1. **Introduction**

Association rule mining concept has been applied to market domain and specific problem has been studied, the management of some aspects of a shopping mall, and an architecture that makes it possible to construct agents capable of adapting the association rules has been used. Data mining refers to extracting knowledge from large quantity of data. Interesting association can be discovered among a large set of data items by association rule mining. The finding of interesting relationship among large amount of business transaction records can help in many business decisions making process. Association rules mining is an important task in the field of data mining, and frequent item set mining is a key step of many algorithms for association rules mining. There had been lots of work done for mining of association rules. When the dataset are large, the rules generated may be very large, but some of them are not interesting to the users, so, it is common to set some parameters to reduce the numbers of rules generated, support and confidence are two common parameters. An association rule R is of the form A → B, where A, B are disjoint subsets of the attribute set I. The support for the rule R is the number of database records which contain A∪B (often expressed as a proportion of the total number of records).

Two properties support and conﬁdence provide the empirical basis for derivation of the inference expressed in the rule, and a measure of the interest in the rule. The support for a rule expresses the number of records within which the association may be observed, while the conﬁdence expresses this as a proportion of the instances of the antecedent of the rule. In practical investigations, it is usual to regard these rules as “interesting” only if the support and conﬁdence exceed some threshold values. Hence the problem may be formulated as a search for all association rules within the database for which the required support and conﬁdence levels are attained. Note that the conﬁdence in a rule can be determined immediately once the relevant support values for the rule and its antecedent are computed. Thus the problem essentially resolves to a search for all subsets of I for which the support exceeds the required threshold. Such subsets are referred to as “large”, “frequent” or “interesting” sets.

1. **Motivation**

The practical benefits of mining association rules and its wide area of application have lead to several proposals for fast mining of association rules. Those proposals, although contribute towards making the process more applicable in practical systems, still suffer from the problem of the huge amount of generated rules that are both confusing and most of the time not useful to the user. The research community focused on providing solutions for the two problems separately. Association mining algorithm focused on maintain the discovered rules that have the same support constraints, namely those which qualify for the same setting of minimum support as the original database before updates.

* 1. **Objectives**

This dissertation work tries to achieve the following objectives.

* To find multiple-level frequent item sets for mining strong association rules in a transaction database.
* To minimize the negative rules.
* Filtering of uninteresting association rules.
* To optimize the performance.
  1. **Brief description**

**Data Mining Techniques:**

There are several major data mining techniques have been developed and used in data mining:

1. **Association Rules Mining**
2. **Clustering Mining**
3. **Classification Mining**
   * 1. **Association Rule Mining**

Association rule mining is a method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness.Based on the concept of strong rules, Rakesh Agrawal et al.[31] introduced association rules for discovering regularities between products in large-scale transaction data recorded by point-of-sale (POS) systems in supermarkets. For example, the rule {Onions, Potatoes}→{burger} found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, they are likely to also buy hamburger meat. There are two important basic measures for association rules, support(s) and confidence(c). The General architecture of Association rule mining is as follows.



**Figure 1.1: Association rule mining architecture.**

* + 1. **Clustering Mining**

Clustering is a process of grouping to similar data. Clustering is the task of identifying a finite set of categories (or clusters) to describe the data [26, 28]. Thus, similar objects are assigned to the same category and dissimilar ones to different categories. Clustering is also called unsupervised learning and segmentation because the data objects are mapped to a set of clusters which can be interpreted as classes as well.

**Figure 1.2: Clustering Mining.**

The process of clustering is similar data points grouped together into clusters. Most popular clustering algorithms suffer from some major issues they are following:

* First, the number of clusters is pre-deterministic, which makes them inadequate for batch processing of huge amount of image database.
* Secondly, the clusters are represented by their centroid and built using a Euclidean distance therefore inducing generally a hyper spherical cluster shape, which makes them not able to capture the real structure of the data base.
* This is especially true in the case of colour clustering where clusters are arbitrarily shaped.
  + 1. **Classification Mining**

Classification is a data mining technique that typically involves three phases, a learning phase, a testing phase and an application phase. A learning model or classifier is built during the learning phase. It may be in the form of classification rules, a decision tree, or a mathematical formula. Since the class label of each training sample is provided, this approach is known as supervised learning. [12, 14] In unsupervised learning (clustering), the class labels are not known in advance. In the testing phase test data are used to assess the accuracy of classifier. If the classifier passes the test phase, it is used for the classification of new, unclassified data tuple’s. This is the application phase. The classifier predicts the class label for these new data samples.

For classiﬁcation algorithms, the two major problems on classifying a data stream are the inﬁnite length and the concept drift. The first one makes the traditional multi-pass classiﬁcation algorithms incapable of classifying a data stream for their requirement of inﬁnite storage and large amount of training time.[16] The second one makes the most static stream classiﬁcation algorithms incapable of classifying a data stream with concept drifts for the under-lying changes occurred in the stream. For a time changing data stream, an incremental updating manner of the classiﬁer is very important. A temporal model is used to capture the evolutions of the stream. In general, the classiﬁcation process is always accompanied by the course of model construction and test. The classiﬁcation model keeps changing with the progression of the stream. If a static classiﬁer is used to classify an evolving data stream, the accuracy of it will drop greatly. For a sudden burst of concept drift in a time-changing stream, an up to date model always provides better accuracy. But for relatively stable time-changing streams, models built with long-term samples will be great. Both of the short-term and long-term behaviours of the stream are important for a classiﬁer. It is decided by the stream itself and cannot be known a priori.

1. **Literature** **Review**

This chapter gives an extensive literature survey on the Association rule mining and their methods. Various research papers and journals have been studied and come to know about Association rule mining.

