Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 12 Issue 15 Version 1.0 Year 2012
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Encapsulation of Soft Computing Approaches within Itemset Mining – A Survey

By Jyothi Pillai & O.P.Vyas

Bhilai Institute of Technology, Durg, Chhattisgarh, India

Abstract - Data Mining discovers patterns and trends by extracting knowledge from large databases. Soft Computing techniques such as fuzzy logic, neural networks, genetic algorithms, rough sets, etc. aims to reveal the tolerance for imprecision and uncertainty for achieving tractability, robustness and low-cost solutions. Fuzzy Logic and Rough sets are suitable for handling different types of uncertainty. Neural networks provide good learning and generalization. Genetic algorithms provide efficient search algorithms for selecting a model, from mixed media data. Data mining refers to information extraction while soft computing is used for information processing. For effective knowledge discovery from large databases, both Soft Computing and Data Mining can be merged. Association rule mining (ARM) and Itemset mining focus on finding most frequent item sets and corresponding association rules, extracting rare itemsets including temporal and fuzzy concepts in discovered patterns. This survey paper explores the usage of soft computing approaches in itemset utility mining.

Keywords : *data mining, soft computing, itemset mining, fuzzy logic, neural networks, genetic algorithm.*

GJCST-C Classification : *H.2.8*

Strictly as per the compliance and regulations of:

Encapsulation of Soft Computing Approaches within Itemset Mining – A Survey

Jyothi Pillai [α] & O.P.Vyas [σ]

Abstract - Data Mining discovers patterns and trends by extracting knowledge from large databases. Soft Computing techniques such as fuzzy logic, neural networks, genetic algorithms, rough sets, etc. aims to reveal the tolerance for imprecision and uncertainty for achieving tractability, robustness and low-cost solutions. Fuzzy Logic and Rough sets are suitable for handling different types of uncertainty. Neural networks provide good learning and generalization. Genetic algorithms provide efficient search algorithms for selecting a model, from mixed media data. Data mining refers to information extraction while soft computing is used for information processing. For effective knowledge discovery from large databases, both Soft Computing and Data Mining can be merged.

Association rule mining (ARM) and Itemset mining focus on finding most frequent item sets and corresponding association rules, extracting rare itemsets including temporal and fuzzy concepts in discovered patterns.

This survey paper explores the usage of soft computing approaches in itemset utility mining.

Keywords : data mining, soft computing, itemset mining, fuzzy logic, neural networks, genetic algorithm.

I. Introduction

Association rule mining (ARM) is one of the most important areas of data mining research which is used for the discovery of frequent itemsets and their corresponding association rules[RT 2011]. An emerging topic in the field of data mining is Utility Mining which is an extension of Frequent Itemset mining. The main objective of Utility Mining is to identify the itemsets with highest utilities, by considering profit, quantity, cost or other user preferences. In many real-life applications, high-utility itemsets consist of rare items also[JV2010]. Soft computing aims to uncover the tolerance for vagueness, partial truth and approximation to achieve tractability, robustness and solutions with low cost. Soft computing methodologies consisting of fuzzy sets, neural networks, genetic algorithms, and rough sets are combined with data mining for knowledge discovery in large databases [SSP 2002]. The resultant technique is a more intelligent system which provides a human-interpretable, low cost solution.

This paper presents a brief overview of exploration of soft computing approaches in itemset utility mining. Section 2 and Section 3 discuss theoretical definitions related to Data Mining, Itemset Utility Mining and Temporal Mining. Section 4 discusses the state of art of soft computing tools. Section 5 presents usage of different soft computing methods in data mining, itemset mining and temporal mining. Section 6 presents conclusion and future work.

II. Data Mining

Data mining is the technique of automatic finding of hidden patterns and information elicitation from huge volume of raw data stored in data bases, data warehouses and other data repositories for making better business decisions, finding sales trends, in developing smarter marketing campaigns, and to predict customer loyalty.

Two categories of Data mining tasks are; Descriptive Mining and Predictive Mining. The Descriptive Mining techniques include Clustering, Association Rule Discovery, and Sequential Pattern Discovery, which is used to find human-interpretable patterns that describe the data in the form of clusters, itemsets, association rules and sequential patterns. The Predictive Mining techniques such as Classification, Regression, Deviation Detection, are used to classify objects or to predict future values of other variables.

One of the most important research areas in the field of Data mining is ARM. Association rules are used to identify relationships among a set of items in a transactional dataset. Apriori algorithm, given by Agrawal, Imielinski and Swami in 1993, is the first association rule mining algorithm, which influenced not only the association rule mining community, but also has impact on other data mining fields. Apriori and all its variants like Partition, Pincer-Search, Incremental, Border algorithm etc. take too much computer time to compute all the frequent item sets and usually consider only the frequency of items in itemsets.

III. Itemset Mining

a) Frequent Itemset Mining

Frequent itemsets are itemsets that occur frequently in a transaction data set. The goal of Frequent Itemset Mining is to identify all the frequent itemsets in a transaction dataset. A frequent itemset is the itemset having frequency support greater a minimum user specified threshold [JV2011].

Author α : Associate Professor, Bhilai Institute of Technology, Durg, Chhattisgarh, India. E-mail : jyothi_rpillai@rediffmail.com

Author σ : Professor, Indian Institute of Information Technology Allahabad, U.P., India. E-mail : dropvyas@gmail.com

Association Rule Mining (ARM)- The problem of mining association rules was first introduced in [RTA1993]. ARM is a popular technique for finding co-occurrences, correlations, frequent-patterns, associations among items in a set of transactions or a database. Rules with confidence and support above user-defined thresholds (minconf and minsup) were found. ARM process can be divided into two steps. The first step involves finding all frequent itemsets in databases. Next, association rules are generated from these frequent itemsets.

b) Rare Itemset Mining

The basic Bottleneck of ARM is Rare Item Problem. In many applications, some items appear very frequently in the data, while others rarely appear. In many practical situations such as security, business strategies, pattern extraction from web page access logs, biology, medicine and super market shelf management, the rare combinations of items in the itemset with high utilities provide very useful insights to the user [JV2010].

c) Utility Mining

Identification of the itemsets with high utilities is called as Utility Mining. The frequency of itemset is not sufficient to reflect the actual utility of an itemset [JV2011]. For example, the sales executives are not interested in frequent itemsets which do not yield significant profit. Mining of high utility itemsets is one of the most challenging recent data mining tasks. The utility value of an item depends on its evaluation e.g. if cola has support 30 and profit of 3%, cake may have support 10 but with a profit of 30%. This indicates that the utility of cake is higher than cola.

