consists of features. Each cluster is treated as a single feature and thus dimensionality is drastically reduced. We have compared the performance of the proposed algorithm with those of the five well-known feature selection algorithms FCBF, Relief, CFS, Consist, and FOCUS-SF on the 35 publicly available image, microar-ray, and text data from the four different aspects of the proportion of selected features, runtime, classification accuracy of a given classifier, and the Win/Draw/Loss record. Generally, the proposed algorithm obtained the best proportion of selected features, the best runtime, and the best classification accuracy for Naive Bayes, C4.5, and RIPPER, and the second best classification accuracy for IB1. The Win/Draw/Loss records confirmed the conclusions. We also found that FAST obtains the rank of 1 for microarray data, the rank of 2 for text data, and the rank of 3 for image data in terms of classification accuracy of the four different types of classifiers, and CFS is a good alternative. At the same time, FCBF is a good alternative for image and text data. Moreover, Consist and FOCUS-SF are alternatives for text data. For the future work, we plan to explore different types of correlation measures, and study some formal properties of feature space.

References Références Referencias

1. Almuallim H. and Dietterich T.G., Algorithms for Identifying Relevant Features, In Proceedings of the 9th Canadian Conference on AI, pp 38-45, 1992.
2. Almuallim H. and Dietterich T.G., Learning boolean concepts in the presence of many irrelevant features, Artificial Intelligence, 69(1-2), pp 279-305, 1994.
3. Arauzo-Azofra A., Benitez J.M. and Castro J.L., A feature set measure based on relief, In Proceedings of the fifth international conference on Recent Advances in Soft Computing, pp 104-109, 2004.
4. Baker L.D. and McCallum A.K., Distributional clustering of words for text classification, In Proceedings of the 21st Annual international ACM SIGIR Conference on Research and Development in information Retrieval, pp 96-103, 1998.
5. Battiti R., Using mutual information for selecting features in supervised neural net learning, IEEE Transactions on Neural Networks, 5(4), pp 537-550, 1994.

6. Bell D.A. and Wang, H., A formalism for relevance and its application in feature subset selection, Machine Learning, 41(2), pp 175-195, 2000.
7. Biesiada J. and Duch W., Features election for high-dimensionaldatała Pear-son redundancy based filter, Advancesin Soft Computing, 45, pp 242C249, 008.
8. Butterworth R., Piatetsky-Shapiro G. and Simovici D.A., On Feature Se-lection through Clustering, In Proceedings of the Fifth IEEE international Conference on Data Mining, pp 581-584, 2005.
9. Cardie, C., Using decision trees to improve case-based learning, In Pro-ceedings of Tenth International Conference on Machine Learning, pp 25-32, 1993.
10. Chanda P., Cho Y., Zhang A. and Ramanathan M., Mining of Attribute Interactions Using Information Theoretic Metrics, In Proceedings of IEEE international Conference on Data Mining Workshops, pp 350-355, 2009.
11. Chikhi S. and Benhammada S., ReliefMSS: a variation on a feature ranking Relief algorithm. Int. J. Bus. Intell. Data Min. 4(3/4), pp 375-390, 2009.
12. Cohen W., Fast Effective Rule Induction, In Proc. 12th international Conf. Machine Learning (ICML'95), pp 115-123, 1995.
13. Dash M. and Liu H., Feature Selection for Classification, Intelligent Data Analysis, 1(3), pp 131-156, 1997.
14. Dash M., Liu H. and Motoda H., Consistency based feature Selection, In Proceedings of the Fourth Pacific Asia Conference on Knowledge Discovery and Data Mining, pp 98-109, 2000.
15. Das S., Filters, wrappers and a boosting-based hybrid for feature Selection, In Proceedings of the Eighteenth International Conference on MachineLearning, pp 74-81, 2001.
16. Dash M. and Liu H., Consistency-based search in feature selection. Artificial Intelligence, 151(1-2), pp 155-176, 2003.
17. Demsar J., Statistical comparison of classifiers over multiple data sets, J. Mach. Learn. Res., 7, pp 1-30, 2006.
18. Dhillon I.S., Mallela S. and Kumar R., A divisive information theoretic feature clustering algorithm for text classification, J. Mach. Learn. Res., 3, pp 1265-1287, 2003.
19. Dougherty, E. R., Small sample issues for microarray-based classification. Comparative and Functional Genomics, 2(1), pp 28-34, 2001.
20. Fayyad U. and Irani K., Multi-interval discretization of continuous-valued attributes for classification learning, In Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence, pp 1022-1027, 1993.