In 2009, Gavin Shaw et al. [1] propose two approaches which measure multi-level association rules to help evaluate their interestingness. These measures of diversity and peculiarity can be used to help identify those rules from multi-level datasets that are potentially useful. Abstract Association rule mining is one technique that is widely used when querying databases, especially those that are transactional, in order to obtain useful associations or correlations among sets of items. Much work has been done focusing on efficiency, effectiveness and redundancy. There has also been a focusing on the quality of rules from single level datasets with many interestingness measures proposed. However, with multi-level datasets now being common there is a lack of interestingness measures developed for multi-level and cross-level rules. Single level measures do not take into account the hierarchy found in a multi-level dataset. This leaves the Support-Confidence approach, which does not consider the hierarchy anyway and has other drawbacks, as one of the few measures available.

In 2010, Xiufeng Piao et al. [2], presents a new algorithm which is based on the correlation and dual confidence. According to authors traditional approaches based on support-confidence framework maybe generate a great number of redundant and wrong association rules. To solve the problems, a correlation measure is defined and added to the mining algorithm for association rules. According to the value of correlation measure, association rules are classified into positive and negative association rules. Therefore, the new algorithm can mine the negative-item-contained rules. The proposed algorithm which is based on the correlation and dual confidence, can mine the positive and negative association rules. The experimental result shows that positive and negative association rules mining algorithm can reduce the scale of meaningless association rules, and mine a lot of interesting negative association rules. Author’s said many databases which use these mining technologies cannot find the hidden patterns, one of these important hidden pattern is the negative association rules, which has a low frequency and strong correlations, showing the property of the strong correlations in the data itemsets which is hard to find. The rules tell us that which data items less to occur together, and they share a very strong correlation, and contain very valuable information, such as rules *A*⇒￢*B,* ￢*A*⇒ *B* ,￢*A*⇒￢*B* ,so mining negative association rules is very important.

In 2011, Li Guang-yuan et al[3], present an algorithm for mining association rules with multiple constraints, the proposed algorithm simultaneously copes with two different kinds of constraints, it consists of three phases, first, the frequent 1-itemset are generated, second, they exploit the properties of the given constraints to prune search space or save constraint checking in the conditional databases. Third, for each item set possible to satisfy the constraint, authors generate its conditional database and perform the three phases in the conditional database recursively. Experimental results show that the method outperform the revised FP-growth algorithm. The problem of discovering all frequent item sets that satisfy constraints is a difficult one, the difficulty stems from the fact that, first, testing for minimum support and maximum support cannot be done simultaneously, since when valid, one is always true for subsets while the other is always true for supersets. Second, despite their selective power, some constraints cannot be checked to filter candidate item sets until a very late stage of the mining process depending upon the type of constraint and the search space traversal strategy used. However, there are some efficient algorithms proposed to deal with this problem, but most of these algorithms only cope with one constraint, authors present an algorithm to mine association rules with multiple constraints, it copes with two different kinds of constraints simultaneously. The Experimental results are conducted on a Pentium IV3.2 GHz personal computer with 2MB main memory, the data set is denotes as V25F20T50I1L100, which V25 denotes that the average size of the transactions is 25, F25 denotes that the average size of the maximal potentially frequent itemsets is 20, T50 denotes that the number of transactions is 50K, I1 denotes that the number of items is 1K, L100 denotes that the number of maximal potentially frequent itemsets is 100.

In 2012, Xiaobing Liu , et al. [4] present a form of the directed item sets graph to store the information of frequent item sets of transaction databases, and give the trifurcate linked list storage structure of directed item sets graph. Furthermore, authors developed the mining algorithm of maximal frequent item sets based on this structure. As a result, one realizes scanning a database only once, and improves storage efficiency of data structure and time efficiency of mining algorithm. Authors introduce a directed item sets graph to store the information of frequent item sets of transaction databases. Next create the trifurcate linked list storage structure of directed item sets graph, and ﬁnally develop the mining algorithm of maximal frequent item sets based on directed item sets graph. The realization of the process in this manner leads to a single scanning of the databases. It also improves storage efficiency of data structures and time efficiency of the mining algorithm itself. Max-Miner and Pincer-Search search the item sets lattice in a breadth ﬁrst manner to ﬁnd MFI. The former algorithm uses a look-ahead strategy to prune branches from the item sets lattice by quickly identifying long frequent item sets. The latter combines both the bottom-up and top-down searches. The Depth-Project algorithm searches the item sets lattice in a depth ﬁrst manner to ﬁnd MFI. To reduce search space, it also uses dynamic reordering of children nodes, superset pruning, an improved counting method and a projection mechanism. Smart Miner uses tail information to guide the depth ﬁrst search which is able to take advantage of the information gathered from previous steps to search for MFI. MAFIA uses a vertical format to represent the database, which al-lows efficient support degree counting and enhances the effect of look-ahead pruning in general. Unlike Depth-Project and MAFIA, GenMax returns the exact MFI. It utilizes a backtracking search for efficiently enumerating all maximal patterns. Besides, it represents the database in a vertical TID set format like VIPER and uses diffset propagation to perform fast support degree counting. GenMax has the better performance in the large data sets, but MAFIA exhibits better performance on small data sets as demonstrated through experiments.

In 2012, Ying-Ho Liu[5], has propose a new algorithm called U2P-Miner for mining frequent U2 patterns from univariate uncertain data, where each attribute in a transaction is associated with a quan-titative interval and a probability density function. The algorithm is implemented in two phases. First, author constructs a U2P-tree that compresses the information in the target database. Then, author use the U2P-tree to discover frequent U2 patterns. Potential frequent U2 patterns are derived by combining base intervals and verified by traversing the U2P-tree.

Author has develops two techniques to speed up the mining process. Since the proposed method is based on a tree-traversing strategy, it is both efficient and scalable. The experimental results demonstrate that the U2P-Miner algorithm outperforms three widely used algorithms, namely, the modified Apriori, modified H-mine, and modified depth-first backtracking algorithms. Author evaluates the performance of the U2P-Miner algorithm on synthetic and real datasets. Then the above-mentioned algorithms are modified to compare their performance with that of U2P-Miner. The H-mine algorithm constructs an H struct for the database, and a linked array called a Header table is maintained to link the occurrences of each item in the transactions. The modified H-mine algorithm is similar to the UH-mine algorithm, which is a variant of the H-mine algorithm for dealing with item set uncertain data.