Utility mining model was proposed in [YHG2006] to define the utility of itemset. Utility is measured by analyzing how useful or profitable an itemset X is to user. The utility of an itemset X, u(X) is the sum of the utilities of itemset X in all the transactions containing X. An itemset X is called a high utility itemset if and only if u(X) >= min_utility, where min_utility is a user-defined minimum utility threshold [YHG2006]. For example, a computer system may be more profitable than a telephone in terms of profit. The main objective of high-utility itemset mining is to find all those itemsets having utility greater or equal to user-defined minimum utility threshold.

IV. Soft Computing

a) Importance of Soft Computing

Soft computing is tolerant of vagueness, imprecision, uncertainty, incomplete truth and approximation. The main components of Soft Computing are: fuzzy logic (FL), neural networks (NN), probabilistic reasoning (PR), genetic algorithms (GA), and chaos theory (ChT), which are summarized:-

i. *Fuzzy Logic* [RV 2011] Lotfi Zadeh conceived the concept of FL. FL is used to deal with uncertain or vague data, considered as fuzzy sets. In FL procedure, attribute values are transformed to fuzzy values and corresponding fuzzy membership or truth values are calculated.
ii. *Neural Networks* [RV 2011] NN is a network of artificial neurons which are simple processing elements which process information with a connectionist approach to computation [RAA2001]. An important property of these networks is their inductive nature, which uses "learning by example" in problems solving.
iii. *Genetic Algorithms* [RV 2011] GA is a flexible, heuristic and inductive search technique based on the theory of natural selection. GA learning consists of following steps: An initial input is created which consists of randomly generated rules. Each rule is represented using a string of bits. The fitness of a rule is evaluated by its classification accuracy on the training samples set. This process of generating new populations based on previous populations of rules is repeated till each rule of a population satisfies a pre defined fitness threshold [RAA2001].
iv. *Rough sets* [RV 2011] RS theory proposed by Pawlak is generally used for classification evaluation of data bases and for discovering structural relationships within uncertain or noisy data.
v. *Probabilistic Reasoning* PR [RAA2001]. PR offers methods to assess the outcome of systems which are affected by probabilistic ambiguity. The probabilistic mechanism provides a precise framework for illustration of a probabilistic knowledge, modeling of random phenomena and to analyze them.
vi. *Chaos Theory* [RAA2001]. A chaotic system is a deterministic system that exhibits random behavior. ChT deals with the non-linear dynamical systems that exhibit extreme sensitivity to initial conditions.

b) Need of Soft Computing in Data Mining

By incorporation of Soft Computing, there is a significant increase in effectiveness of artificial intelligence systems. All techniques have their own uniqueness based upon which they can be properly used in data mining process.

i. *Fuzzy Logic in Data mining*

The role of fuzzy sets based on different data mining functions are categorized below [SSP 2002]-

a. *Classification*

FL systems are used in several areas for classification, such as in business, health care and finance.

b. *Clustering*

Fuzzy clustering algorithms have been developed to mine telecommunications, customer and

prospect databases for gaining residential and business customer market share [SSP 2002].

c. *Association Rules*

Because of the affinity of FL with human knowledge representation, FL is considered as a key component of data mining systems.

d. *Functional Dependencies*

FL is also used for analyzing deductions based on functional dependencies (FDs) among variables, in database relations. Fuzzy relational databases generalize their classical and imprecise counterparts by supporting fuzzy information storage and retrieval [SSP 2002].

e. *Data Summarization*

FL techniques are used for data summarization.

ii. *Neural Networks in Data mining*

NNs can be efficiently encapsulated with data mining methods to increase the efficiency of the output of different data mining techniques. ANNs act as feasible computational models for different problems such as pattern classification, speech recognition, curve fitting, approximation capability, image data compression, associative memory, and modeling and control of non-linear unknown systems and are successfully utilized in various areas, such as science, engineering, medical, business, banking, telecommunication[RAA2001].

iii. *Genetic Algorithm in Data mining*

GA processing objects operate directly to set, queue, matrices, charts, and other structure. GA adopts probability rules to lead search direction. Genetic programming concepts have been used for developing Knowledge discovery systems. For better attribute interaction, GAs can be used.

iv. *Rough Sets in Data mining*

The main aim of RS is stimulation of approximation of concepts. Mathematical tools are offered by RS to extract hidden patterns in data and therefore are used in data mining. In data mining, RS can be used as a framework where precise data is not necessary and in the areas where approximate data is of great help. In data processing RST can be used for computing lower and upper approximation [RV 2011].

v. *Probabilistic Reasoning in Data mining*

Statistics or Probabilistic Theory forms a basis for good management and also plays a very important role in the data mining methods [RAA2001].

vi. *Chaos Theory in Data mining*

The predictability can be done using chaotic analysis and also prediction strategies of system's behavior can be formulated. ChT deals efficiently with noisy nonlinear systems. Chaotic computing gives a tool to determine a new perspective of nonlinear data analysis [RAA2001].

V. Literature Survey

a) Application of Soft Computing in Data Mining

By combining the advantages of both Data mining and soft computing paradigms, the techniques can be used for discovering knowledge in databases. In this section, a literature survey of integration of various soft computing methodologies and data mining is presented.

i. *Fuzzy Logic*

In retrieval of information, the main complexity is identifying relevant information, i.e. the nearest or the most similar according to user's need or expectation. This problem motivated to use fuzzy sets in knowledge representation thus enabling the user to express his prospect in a language not far from natural. Another reason is the approximate matching between the user's requirements and existing values in the database, on the basis of similarities and degrees of satisfiability.