Global Journals Inc. (US) Guidelines Handbook 2013

www.GlobalJournals.org

FELLOW OF ASSOCIATION OF RESEARCH SOCIETY IN COMPUTING (FARSC)

- 'FARSC' title will be awarded to the person after approval of Editor-in-Chief and Editorial Board. The title 'FARSC" can be added to name in the following manner. eg. **Dr. John E. Hall, Ph.D., FARSC or William Walldroff Ph. D., M.S., FARSC**
- Being FARSC is a respectful honor. It authenticates your research activities. After becoming FARSC, you can use 'FARSC' title as you use your degree in suffix of your name. This will definitely will enhance and add up your name. You can use it on your Career Counseling Materials/CV/Resume/Visiting Card/Name Plate etc.
- 60% Discount will be provided to FARSC members for publishing research papers in Global Journals Inc., if our Editorial Board and Peer Reviewers accept the paper. For the life time, if you are author/co-author of any paper bill sent to you will automatically be discounted one by 60%
- FARSC will be given a renowned, secure, free professional email address with 100 GB of space eg.johnhall@globaljournals.org. You will be facilitated with Webmail, Spam Assassin, Email Forwarders, Auto-Responders, Email Delivery Route tracing, etc.
- FARSC member is eligible to become paid peer reviewer at Global Journals Inc. to earn up to 15% of realized author charges taken from author of respective paper. After reviewing 5 or more papers you can request to transfer the amount to your bank account or to your PayPal account.
- Eg. If we had taken 420 USD from author, we can send 63 USD to your account.
- FARSC member can apply for free approval, grading and certification of some of their Educational and Institutional Degrees from Global Journals Inc. (US) and Open Association of Research, Society U.S.A.
- After you are FARSC. You can send us scanned copy of all of your documents. We will verify, grade and certify them within a month. It will be based on your academic records, quality of research papers published by you, and 50 more criteria. This is beneficial for your job interviews as recruiting organization need not just rely on you for authenticity and your unknown qualities, you would have authentic ranks of all of your documents. Our scale is unique worldwide.
- FARSC member can proceed to get benefits of free research podcasting in Global Research Radio with their research documents, slides and online movies.
- After your publication anywhere in the world, you can upload you research paper with your recorded voice or you can use our professional RJs to record your paper their voice. We can also stream your conference videos and display your slides online.
- FARSC will be eligible for free application of Standardization of their Researches by Open Scientific Standards. Standardization is next step and level after publishing in a journal. A team of research and professional will work with you to take your research to its next level, which is worldwide open standardization.

- FARSC is eligible to earn from their researches: While publishing his paper with Global Journals Inc. (US), FARSC can decide whether he/she would like to publish his/her research in closed manner. When readers will buy that individual research paper for reading, 80% of its earning by Global Journals Inc. (US) will be transferred to FARSC member's bank account after certain threshold balance. There is no time limit for collection. FARSC member can decide its price and we can help in decision.

MEMBER OF ASSOCIATION OF RESEARCH SOCIETY IN COMPUTING (MARSC)

- 'MARSC' title will be awarded to the person after approval of Editor-in-Chief and Editorial Board. The title 'MARSC" can be added to name in the following manner. eg. Dr. John E. Hall, Ph.D., MARSC or William Walldroff Ph. D., M.S., MARSC
- Being MARSC is a respectful honor. It authenticates your research activities. After becoming MARSC, you can use 'MARSC' title as you use your degree in suffix of your name. This will definitely will enhance and add up your name. You can use it on your Career Counseling Materials/CV/Resume/Visiting Card/Name Plate etc.
- 40% Discount will be provided to MARSC members for publishing research papers in Global Journals Inc., if our Editorial Board and Peer Reviewers accept the paper. For the life time, if you are author/co-author of any paper bill sent to you will automatically be discounted one by 60%
- MARSC will be given a renowned, secure, free professional email address with 30 GB of space eg.johnhall@globaljournals.org. You will be facilitated with Webmail, Spam Assassin, Email Forwarders, Auto-Responders, Email Delivery Route tracing, etc.
- MARSC member is eligible to become paid peer reviewer at Global Journals Inc. to earn up to 10% of realized author charges taken from author of respective paper. After reviewing 5 or more papers you can request to transfer the amount to your bank account or to your PayPal account.
- MARSC member can apply for free approval, grading and certification of some of their Educational and Institutional Degrees from Global Journals Inc. (US) and Open Association of Research, Society U.S.A.
- MARSC is eligible to earn from their researches: While publishing his paper with Global Journals Inc. (US), MARSC can decide whether he/she would like to publish his/her research in closed manner. When readers will buy that individual research paper for reading, 40% of its earning by Global Journals Inc. (US) will be transferred to MARSC member's bank account after certain threshold balance. There is no time limit for collection. MARSC member can decide its price and we can help in decision.

AUXILIARY MEMBERSHIPS

ANNUAL MEMBER

- Annual Member will be authorized to receive e-Journal GJCST for one year (subscription for one year).
- The member will be allotted free 1 GB Web-space along with subDomain to contribute and participate in our activities.
- A professional email address will be allotted free 500 MB email space.

PAPER PUBLICATION

- The members can publish paper once. The paper will be sent to two-peer reviewer. The paper will be published after the acceptance of peer reviewers and Editorial Board.

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

ISSN 9754350