In 2013, Idheba Mohamad Ali O. Swesi et al. [6] has proposed model which is integration between two algorithms, the Positive Negative Association Rule (PNAR) algorithm and the Interesting Multiple Level Minimum Supports (IMLMS) algorithm, to propose a new approach (PNAR\_IMLMS) for mining both negative and positive association rules from the interesting frequent and infrequent item sets mined by the IMLMS model.

The experimental results show that the PNAR\_IMLMS model provides significantly better results than the previous model.

The purpose of association rule mining is to find certain associations between a set of items in a database. As infrequent item sets become more significant for mining the negative association rules that play an important role in decision making, this study proposes a new algorithm for efficiently mining positive and negative association rules in a transaction database. The algorithm is called PNAR\_IMLMS and is appropriate for mining positive association rules from frequent item sets and negative association rules from both frequent and infrequent item sets discovered by the IMLMS model. The IMLMS model adopted an effective pruning method to prune uninteresting item sets. An interesting measure VARCC is applied that avoids generating uninteresting rules that may be discovered when mining positive and negative association rules.

In 2013, Huang QingLan & DuanLongZhen.[7] described the method which is combined with the concept of hierarchical concept, the data of the generalization sets processing, and uses SOFM neural network generalization into the database after the transaction, by way of introducing an internal threshold so no need to set the minimum support threshold, to generate the local frequent item sets as global candidates item sets to generate global frequent item sets, thereby enhancing the efficiency of multi-level association rules and accuracy. And by simulating the case shows that the method can not only efficient mining single-layer and cross-layer association rules, but also the association rules is new ,easy to understand and meaningful.

Based on theoretical research of the multi-level association rules mining, Top - k frequent pattern mining and SOFM clustering and classification methods, according to the problem of that the traditional multi-level association rules mining based on the Apriori algorithm is not adaptable to a large database. The experiment shows that the algorithm is feasible and effective. Authors also focused on parameter Settings difficult problems in the method and this method‘s application in Web mining fields for further study.

In 2013, Vidhu Singhal & Gopal Pandey.[8] described about the Association rule and clustering and the details are, A web based recommendation system include browsing history database enclosed with information related to the web pages that a user browsed. Authors presents the Prediction of User navigation patterns of WUM using Association Rule and Clustering from web log data. In the first stage, separating the potential users is processed, and in the second stage clustering process is used to group the users with similar interest, and in the third stage association and clustering is used to navigate the user future requests. The experimental results are really encouraging and produce valuable information.

The main objective is to help achieve better prediction accuracy for Web page access. Recommending a next page, the Web user’s access is very vital for diverse Web applications. The main technology implemented for this purpose is through using Web usage mining pattern discovery techniques. By keeping the models limitations to a minimum and relying on their advantages according to different constraints, it become possible to achieve more accurate prediction results.

In 2014, Aritra Roy, & Rajdeep Chatterjee. [11] presented an extensive analysis of the ARWDC approach for different sizes of Reuter’s datasets. Furthermore the performance of our approach is evaluated with the help of evaluation measures such as, Precision, Recall and F-measure compared to the existing clustering algorithms like Bisecting K-means and FIHC. The experimental results show that the efficiency, scalability and accuracy of the ARWDC approach has been improved significantly for Reuters datasets.

Although standard clustering techniques such as k-means can be applied to document clustering, they usually do not satisfy the special requirements for clustering documents: high dimensionality, high volume of data, ease for browsing, and meaningful cluster labels. In addition, many existing document clustering algorithms require the user to specify the number of clusters as an input parameter. Incorrect estimation of the value always leads to poor clustering accuracy. Furthermore, many clustering algorithms are not robust enough to handle different types of document sets in a real-world environment. In some document sets, cluster sizes may vary from few to thousands of documents.

Authors have conducted an extensive analysis of association rules-based web document clustering ARWDC approach. The largest dataset, Reuters, is chosen to exam the efficiency and scalability of our algorithm. The experimental results show that at different sizes of Reuter’s datasets, the ARWDC approach improved scalability. Furthermore when compared with other clustering algorithms like Bisecting K-means and FIHC, the accuracy and efficiency are improved. Moreover, ARWDC approach associated a meaningful label to each final cluster. Then the user can easily find out what the cluster is about since the label can provide an adequate description of the cluster based on Association Rules. However, it is time-consuming to determine the labels after the clustering process is finished. From all experiments, authors conclude that ARWDC approach has favorable quality in clustering documents using Association Rules. The importance of document clustering will continue to grow along with the massive volumes of web documents. With the standardization of XML as an information exchange language over the web, documents formatted in XML have become quite popular. Moreover, most of the clustering algorithms of MEDLINE abstracts are based on pre-defined categories. In future, authors intend to apply ARWDC approach for automatically clustering the MEDLINE abstracts formatted in XML to help biomedical researchers in quickly finding relevant and important articles related to their research field without need to predefine categories.

In 2014, Deepak A. Vidhate et al. [12] Presents Multilevel Relationship Algorithm (MRA) approach. This approach is the combination of Multilevel Apriori algorithm and Bayesian probability estimation. The algorithm consists of three sub modules: MRA Stage I, MRA stage II and Interdependency Module.



**Figure 2.1: MRA Architecture**

In MRA Stage I, it finds Level 1 association amongst items or internal relationship between the same item types.InMRA Stage II*:* it uses individual knowledge base and level 1 association which is generated in stage 1 to find out the frequent item sets. In Stage III: it finds Interdependency by Bayesian Probability, Bayesian Probability is used to determine dynamic behavior of a particular season and External Dependencies amongst Items. New patterns are generated by Bayesian probability through which learning rules are predicted & interpreted.