The thesis report of Jianxiong Luo [JL1999] explores integrating FL with two data mining methods (association rules and frequency episodes) for intrusion detection. In intrusion detection, many quantitative features are involved and also security is fuzzy.

Au and Chan [WK1999] use an adjusted difference between experimental and probable frequency counts of attributes for finding out fuzzy association rules in relational datasets. The algorithm discovers both positive and negative rules and is able to cope with fuzzy class boundaries and missing values.

The authors in [BDLMR2007] focus on the applications of fuzzy techniques for information retrieval and data mining in real-world situations such as medical, educational, chemical and multimedia have been illustrated.

In real-time systems, for example in e-banking, assessing and determining any phishing websites is a complex and dynamic problem because of ambiguities involved. Aburrous et al present a intelligent, flexible and efficient system approach to deal with 'fuzziness' in the e-banking phishing website using fuzzy data mining techniques [AHDT2010].

In [KMA2012], the authors present an overview of the applications of fuzzy decision tree in heterogeneous fields. It is used dynamically in various fields such as intrusion detection, querying processes, cognitive process analysis (Human Computer Interaction), biometrics authentication, stock-market, parallel processing support, information retrieval and also in data mining.

ii. *Neural Network*

The paper [HRH1996] presents a method to find out symbolic classification rules using NNs.

Hongjun Lu et al propose an approach which can extract concise symbolic rules accurately using NN. The NN is trained for achieving required accuracy rate. Then through network pruning algorithm, repeated connections of the network are removed. The hidden layers of the network are analyzed and classification rules are generated and high quality rules are generated from the data sets.

In [X2008] the usage of NNs in data mining is researched in detail. NN can be considered as a parallel processing network which is formed by simulating the intuitive thinking of human. In data mining frequently used fuzzy NNs are fuzzy Back Propagation network, fuzzy perception model, fuzzy inference network, fuzzy clustering Kohonen network and fuzzy ART model.

By combining data mining and NNs, information is harvested from datasets by data warehousing firms [YA2009]. NNs can be used in all data mining tasks; generating association rules, classifications, clustering, prediction and forecasting. In data mining NNs act as a promising field for detecting and generating relationships among variables of large data sets.

NNs are motivated by brain functions, particularly pattern recognition and associative memory. The design of the NN architecture for the credit card detection system was based on unsupervised method, which was applied to the transactions data to generate four clusters of low, high, risky and high-risk clusters[F2011]. NN can be employed in banks to detect fraudulent usage of card more efficiently.

Anuj et al discuss that it is more expensive to connect a new customer than to maintain an existing loyal customer [AP2011]. The authors propose a NN based approach for predicting customer churn in cellular wireless services subscription and conclude a promising solution for customer churn management. The experimental results show that NN based method can predict customer churn with more than 92% accuracy.

Kamruzzaman et al propose a novel four-phase data mining algorithm using ANNs, referred as ESRNN (Extraction of Symbolic Rules from ANNs), for extracting symbolic rules [KJ2011]. The algorithm uses back propagation learning. Network architecture is defined and refined in the first phase and second phases. By using heuristic clustering algorithm, the nodes in hidden layers are discretized in third phase. Then symbolic rules are extracted from frequent patterns using extraction algorithm.

Mohammad Iquebal Akhter et al discusses in detail the function of ANN in preventing fraud in telecommunication services [MM2012]. A Fraud Detection System using ANN gathers historical data which is preprocessed and is used for training the NN for building a model which incorporates frequent fraud patterns. Finally, the model is applied to new business to learn new fraud patterns as the types of fraud evolved.

Madhusmita Swain et al introduced NNs for simplifying classification problem, IRIS plant classification [MSSA2012]. The problem identifies IRIS plant species on basis of plant attribute measurements. The authors used back propagation learning algorithm to train Multilayer feed- forward networks for identification of IRIS plants based on measurements such as length and width of sepal and length and width of petal. The authors conclude that Multi Layer Feed Forward NN (MLFF) is faster in terms of learning and is more accurate.

iii. *Genetic Algorithm*

A family of computational models which are inspired by evolution are GAs [E2011]. GA implementation begins with a population of random chromosomes. To create next generation of chromosomes from current population, GA uses three main types of rules:

1. The individuals called parents are selected through Selection rules, which contribute to next generation population.
2. Two parents are combined using Crossover rules to form next generation children.
3. Random changes are applied to individual parents using Mutation rules for forming children.

Ramesh Kumar et al presented a novel algorithm for rule prioritizing, which are generated by apriori algorithm through GA [RI2011].

E.P. Ephzibah proposed a new way to improve the performance of a model by combining GAs and FL, for feature selection and classification [E2011]. The proposed automated pattern classification system identifies and selects a subset of pattern from a larger set of features using fuzzy rule-based classification system. By the application of FL, the system's performance improved for diagnosing diabetes in patients.

Roohollah Etemadi et al propose a GA approach based on k-means clustering algorithm which can select cluster centers in a better manner [R2012]. All data objects are firstly clustered through k-means algorithm. Secondly, for each data object a pattern is generated by considering the generated clusters. On comparing with other related algorithms, the authors state that the proposed algorithm is more efficient than k-means algorithm and other algorithms.

Basheer M. Al-Maqaleh et al have explored the usage of GA, for finding predictive, complete and comprehensible classification rules from large database [BH2012]. The classification results of the proposed algorithm are compared with the performance of two algorithms; C4.5 and DTGA (DT and GA). DTGA has two rule inducing phases. In first phase, C4.5, a base classifier is used to generate rules from training data set,

then in next phase GA refines them for providing more accurate and high-performance prediction rules. According to authors, the proposed algorithm achieves better and accurate predictive results as compared to other two competent learners.

iv. *Rough Set*

RST deals with classificatory study of information systems. Z. Pawlak proposed this mathematical approach which is a powerful tool for dealing with vague data. Using RS method without deteriorating the quality of approximation, minimal attribute sets, and minimal length decision rules corresponding to lower or upper approximation can be extracted [W2012].

Prasanta et al proposed an approach based on RST which mine concise rules from inconsistent data [PRBB2011]. Firstly, lower and upper approximation is computed for each concept. Then a learning algorithm is adopted for building classification rules for each concept which satisfies classification accuracy. Test results show that the approach produced effective and minimal rules and offers more accurate results applied on several real life datasets.

In many fields such as inductive reasoning, classification, pattern recognition, cluster analysis, automatic learning algorithms, RST plays a significant role and is used in different domains like Medicine, Banking, Marketing and Engineering. In [S2011], A.S. Salama described some topological properties of RS which will help get rich results and discover hidden relations between data and also help in producing accurate programs.

Abdul Nassar proposes that using RST concept, clusters can be generated without any additional information for example probability distribution or fuzzy membership function [A2011]. By considering Lower approximation important rules of the target set can be generated. A reduct rule set of high importance can be generated by considering generated rules as attributes and a new decision table can be constructed.

Wen-Yau proposed a clustering technique which uses GA and RST [W2012]. After clustering, Apriori algorithm is used to discover association rules between products of same cluster and then marketing people can suggest related products to the targeting group. RS is used to generate rules and these rules are applied to various GA parts.

b) Application of Soft Computing in Itemset Mining

A literature survey of exploration of different soft computing approaches in itemset mining is discussed in this section.

i. *Fuzzy Logic*

Wai-Ho introduced a novel technique, called FARM (Fuzzy Association Rule Miner) to mine fuzzy association rules [WK1999] which uses linguistic terms for representing revealed regularities and exceptions, based on fuzzy set theory. The rules generated are called fuzzy association rules. FARM also discovers interesting associations between different quantitative values. One more advantage of FARM is that it can reveal both positive and negative association rules. A positive association rule indicates presence of another attribute value along with a certain attribute value whereas a negative association rule indicates absence of another attribute value along with a certain attribute value. Wai-Ho et al discuss that experimental results show FARM to be capable of discovering meaningful and useful fuzzy association rules.

Yi-Chung Hu et al proposed a learning algorithm, which acts as a knowledge acquisition tool for classification problems to efficiently generate fuzzy association rules [YRG2002]. In first phase, from training samples, large fuzzy grids are generated by fuzzy partitioning of each attribute and in second phase, for classification problems, fuzzy association rules by large fuzzy grids are generated. Experimental results on iris data indicate that the proposed algorithm accurately derive fuzzy association rules for classification problems.

One of the most essential areas of the application of fuzzy set theory is Fuzzy rule-based systems [CMM2004]. The advantages of using fuzzy systems for knowledge discovery processes are; information dealing with uncertain data, considering multi-variable relationships; human understandable results, easy information modification by an expert, easy adaptability to the given problem and high automated process. Fuzzy systems improve the interpretation and understandability of consumer models. In [CMM2004], Casillas et al presented a new approach for consumer behaviour modelling which is based on fuzzy association rules (FARs), centered on consumer attitude towards Internet and confidence in Internet shopping.

Sulaiman et al propose a new Fuzzy Healthy Association Rule Mining Algorithm (FHARM) which introduces new quality measures for generating more interesting and quality rules effectively and efficiently [SMCF2006]. Using FHARM, edible attributes are extracted from transactional input data and transformed to Required Daily Allowance (RDA) numeric values. The RDA values from database are then converted to fuzzy values. Analysis of normalized fuzzy transactional database is performed for getting nutritional information.

O. Dehzangi et al proposed a new approach to generate a set of rules for each class using data mining principles by reducing the number of generated rules [DZTF2007]. Using selection criteria, a precise number of rules for each class are selected and then a compact rule-base is constructed. The presented method improved the classification rate and deal effectively with noisy training examples.

Ashish Mangalampalli et al put forward a naive fuzzy ARM algorithm which performs faster and efficiently on very large datasets [AV2009]. Fuzzy ARM algorithm has following steps: Firstly, the crisp dataset is converted into a fuzzy dataset. Then fuzzy ARM algorithms are used which consider the fuzzy membership of an itemset in a given transaction along with its presence or absence.

Rajendran et al proposed a Novel Fuzzy Association Rule Mining (NFARM) method which deals with the detection of brain tumor in the CT scan brain images [RM2010]. In FARM, FL is used to transform numerical to fuzzy attributes. Discovered NFARM rules are tested on new test image to detect the brain tumor. The authors state that NFARM gives better performance and helps physicians in diagnosing the cancerous cells by providing better diagnosis system containing diagnosis keywords.

Radha et al proposed a classification method for generating fuzzy rules from training data [RR2010]. Using fuzzy C-Means algorithm, Quantitative attributes are divided into several fuzzy sets and accordingly membership values are generated. Then a supervised association rule algorithm is employed for discovering interesting FARs. Generated Fuzzy rules are used to build classification system. C4.5, Naïve Bayes, and ID3 classifiers are used for classification and accordingly fuzzy classified association rules are discovered. The authors discuss that the number of generated rules is reduced due to the usage of fuzzy linguistic values.

Maybin Muyeba et al presented a novel approach to mine weighted FARs effectively and address the issue of invalidation of downward closure property (DCP) in weighted ARM, where each item is assigned a weight according to its significance with respect to some user defined criteria [MSC2010].
Prakash et al present a qualitative fuzzy ARM (FARM) approach for mining FARs for the quantitative attributes [PP2011]. The authors evaluated the performance of qualitative FARM by experimenting with real data sets. Results prove that the qualitative approach discover more accurate association rules in less time with increased execution speed.