In 2011, Peter P. Wakabi–Waiswa & Venansius Baryamureeba[34] has proposed the Mining Optimized Association Rules Algorithm (MOAR) which maintains two populations: the internal population P, and a Pareto−store, *Ṗ*. within each generation individuals are selected from the set of the non–dominated solutions for reproduction using crossover and mutation before updating the Pareto–store. Author has adopted a modified Michigan approach whereby the encoding/decoding scheme associates two bits to each attribute in the database. If these two bits are 00 then the attribute next to these two bits appears in the antecedent part and if it is 11 then the attribute appears in the consequent part. The other two combinations, 01 and 10 indicate the absence of the attribute from the rule. Following this approach the algorithm can handle variable length rules with more storage efficiency, adding only an overhead of 2k bits, where k is the number of attributes in the database. The reproduction mechanism involves rule selection and the application of the crossover operators. The rule selection method used by MOAR follows the “universal suffrage” approach. With this approach each association rule is represented by a single individual. The individuals to be mated are elected by training data instances. Each instance votes for a rule that it covers in a stochastic fitness–based way. Author’s modified the standard crossover operator to either generalize the crossover operator if the rule is too specific, or to specialize it if the rule is too general. A rule is considered too specific if it covers too few data instances i.e. when too few data instances satisfy both the antecedent and the consequent of the rule. In contrast, a rule is considered too general when it covers too many data instances. Author’s make use of the bitwise OR and the bitwise AND to implement generalization and specialization, respectively. The bitwise OR and the bitwise AND are applied to the antecedent part of the rule. Author’s also adopted the non−uniform mutation operator. The non-uniform mutation operator adapts to the environment by varying as the fitness of the individuals in the population change. Author’s made a modification to the non–uniform mutation operator to enable it to generalize or specialize a rule condition.

The resulting mutation operator randomly selects a condition from the rule. If that condition involves a nominal attribute, then the value is randomly changed from one value to another. If the attribute is continuous, the operator randomly changes the condition’s interval values. MOAR use an elitist individual replacement approach that ensures that more fit genotypes are always introduced into the population.

**Association Rule**

The formal statement of association rule mining problem was firstly stated in by Agrawal et al. in 2009.[32] Let I= I1, I2, … , Im be a set of m distinct attributes, T be transaction that contains a set of items such that T I, D be a database with different transaction records Ts. An association rule is an implication in the form of X → Y, where X, Y I are sets of items called item sets, and X ∩ Y = ø. X is called antecedent while Y is called consequent, the rule means X implies Y. There are two important basic measures for association rules, support(S) and confidence(C). Since the database is large and users concern about only those frequently purchased items, usually thresholds of support and confidence are predefined by users to drop those rules that are not so interesting or useful. The two thresholds are called minimal support and minimal confidence respectively, additional constraints of interesting rules also can be specified by the users. The two basic parameters of Association Rule Mining (ARM) are: support and confidence. Support(s) of an association rule is defined as the percentage/fraction of records that contain X U Y to the total number of records in the database. The count for each item is increased by one every time the item is encountered in different transaction T in database D during the scanning process. It means the support count does not take the quantity of the item into account. For example in a transaction a customer buys three bottles of beers but only support count number of {beer} is increased by one, in another word if a transaction contains a item then the support count of this item is increased by one. Support(s) is calculated by the following formula:



Support (XY) = Support count of XY\_\_\_\_

Total number of transaction in D

Support of an item is a statistical significance of an association rule. Suppose the support of an item is 0.1%, it means only 0.1 percent of the transaction contain purchasing of this item. The retailer will not pay much attention to such kind of items that are not bought so frequently, obviously a high support is desired for more interesting association rules.

Confidence of an association rule is defined as the percentage/fraction of the number of transactions that contain X U Y to the total number of records that contain X, where if the percentage exceeds the threshold of confidence an interesting association rule X→Y can be generated.

Confidence (XjY) = Support (XY)

Support(X)

Confidence is a measure of strength of the association rules, suppose the confidence of the association rule X→Y is 80%, it means that 80% of the transactions that contain X also contain Y together, similarly to ensure the interestingness of the rules specified minimum confidence is also pre-defined by users.

* + 1. **Association Rule Mining in Large Database**

Association rule mining used to mine the sales transactions between items in large database recognized as a most significant area of database research. Measuring a large database there are different techniques are used.[13] Pruning strategy and interestingness is one of the measuring techniques for measuring large database. Large database consists of many fields. Each field consists of their own process. They different depends on their field of work. Suppose a customer transaction of a large database each transaction consists of items purchased by a customer in a visit items purchased by a customer in a visit , time of purchase, category of payment, net amount etc. so it is a tedious process to maintain for huge amount of customer transaction. An efficient algorithm implemented in association rule mining. Apriori algorithm is best for association rule mining in large database.[16,20] This algorithm generates all significant association rules between items in the large database. Today, most research related work on data mining in association rules are encouraged by an wide range of application areas, such as financial transactions, engineering, health care, GIS, and broadcastings. Association rule mining used to originate interesting association or correlation relationships among a large set of items in the large database. In large database Application of association rule mining in market basket analysis are:-

* To analyses the point of sales transaction.
* From uses information on what customers buy to provide insights into who they are and why they make certain purchases.
* From which products are purchased together and which are most willing to support.
  1. **Association Rule Mining in Medical Database**

Associative classification rule mining is a combination of association rule mining integrated with classification rule mining. It is used in medical data base. Association rules can expose biologically significant associations between different genes or between environmentally friendly belongings and gene expression.[22] An association rule has the form LHS⇒RHS, where LHS and RHS are disjoint sets of items, the RHS set being likely to happen at whatever time the LHS set occurs. Items in gene expression data can include genes that are extremely articulated or inhibited, as well as related facts labeling the cellular atmosphere of the genes. Association rules relate disease data measures the patient risk factors and appearance of the disease in medical terms. Association rule medical consequence is estimated with the usual support and confidence metrics. Association rules are used to compare analytical rules mined with decision trees, a well-known machine learning method.[23]

* 1. **Association Rule Mining in Distributed Database**

Databases or data warehouses may store a large amount of data (large database) to be mined. Mining association rules in large databases may require extensive processing power. Distributed system is to solve this problem in large database mining. Many large databases are distributed, more feasible to use Distributed algorithms by distributed system. Distributed computing of large item sets encounters certain different complications.[24] To solve this complications by using different distributed algorithms. Such as:-

* Distributed association rule learning.
* Distributed hierarchical clustering.
* Collective PCA and PCA-based clustering.
* Collective decision tree learning.
* Collective Bayesian network learning.