A novel approach is presented by Vedula Venkateswara Rao et al in [VES2012] for effectively mining frequent Item sets and generating association rules (ARs) based on fuzzy Apriori and weighted fuzzy Apriori. In weighted association rule mining (WARM), each item is assigned a weight with respect to its importance to some user defined criteria. Both binary data and fuzzy data are used in the proposed approach and Frequent Item Sets are generated. The Fuzzy Apriori algorithm (Apriori-Total) proposed in [VES2012] is founded on a tree structure called the T-tree to store frequent item set information.

K. Suriya Prabha et al proposed an approach that integrates FL and tree-based algorithm. The approach constructs a compact sub-tree for finding fuzzy frequent item [SL2012]. The authors conclude that the presented approach is quite efficient than other algorithms when evaluated on the basis of execution time, memory usages and search space for generating fuzzy frequent itemsets.

Ferdinando et al present a novel method for detecting association rules from datasets based on fuzzy transforms [FS2012]. AprioriGen algorithm is used for extracting fuzzy association rules which are represented in the form of linguistic expressions. A pre-processing phase is performed for determining optimal fuzzy partition of quantitative attributes domains.

Roohollah Etemadi states that one of the most well-known clustering methods is K-means algorithm which forms the base for other clustering approaches [R2012]. K-means and k-methods are heuristic partitioning algorithms where as Fuzzy k-means and Fuzzy k-methods are equivalent fuzzy type algorithms. In these partitioning methods, firstly k number of partitions is generated from data where each partition will contain at least one data. If crisp partitioning is performed, then a particular data will be present in only one cluster but if fuzzy partitioning is assumed then a particular data may be present in different clusters.

ii. *Neural Networks*

[VI2012] NNs have the capability to interpret meaning from complicated or vague data and hence can be used for extracting patterns and detecting trends which are difficult for humans or other computer techniques to notice

P. Sermswatsri et al proposes a more efficient method of frequent pattern mining by using Associative Classification method and NN [PS2006]. The proposed NN Associative Classification (NAC) method can be used to build more accurate and efficient classifiers. The authors conclude that the experimental results of NAC on datasets show improved accuracy rates.

Divya Bhatnagar et al propose an efficient technique for frequent itemsets mining in large databases using Optical NN Model [DNS2011]. The proposed technique removes the need for generating candidate sets for ARM to find frequent itemsets. The time complexity and space complexity of this technique is very low as optical NN can perform several computations simultaneously.

In [ASP2011], an efficient algorithm named Multi Level Feed Forward Mining (MLFM) is proposed by Amit Bhagat et al, for mining of multiple-level association rules efficiently from large transaction databases. The authors have used supervised NN in parallel for discovering frequent itemsets at each concept levels in a single scan of database. At each concept level MLFM reads items and divide them to various concept levels of hierarchy and passes it to the NN for generating frequent itemsets. Data at all the levels is given as input

by scanning the database only once, and thus produces fast output

NN Associative Classification system proposed by Prachitee B. Shekhawat builds a classifier with the help of Back propagation NN [PS2011].

iii. *Genetic Algorithm*

Xiaowei Yan et al designed an evolutionary mining strategy based on a GA called ARMGA model [XCS2007]. The authors discuss that, ARMGA model is efficient for global searching when search space is very large. Generally for rules mining, GAs are classified into two categories, according to encoding of rules in the population of chromosomes. In one encoding method called the Michigan Approach, each rule is encoded into an individual. In another method referred as the Pittsburgh Approach, the set of rules are encoded into a chromosome. ARMGA model is based on the Michigan strategy, where each association rule is encoded in a single chromosome.

In [ACC2007], Ansaf Salleb-Aouissi et al proposed QUANTMINER, a mining quantitative association rules system, which is based on GA that dynamically discovers "good" intervals in association rules.

Peter P. et al present a Pareto-based multiobjective evolutionary ARM method based on GAs [PV2008]. Predictive accuracy, comprehensibility and interestingness are used as different levels of interestingness of ARM problem

Xiaowei Yan et al designed a GA-based policy for discovering association rules without specifying actual minimum support [XCS2009].

Anandhavalli M. et al deal with a challenging ARM problem of finding optimized association rules [ASAG2009]. The authors by using GA find all the possible optimized rules from given data set. By using Apriori, frequent itemsets are generated.

Soumadip Ghosh et al propose a model in which the GA is applied on large data sets to find frequent itemsets [SSDP2010].

Vijaya Prakash et al proposed a technique to find all the frequent itemsets present in large data sets using GA [VGS2011]. The authors state that the generation of Frequent Itemset can be improved by using GA and also time complexity is reduced.

Rupesh Dewang et al propose "A new method for generating all positive and negative Association Rules" (NRGA) [RJ2011]. In First phase of NRGA, frequent itemsets and positive rules are generated using Apriori Algorithm. Then NRGA is used for generating all negative rules and finally GA is applied to optimize the generated rules.

Peter P. et al present a general ARM model for extracting useful information from very large databases [PVS2011]. The proposed model finds generalized association rules between items in a large database of transactions at any level of the taxonomy (is-a hierarchy) on the items.

The research paper [SJSK2012] presented by Sanat Jain et al is concerned with finding all positive and negative association rules from databases efficiently and optimization of generated rules is done using GA.

J. Malar Vizhi et al propose a new GA for generating high quality Association Rules. The authors used Michigan approach for representing the strong interesting association rules as chromosomes. Each chromosome is used to represent a separate strong association rule.

iv. *Rough Sets*

T. Y. Lin proposed a technique which applies RST to very large relational databases [T1996]. The proposed method integrates RST and the technique of extracting clean data subsets from noisy data banks for effectively mining soft rules.

Jiye Li et al introduced a rough set based process for ARM for selecting the most appropriate rules by using a rule importance measure [JN2005]. The authors introduced a rough set based model for providing an automatic and efficient way for ranking important rules in decision making applications.

X-Y. SHAO et al presents a methodology which integrates data mining tasks (like fuzzy clustering and ARM) and RST for discovering customer group-based configuration rules from the products purchased.

Tinghuai Ma et al provide a reduction algorithm for attribute reduction and pruning, using RST [TM2006]. The proposed reduction algorithm finds all reductions and is suitable for any uncertain knowledge reasoning.

In [EC2008], a new classification technique called 'Reduced MEPAR-miner Algorithm' is introduced by Emel Kizilkaya Aydogan et al, based on RST and Multi-Expression Programming for ARM, MEPAR-miner algorithm. In the preprocessing stage, the rough sets are used to reduce the feature space dimensionality and then to extract the classification rules, MEPAR-miner algorithms are used.

Anjana Pandey et al proposed RSMAR which uses Rough Set approach for Mining of Multidimensional Association Rules [AK2009]. The RSMAR algorithm consists of two steps. Firstly, the tables are combined to a single table for generating the rules which expresses the association between two or more domains belonging to different database tables and then on selected dimension, the mapping code is applied. In second step frequent itemsets are generated through equivalence classes and also the mapping code is transformed into real dimensions.

Jigyasa Bisaria et al have analyzed the sequential pattern mining problem through computational aspect and time constraint [JNP2009]. The authors have used RST to partition the sequential patterns search space in the proposed novel algorithm

which allows pre-visualization of patterns and also allows time constraint adjustment.

Anjana Pandey et al proposed an algorithm RS Model for Discovering Hybrid Association Rules [RSHAR] algorithm, for mining hybrid association rules using rough set approach [AP2009]. In RSHAR algorithm, the participant tables are combined into a general table for generating rules to express the relationship between two or more domains belonging to different database tables and then on selected dimension, the mapping code is applied. Then frequent itemsets are generated through equivalence classes and also the mapping code is transformed into real dimensions.

In [DHM2002], Daniel Delic et al emphasis on the comparison of association rules procedure and rough sets procedure. The proposed association rules method focus on the analysis of data bases containing boolean-valued attributes only. The authors conclude that there is a considerable reduction in computing time in the rough set algorithm.

A. Anitha et al proposed to combine upper approximation based rough set clustering and Apriori selective ARM for e-learning recommendation [AK2011]. In making e-learning recommendations, similar learning patterns are considered instead of all clicks stream sequences. The proposed algorithm resulted in dense clusters with less computational complexity and reduced number of extracted rules, which are highly relevant and meaningful.

VI. Conclusion

There has been substantial commercial interest as well as active research in data mining area for developing new and improved approaches for extracting information, relationships, and patterns from large datasets. Soft computing may be viewed as a foundation component for the emerging field of conceptual intelligence [RAA2001]. Hence Soft computing techniques can be encapsulated in Data mining for knowledge discovery in large databases. This paper presents a brief overview of various soft computing approaches used in itemset mining.

In future we will incorporate soft computing methodologies and itemset mining for mining high utility itemsets.

References Références Referencias

1. [A2011]Abdul Nassar . A.A, **Application of Rough Sets in Data Mining,** A Project Report Submitted in partial fulfillment of the requirements for the award of the degree of Master of Technology in Computer Science and Engineering, 2011.
2. [AHDT2010] M. Aburrous, M.A. Hossain, K. Dahal, F. Thabtah, **Intelligent phishing detection system for e-banking using fuzzy data mining,** Expert Systems with Applications 37 (2010), Elsevier, pp7913–7921.
3. [AK2011] A.Anitha, N.Krishnan, **A Dynamic Web Mining Framework for E-Learning Recommendations using Rough Sets and Association Rule Mining,** International Journal of Computer Applications (0975 – 8887), Volume 12– No.11, January 2011, pp 36-41.
4. [AP2009] Anjana Pandey, K.R.Pardasani, **Rough Set Model for Discovering Hybrid Association Rules,** IJCSIS June 2009 Issue, Vol. 2.
5. [AP2011] Anuj Sharma, Prabin Kumar Panigrahi, **A Neural Network based Approach for Predicting Customer Churn in Cellular Network Services,** International Journal of Computer Applications (0975 – 8887), Volume 27– No.11, August 2011, pp 26-31.
6. [ASAG2009] Anandhavalli M., Suraj Kumar Sudhanshu, Ayush Kumar, Ghose M.K., **Optimized association rule mining using genetic algorithm,** Bioinfo Publications, Advances in Information Mining, ISSN: 0975–3265, Volume 1, Issue 2, 2009, pp 01-04.
7. [ASP2011] Amit Bhagat, Sanjay Sharma, K.R. Pardasani, **Ontological Frequent Patterns Mining by potential use of Neural Network,** International Journal of Computer Applications (0975–8887), Volume 36- No. 10, December 2011, pp 44-53.
8. [AV2009] Ashish Mangalampalli, Vikram Pudi, **Fuzzy Association Rule Mining Algorithm for Fast and Efficient Performance on Very Large Datasets,** Fuzzy Systems, FUZZ-IEEE 2009. IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Jeju Island, Korea, 20-24 Aug., 2009, pp 1163 – 1168.
9. [BDLMR2007] B. Bouchon-Meunier, M. Detyniecki, M.-J. Lesot, C. Marsala, M. Rifqi, **Real world fuzzy logic applications in data mining and information retrieval**. Springer, pp 219-247, 2007.
10. [BH2012] Basheer M. Al-Maqaleh , Hamid Shahbazkia, **A Genetic Algorithm for Discovering Classification Rules in Data Mining,** International Journal of Computer Applications (0975 – 8887), Volume 41– No.18, March 2012, pp 40 -44.
11. [CMM2004] J. Casillas, F.J. Martínez-López, F.J. Martínez, **Fuzzy Association Rules For Estimating Consumer Behaviour Models And Their Application To Explaining Trust In Internet Shopping,** Fuzzy Economic Review, Volume: 9, Nov. 2004, pp 3-26.
12. [DHM2002] Daniel Delic, Hans-J. Lenz, Mattis Neiling, **Rough Sets and Association Rules - which is efficient?,** 14th Conference on Computational Statistics (CompStat2002) Berlin, HU, August 24-28, 2002, pp 527- 532.
13. [DNS2011] Divya Bhatnagar, Neeru Adlakha, A. S. Saxena, **Mining Frequent Itemsets without Candidate Generation using Optical Neural Network**, IJCA Special Issue on "Artificial

Intelligence Techniques - Novel Approaches & Practical Applications", AIT, 2011, pp 14-18.