Centralized data mining to discover useful patterns in distributed databases isn't always feasible because merging data sets from different sites sustains huge network communication costs. Distributed higher-order association rule mining algorithm is to determine propositional rules established on higher-order associations in a distributed surroundings and also detect a critical suppositions made in existing association rule mining algorithms that preclude them from scaling to complex distributed surroundings in which the complete global schema is indefinite, data is inconsistent in a hybrid non-vertical, non-horizontal form, and errors occur in record linkage.

* 1. **Association Rule Mining in Relational Database**

An ever growing number of organizations are installing large data warehouses using relational database technology. There is a huge demand for mining bits of knowledge from these data warehouses. Association rule mining is used to make a decision to solve this problem. Relational association rules and supervised learning methods help to identify the probability of illness in a certain disease. This interface can be simply extended by adding new symptoms types for the given disease, and by defining new relations between these symptoms.[27]

Many industrial databases applications make use of relational databases. It is used to store, manipulate and re-claim regulated data from large database. Through association rule mining from relational databases utilize database indexing and query optimization procedures applied in relational database management systems to develop performance and improve efficiency. Association Rule mining in the relational database is the process of recognizing the dependency of one item(s) with respect to the existence of other item(s). This helps to study the buying patterns of their customers.



**Figure 2.2: A Multi-relational Database Environment.**

* 1. **Association Rule Mining in Spatial Database**

Spatial Data Mining is the discovery of fascinating patterns from large geospatial databases. It refers to the extraction of knowledge, spatial associations or other fascinating patterns not clearly stored in spatial databases. In data mining association rules are encompassing spatial relations among spatial substances. Spatial database contains objects which are described by a spatial scene and/or extension as well as by several non-spatial attributes. Spatial data mining algorithms have to consider the neighbors of substances in order to mine useful knowledge.[28] It is indispensable because the attributes of the neighbors of some substance of curiosity may have a momentous inspiration on the substance itself. The application of data mining techniques in spatial database to census data, and more generally, to official data, has great potential in supporting worthy public strategy and in sustaining the actual operational of an independent society. Spatial data mining approaches and procedures have been suggested for the mining of hidden knowledge, spatial relations, or other patterns not clearly stored in spatial databases.

Spatial data mining is used in:-

* NASA Earth Observing System (EOS) for Earth science data.
* Census Bureau, Dept. of Commerce for census data.
* Dept. of Transportation (DOT) for traffic data National Inst. of Health (NIH) for cancer clusters.

1. **System Analysis and Design**

Validation of association rule mining is very important and critical phase of rule generation. The dependence of rule generation has two factors one is minimum support value and other is minimum confidence value. The process of validation satisfy the given condition and used for the process of rule generation. The generated rules come along with some negative rule and some positive rule in form of rule set. Now the process of strong rule generation used various constraints function such as monotonic and non-monotonic function for range validation of association rule mining. Some authors have also used some multi-level approach for generation of association rule mining.

* 1. **Problem** **Definition**

The Knowledge Discovery in Databases (KDD) field is concerned with the development of methods and techniques for making sense of data. Association rule mining identifies collections of data attributes that are statistically related in the underlying data. An association rule is an expression of the form X=>Y where X and Y are disjoint sets of items. In a dataset D, consisting of data instances where every instance is a set of items, the rule X=>Y has support sup, equal to the percentage of the instances of D that contain both X and Y.

The association rule mining suffered following problem.

1. Scanning and pruning problem of dataset
2. Generation of negative and positive rules.
3. Superiority measure problem.
4. Number of passes over the database.
5. Sampling problem of data set
   1. **Feasibility Study**

The feasibility study is an evaluation and analysis of the potential of a proposed project, which is based on extensive investigation and research to support the process of decision making. A well-designed feasibility study should provide a historical background of the project. A project must be feasible in all three ways to merit further development.

* + 1. **Technical Feasibility:**

A large part of determining recourse has to do with assessing technical feasibility. The analyst must find out whether current resources can be upgraded or added to in a manner that fulfills the request under consideration. Sometimes “add-ons” to existing systems are costly and not worthwhile, because they-meet needs inefficiently. If existing systems cannot be added onto, the next question becomes whether there is technology in existence that meets the specifications. The project is analyzed along with the technical resources which are required for developing the proposed system. The technical resources are found to be feasible.

* + 1. **Economic Feasibility:**

Economic feasibility is the second part of resource determination. The basic resources to consider are time and the cost of doing a full systems study including the estimated cost of hardware, and the estimated cost of software. The concerned business must be able to see the value of investment it is pondering before committing to an entire system study. If short term costs are not overshadowed by long term gains or produce no immediate reduction in operating costs, the system is not economically feasible and the project should not proceed any further. The resources required for developing the system are identified as software and hardware. The requirement of software and hardware are found to be economical.

* + 1. **Operational Feasibility:**

Consider for a moment that technical and economical resources are both judged adequate. The systems analyst must still consider the operational feasibility of the requested project. Operational feasibility is dependent on the human resources available for the project and involves projecting whether the system will operate and be used once it is installed.

* 1. **Requirement and Input Output Specifications**

Following are the Hardware Requirements:

* Processor: Minimum Intel Pentium IV or AMD Xeon
* RAM: Minimum 512 Mb

Following are the Software/Development Tool Requirement:

* MATLAB R20127.14.0.334

**Input Specification:**

Min Max constraints provided by the user will be used to generate results.

**Output Specification:**

The rules generated by the various algorithms will be used for result analysis.

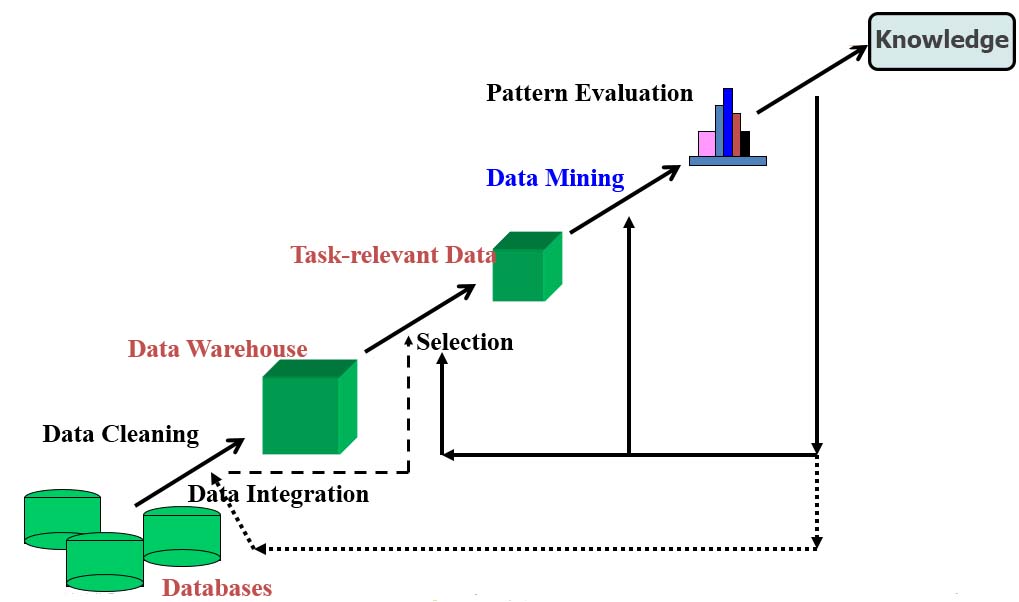
* 1. **System** **Design**
     1. **Proposed Design**

There is a large amount of data present in physical world and useful knowledge is hidden within these data. The extraction of knowledge from these dataset is a crucial thing because knowledge helps us in decision making. Data mining is a detailed process of analyzing large amounts of data and picking out the relevant information. It refers to extracting or mining knowledge from large amounts of data. It involves the following steps: cleaning and integrating data from data sources like databases, flat files, pretreatment of selecting and transformation target data, mining the required knowledge and finally evaluation and presentation of knowledge. Data mining can be described as an important tool to discover the knowledge. So, the whole process of producing knowledge from raw data is called knowledge discovery in databases (KDD). And data mining is just a part of KDD process. Association rules provide interesting correlations among data items. From these association rules, useful knowledge can be derived. Association rule mining is an aspect of data mining.



**Figure 3.1: Data mining process.**

Data mining methods are often used to detect patterns in a large set of data. These patterns are then used to identify future instances in a similar type of data. Numbers of data mining techniques are experimented to identify new malicious binaries. Here are three learning algorithms to train a set of classifiers on some publicly available malicious and benign executables. However, the algorithm also resulted in higher false positive rates when compared to signature-based method. The key to any data mining framework is the extraction of features, which are properties extracted from examples in the dataset. Schultz et al. extracted some static properties of the binaries as features. These include system resource information (the list of DLLs, the list of DLL function calls, and the number of different function calls within each DLL) obtained from the program header, and consecutive printable characters found in the files. The most informative feature they used was byte sequences, which were short sequences of machine code instructions generated by the hex dump tool. The features were used in three different training algorithms. There was an inductive rule-based learner that generated Boolean rules to learn what a malicious executable was; a probabilistic method that applied Bayes rule to compute the likelihood of a particular program being malicious, given its set of features; and a multi-classifier system that combined the output of other classifiers to give the most likely prediction.



**Figure 3.2: Pre-processing and Mining.**

The scanning of database divided into multiple levels as frequent level and infrequent level of data according to MLMS. The frequent data logically assigned 1 and infrequent data logically assigned 0 for MLMS process. The divided process reduces the uninteresting item in given database. The proposed algorithm is a combination of MLMS and min-max algorithm along this used level weight for the separation of frequent and infrequent item. The weight value act as Support length key is a vector value given by the transaction data set. The support value passes as a vector for finding a near level between MLMS candidates key. After finding a MLMS candidate key the nearest level divide into two levels, one level take a higher odder value and another level gain infrequent minimum support value for rule generation process. The process of selection of level also reduces the passes of data set. After finding a level of lower and higher of given support value, compare the value of level weight vector. Here level length vector work as a fitness function for selection process of min-max algorithm. Here steps of process of algorithm is presented step by step and finally draw a flow chart of complete process.