14. [DZTF2007] O. Dehzangi, M. J. Zolghadri, S. Taheri, S.M. Fakhrahmad, **Efficient fuzzy rule generation: A new approach using data mining principles and rule weighting,** Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007), August 24-August 27, ISBN: 0-7695-2874-0, Vol. 2, pp 134 – 139.
15. [E2011] E.P.Ephzibah, **Cost Effective Approach On Feature Selection Using Genetic Algorithms And Fuzzy Logic For Diabetes Diagnosis**, International Journal on Soft Computing (IJSC), Vol.2, No.1, February 2011, pp 1-10.
16. [EC2008] Emel Kizilkaya Aydogan , Cevriye Gencer, **Mining classification rules with Reduced MEPAR-miner Algorithm,** Applied Mathematics and Computation 195 (2008) pp 786–798, www.elsevier.com/locate/amc
17. [F2011] Francisca Nonyelum Ogwueleka**, Data Mining Application in Credit Card Fraud Detection System** Journal Of Engineering Science And Technology, Vol. 6, No. 3 (2011) 311 - 322.
18. [FS2012] Ferdinando Di Martino, Salvatore Sessa, **Detection of Fuzzy Association Rules by Fuzzy Transforms,** Advances in Fuzzy systems, 2012.
19. [HRH1996] Hongjun Lu, Rudy Setiono, Huan Liu, **Effective Data Mining Using Neural Networks,** IEEE Transactions On Knowledge And Data Engineering, VOL. 8, NO. 6, DECEMBER 1996, pp 957-961.
20. [JL1999] Jianxiong Luo, **Integrating Fuzzy Logic With Data Mining Methods For Intrusion Detection,** A Thesis Submitted to the Faculty of Mississippi State University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer Science in the Department of Computer Science Mississippi State, Mississippi, August-1999.
21. [JN2005] Jiye Li, Nick Cercone, **A Rough Set Based Model to Rank the Importance of Association Rules,** In Proceedings of 10th International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, RSFDGrC 2005, LNAI Volume 3642, pp 109-118.
22. [JNP2009] Jigyasa Bisaria, Namita Shrivastava, K.R. Pardasani, **A Rough Sets Partitioning Model for Mining Sequential Patterns with Time Constraint,** International Journal of Computer Science and Information Security(IJCSIS),Vol. 2, No. 1, 2009.
23. [JV2010] Jyothi Pillai, O.P. Vyas, **Overview of Itemset Utility Mining and its Applications,** International Journal of Computer Applications (0975 – 8887), Volume 5– No.11, August 2010.
24. [KJ2011] S. M. Kamruzzaman, A. M. Jehad Sarkar, **A New Data Mining Scheme Using Artificial Neural Networks,** Sensors 2011, 11, doi: 10.3390/s110504622, ISSN 1424-8220, www.mdpi.com/journal/sensors, pp 4622-4647.
25. [KMA2012] Kavita Sachdeva, Madasu Hanmandlu, Amioy Kumar, **Real Life Applications of Fuzzy Decision Tree,** Intl Journal of Computer Applications (0975–8887), Volume 42– No.10, March 2012,pp 24-28
26. [MB2012] J.Malar Vizhi, Dr. T.Bhuvaneswari, **Data Quality Measurement With Threshold Using Genetic Algorithm,** International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com, Vol. 2, Issue4, July-August 2012, pp 1197-1203.
27. [MM2012] Mohammad Iquebal Akhter, Mohammad Gulam Ahamad, **Detecting Telecommunication Fraud using Neural Networks through Data Mining**, International Journal of Scientific & Engineering Research IJSER, Volume 3, Issue 3, March-2012 1 ISSN 2229-5518, http://www.ijser.org, pp 1-5.
28. [MSC2010] Maybin Muyeba, M. Sulaiman Khan, Frans Coenen, **Effective Mining of Weighted Fuzzy Association Rules, 2010,** IGI Global, **pp 47-64,** DOI: 10.4018/978-1-60566-754-6.ch004**.**
29. [MSSA2012] Madhusmita Swain, Sanjit Kumar Dash, Sweta Dash, Ayeskanta Mohapatra, **An Approach For Iris Plant Classification Using Neural Network,** International Journal on Soft Computing (IJSC) Vol.3, No.1, February 2012, pp 79-89.
30. [PP2011] S.Prakash, R.M.S.Parvathi, **Qualitative Approach for Quantitative Association Rule Mining using Fuzzy Rule Set,** Journal of Computational Information Systems, http://www.Jofcis.com, 1553-9105, Binary Information Press, June2011,1879-1885.
31. [PRBB2011] Prasanta Gogoi, Ranjan Das, B Borah, D K Bhattacharyya, Efficient Rule Set Generation using Rough Set Theory for Classification of High Dimensional Data, International Journal of Smart Sensors and Ad Hoc Networks (IJSSAN) ISSN No. 2248-9738 (Print) Volume-1, Issue-2, 2011, pp 13-20.
32. [PS2006] P. Sermswatsri & C. Srisa-an, **A neural-networks associative classification method for association rule mining,** WIT Transactions on Information and Communication Technologies, Vol.37, pp 93-102, WIT Press, Southampton (2006).
33. [PS2011] Prachitee B. Shekhawat, Sheetal S. Dhande, **A Classification Technique using Associative Classification,** International Journal of Computer Applications (0975 – 8887), Volume 20– No.5, April 2011, pp 20-28.
34. [PV2008] Peter P. Wakabi-Waiswa, Venansius Baryamureeba, **Extraction of Interesting Association Rules Using Genetic Algorithms,** International Journal of Computing and ICT Research, Vol. 2, No. 1, pp. 26 – 33, June 2008, http:www.ijcir.org.
35. [PVS2011] Peter P. Wakabi-Waiswa, Venansius Baryamureeba, K. Sarukesi, **Generalized Association Rule Mining Using Genetic Algorithms,**

Computer Science, Proceedings of Seventh International conference in natural computation, IEEE 2011, 1116-1120, pp 59-69.

36. [R2012] Roohollah Etemadi, **A Novel Evolutionary Algorithm For Data Clustering In N Dimensional Space,** Indian Journal of Computer Science and Engineering (IJCSE), ISSN : 0976-5166, Vol. 2, No. 6, Dec 2011-Jan 2012, pp 902-908.
37. [RI2011] M. Ramesh Kumar, K. Iyakutti, **Application of Genetic algorithms for the prioritization of Association Rules,** IJCA Special Issue on "Artificial Intelligence Techniques - Novel Approaches & Practical Applications", AIT, 2011, pp 35-38.
38. [RJ2011] Rupesh Dewang, Jitendra Agarwal, **A New Method for Generating All Positive and Negative Association Rules,** International Journal on Computer Science and Engineering (IJCSE), ISSN: 0975-3397 Vol. 3, No. 4, Apr 2011, pp 1649- 1657.
39. [RM2010] P. Rajendran, M. Madheswaran, **Novel Fuzzy Association Rule Image Mining Algorithm for Medical Decision Support System**, International Journal of Computer Applications (0975 - 8887), Volume 1 – No. 20, 2010, pp 87-94.
40. [RR2010] R.Radha, S.P.Rajagopalan, **Generating Membership Values And Fuzzy Association Rules From Numerical Data,** (IJCSE) International Journal on Computer Science and Engineering, Vol. 02, No. 08, 2010, pp 2705-2715.
41. [RV2011] V. V. R. Raman, Veena Tewari, **Data-mining and Soft-computing: Basis for Technology-based Management,** International Conference on Technology and Business Management March 28-30, 2011, 1221-1226.
42. [S2011] A. S. Salama, **Some Topological Properties of Rough Sets with Tools for Data Mining**, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 2, May 2011, ISSN (Online): 1694-0814, www.IJCSI.org, pp 588-595.
43. [SCRPA2010] M. Sulaiman Khan, Frans Coenen, D. Reid, R. Patel, L. Archer, A sliding windows based dual support framework for discovering emerging trends from temporal data, Knowledge Based Systems, Volume 23, Number 4, May 2010, 316-322.
44. [SJSK2012] Sanat Jain, Swati Kabra, **Mining & Optimization of Association Rules Using Effective Algorithm,** International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, Volume 2, Issue 4, April 2012, pp 281-285, www.ijetae.com.
45. [SL2012] K. Suriya Prabha, R. Lawrance, **Mining Fuzzy Frequent itemset using Compact Frequent Pattern(CFP) tree Algorithm,** International Conference on Computing and Control Engineering (ICCCE 2012), 12-13 April, ISBN 978-1-4675-2248-9.
46. [SMCF2006] M. Sulaiman Khan, Maybin Muyeba, Christos Tjortjis, Frans Coenen, **An effective Fuzzy Healthy Association Rule Mining Algorithm (FHARM),** In Lecture Notes Computer Science, vol. 4224, pp.1014-1022, ISSN: 0302-9743, 2006.
47. [SSDP2010] Soumadip Ghosh, Sushanta Biswas, Debasree Sarkar, Partha Pratim Sarkar, **Mining Frequent Itemsets Using Genetic Algorithm**, International Journal of Artificial Intelligence & Applications (IJAIA), Vol.1, No.4, October 2010, pp 133-143.
48. [SSP 2002] Sushmita Mitra, Sankar K. Pal, Pabitra Mitra, **Data Mining in Soft Computing Framework: A Survey**, IEEE Transactions On Neural Networks, VOL. 13, NO. 1, JANUARY 2002, pp 3-14.
49. [T1996] T. Y. Lin, **Rough Set Theory in Very Large Databases,** Proceedings of Symposium on Modeling, Analysis and Simulation, IMACS Multiconference (Computational Engineering in Systems Applications, Lille, France July 9-12, 1996, Volume 2 of 2, 1095-1100.
50. [TM2006] Tinghuai Ma, Meili Tang, **A New Reduction Implementation Based on Concept,** Journal of Information and Computing Science, **ISSN** 1746-7659, England, UK, Vol. 1, No. 5, 2006, pp. 290-294.
51. [VGS2011] R. Vijaya Prakash, Govardhan, S. S.V.N. Sharma, **Mining Frequent Itemsets from Large Data Sets using Genetic Algorithms,** IJCA Special Issue on "Artificial Intelligence Techniques - Novel Approaches & Practical Applications", AIT, 2011, pp 38-43.
52. [VI2012] Vinodhini Katiyaar, Ina Kapoor Sharma **Use of Data Mining & Neural Network in Commercial Application** International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 3 (3) , 2012,ISSN 0975-9646, pp 4041 – 4049.
53. [W2012] Wen-Yau Liang, **A Hybrid Evolutionary Computing For Market Segmentation,** Business and Information 2012, Sapporo, July 3-5), pp F171-F177.
54. [X2008] Xianjun Ni, **Research of Data Mining Based on Neural Networks**,World Academy of Science, Engineering and Technology 39 2008.
55. [XCS2007] Xiaowei Yan, Chengqi Zhang, Shichao Zhang, **ARMGA: Identifying Interesting Association Rules With Genetic Algorithms**, Applied Artificial Intelligence, 19:677–689, 2007, ISSN: 0883-9514 print/1087-6545 online, pp 677-689.
56. [XCS2009] Xiaowei Yan , Chengqi Zhang , Shichao Zhang , **Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support,** Expert Systems with Applications 36 (2009) 3066–3076, www.sciencedirect.com.

57. [YRG2002] Yi-Chung Hu, Ruey-Shun Chen, Gwo-Hshiung Tzeng,**Mining fuzzy association rules for classification problems,** Computers & Industrial Engineering 43 (2002), Elsevier Science Ltd., pp 735–750.

This page is intentionally left blank