**Steps of algorithm (MLMS-GA)**

1. Scanning of database used flowing steps

Some standard notation of pseudo code of algorithm such as D dataset, K level MLMS, Ls generation candidate

K = MLMS dataset (D)

n = Number of multiple level block

**For** i = 1 to n loop

Scan\_k (Ki ∈k)

Li = gen\_\_itemsets (ki)

**For** (i = 2; Lj i≠ φ, j = 1,2....,n; i++)

Ci G = ∪j = 1,2,...nLij

End;

**For** i = 1 to n

scan\_kmap (ki∈K)

**For** all items C ∈CG generate block (C, ki)

End;

LG = {c ∈CG|}

2. Generate multiple support vector value for selection process

for all transaction LG do

generate count table TC

L1 =(frequent 1-itemsets);

C2 =L1 ∞ L1;

L2 ={cEC2 | sup(c)≥MinSupNum};

For(k=3;Lk-1 ≠Ø ;k++)do begin

For (j=k;j≤m;j++)do

Generate CIVijk-1;

Ck=candidate\_gen(Lk-1)

Lk ={cECk| sup(c)≥MinSupNum};\

End

3. Set of rule is generated

Return L = Ư Lk;

Candidate\_gen(frequent itemset Lk-1)

* 1. for all(K-1)-itemsetlE Lk-1 do
  2. for all ijE Lk-1 do
  3. //S is the result of the formula(2)

If for every r(1≤r≤k) such that S[r]≥k-1 then

L1 = (frequent 1-itemsets);

C2 =L1 ∞ L1;

L2 = {cEC2 | sup(c)≥MinSupNum};

For (k=3;Lk-1 ≠Ø ;k++)do begin

For (j=k;j≤m;j++)do

Generate CIVijk-1;

Ck=candidate\_gen(Lk-1)

4. Check MLMS value of table

5. If rule is not MLMS go to selection process

6. Else optimized rule is generated.

7. Exit

1. **Data Encoding**

The process of data in min-max algorithm needs some data encoding technique for representation of data. In this technique used binary encoding technique.

1. **Fitness function**

The population selection of Min-max Algorithm is a design of Fitness Function:

Ai = {frequent item support}

Wi1 = {level of Wight value of MLMS}

Bi = {those value or Data infrequent}

The selection strategy based on the basis of individual fitness and concentration pi is the probably of selection of individual whose fitness value is greater than one and m(s) is a those value whose fitness is less than one but near to the value of 1. The Min-max operators determine the search capability and convergence of the algorithm. Min-max operators hold the selection crossover and mutation on the population and generate the new population. In this algorithm it restore each chromosome in the population to the corresponding rule, and then calculate selection probability pi for each rule based on above formula. In which single point are used. It divide multiple level domain of each attribute into a group and classifies the cut point of each continuous attributes into one group .And the crossover carried out between the corresponding groups of two individuals by a certain rate. Any bit in the chromosomes is mutated by a certain rate that is, changing “0” to “1” & “1” to “0”. Now the complete process of algorithm shows block diagram of proposed algorithm using min-max algorithm.

**Load Data**

**Set Constraints Parameters**

**Start Scanning**

**Get Frequent Item**

**Partition Into Multilevel of All Frequent Items**

**If Level = 0**

**No**

**Yes**

**Set Population**

**Selection**

**Cross Over**

**Probability of Mutation**

**P=0.007**

**If Rule is Optimal?**

**No**

**Yes**

**Optimized Rule Set**

**Figure 3.3: Proposed model for association rule mining algorithm for large database.**

1. **System Implementation & Testing**

This dissertation work tries to perform experimental process of proposed method for Min–Max Constraints. The proposed method implemented in Matlab R20127.14.0.334 and tested with very reputed data set from UCI machine learning research center. The research work measured the total number of generated rules and the Elapsed time taken by each method. To evaluate these performance parameters four datasets have been used from UCI machine learning repository namely Chess, Chess less, Abalone and t 30 (Thoracic Surgery).

1. **Setting Environment**

MATLAB software package is used or the performance evaluation of Min-Max Constraints algorithm. MATLAB is a software package for high- performance numerical computation and visualization. It provides an interactive environment with hundreds of built-in function for technical computation, graphics and animation. Best of all, it also provides easy extensibility with its own high- level programming language. The MATLAB stands for matrix laboratory. There are also several optional "toolboxes” available from the developers of MATLAB. These toolboxes are collections of functions written for special applications such as symbolic computation, image processing, statistics, control system design, neural networks etc. the list of toolboxes keeps growing with time. There are now more than 50 such toolboxes. One of the best features of MATLAB is its platform independence. Once you are in MATLAB, for the most part, it does not matter which computer you are on. Almost all commands work the same way. The only commands that differ are the ones that necessarily depend on the local operating system, such as editing and saving M- files. Programs written in the MATLAB language work exactly the same way on all computers. MATLAB started as an interactive software package developed to perform numerical calculations on vectors and matrices. It is much more powerful because it supports two and three dimensions graphics, and also supports high-level programming languages .which makes it quite easy to code complicated algorithms involving vectors and matrices. It can numerically solve nonlinear initial and boundary value differential equations and MATLAB contains a wide variety of toolboxes which provide a help in science and engineering environment, MATLAB also provide a facility to create a MATLAB script that carry out a sequence of commands.

The efficiency of programs that are used often and by several different people can be improved by simplifying the input and output data management. The use of Graphic User Interfaces (GUI), which provides facilities such as menus, pushbuttons, sliders etc, allows Programs to be used without any knowledge of MATLAB. They also provide means for efficient data management. A graphic user interface is a MATLAB script file customized for repeated analysis of a specific type of problem. There are two ways to design a graphic user interface**.** It is a user interface built with graphical objects, such as buttons, text fields, sliders, and menus. In general, these objects already have meanings to most computer users. For example, when you move a slider, a value change; when you press an OK button, your settings are applied and the dialog box is dismissed. Of course, to leverage this built-in familiarity, you must be consistent in how you use the various GUI –building components. Applications that provide GUIs are generally easier to learn and use since the person using the application does not need to know what commands are available or how they work. The action that results from a particular user action can be made clear by the design of the interface. MATLAB implements GUIs as figure window containing various styles of control objects. You must program each object to perform the intended action when activated by the user of the GUI.

A GUI is a graphical display which allows a user to interact with a program. It usually consists of three elements:

Components: These are the tools which enable the user to interact with a program, and include push bottoms, sliders, radio buttons, check boxes, editable text, pop-up menus, list boxes, and toggle buttons. Such an interaction is called an event, and a program which responds to events is event driven. Components can also include plots and/or tables which allow the program to interact with the user that show results of the program. All of these components must be arranged within a figure, which is a graphical window which is separate from the MATLAB window which arises when MATLAB is first executed.

1. **Data Set’s**

In this section, firstly the description of the dataset is revealed. Four datasets are used in this dissertation, namely Chess, Chess less, Abalone and t 30 (Thoracic Surgery) dataset which are taken from UCI Machine Learning Repository.

* + - 1. **Chess and Chess Less Data Set**

[chess\_flann\_new].

legal\_move(state1,NS,black).

legal\_move(state1,NS,white).

[chess\_flann\_wyl].

legal\_move(state1,Newstate,black).

legal\_move(state1,Newstate,white).

[chess\_russell\_wyl].

legal\_move(state1,Newstate,black).

legal\_move(state1,Newstate,white).

[chess\_vijay\_1].

legal\_move(state1,Newstate,black).

legal\_move(state1,Newstate,white).

[chess\_vijay\_2].

legal\_move(state1,Newstate,black).

legal\_move(state1,Newstate,white).

[chess\_vijay\_3].

legal\_move(state1,Newstate,black).

legal\_move(state1,Newstate,white).

1. Chess\_flann\_new: Written by flann@cs.orst.edu. Employs a geometric representation for states, with each square designated by an X, Y coordinate and square connectivity computed by vectors. Generates legal moves by first generating peusdo moves then eliminating those that result in the moving player being in check.
2. Chess\_flann\_wyl: Written by flann@cs.orst.edu. Employs a relational representation for states, with each square given a unique name and square connectivity computed by an enumeration of connected relations. Generates legal moves by first generating peusdo moves then eliminating those that result in the moving player being in check.
3. Chess\_russell\_wyl: Originally written by Stuart Russell in MRS, translated into prolog by flann@cs.orst.edu. Employs a geometric representation for states, with each square designated by an X, Y coordinate and square connectivity computed by vectors. Generates legal moves by determining whether the moving side is in check or uncheck. If the moving side is in check, moves are generated that destroy the check threat. If the moving side is not in check, moves are generated that do not create a check threat. Note that if the moving side is in check from multiple threats then the domain theory generates incorrect moves.
4. Chess\_vijay\_1: Written by vijay@cs.orst.edu. Employs a relational representation for states, with each square given a unique name and square connectivity computed by an enumeration of connected relations.
5. Chess\_vijay\_2: Written by vijay@cs.orst.edu. Employs a geometric representation for states, with each square designated by an X, Y coordinate and square connectivity computed by vectors.
6. Chess\_vijay\_3: Written by vijay@cs.orst.edu. Employs a special linear representation for states, with each square designated by a single number and square connectivity computed by a single delta value.
   * + 1. **Abalone Data Set**

The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope a boring and time-consuming task. Other measurements, which are easier to obtain, are used to predict the age. Further information, such as weather patterns and location (hence food availability) may be required to solve the problem.  From the original data examples with missing values were removed (the majority having the predicted value missing), and the ranges of the continuous values have been scaled for use with an ANN (by dividing by 200). The number of attribute is 8 and given instance is 4177.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sr No.** | **Name** | **Data type** | **Meas.** | **Description** |
| 1 | Gender | nominal |  | M, F, and I (infant) |
| 2 | Length | continuous | Mmm | Longest shell measurement |
| 3 | Diameter | continuous | Mmm | perpendicular to length |
| 4 | Height | continuous | Mmm | with meat in shell |
| 5 | Whole weight | continuous | Grams | whole abalone |
| 6 | Shucked Weight | continuous | Grams | weight of meat |
| 7 | Viscera weight | continuous | Grams | gut weight (after bleeding) |
| 8 | Shell weight | continuous | Grams | after being dried |
| 9 | Rings | Integer |  | +1.5 gives the age in years |

**Table 4.1: Abalone data set attributes information.**

|  |
| --- |
| M,0.455,0.365,0.095,0.514,0.2245,0.101,0.15,15  M,0.35,0.265,0.09,0.2255,0.0995,0.0485,0.07,7  F,0.53,0.42,0.135,0.677,0.2565,0.1415,0.21,9  M,0.44,0.365,0.125,0.516,0.2155,0.114,0.155,10  I,0.33,0.255,0.08,0.205,0.0895,0.0395,0.055,7  I,0.425,0.3,0.095,0.3515,0.141,0.0775,0.12,8  F,0.53,0.415,0.15,0.7775,0.237,0.1415,0.33,20  F,0.545,0.425,0.125,0.768,0.294,0.1495,0.26,16  M,0.475,0.37,0.125,0.5095,0.2165,0.1125,0.165,9  F,0.55,0.44,0.15,0.8945,0.3145,0.151,0.32,19  F,0.525,0.38,0.14,0.6065,0.194,0.1475,0.21,14  M,0.43,0.35,0.11,0.406,0.1675,0.081,0.135,10  M,0.49,0.38,0.135,0.5415,0.2175,0.095,0.19,11  F,0.535,0.405,0.145,0.6845,0.2725,0.171,0.205,10  F,0.47,0.355,0.1,0.4755,0.1675,0.0805,0.185,10 |

**Table 4.2: Sample of abalone data set.**

* + - 1. **Thoracic Surgery Data Set**

The data was collected retrospectively at Wroclaw Thoracic Surgery Centre for patients who underwent major lung resections for primary lung cancer in the years 2011. The Centre is associated with the Department of Thoracic Surgery of the Medical University of Wroclaw and Lower-Silesian Centre for Pulmonary Diseases, Poland, while the research database constitutes a part of the National Lung Cancer Registry, administered by the Institute of Tuberculosis and Pulmonary Diseases in Warsaw, Poland. The no. of attribute is 17 and the no. of u=instance is 470.

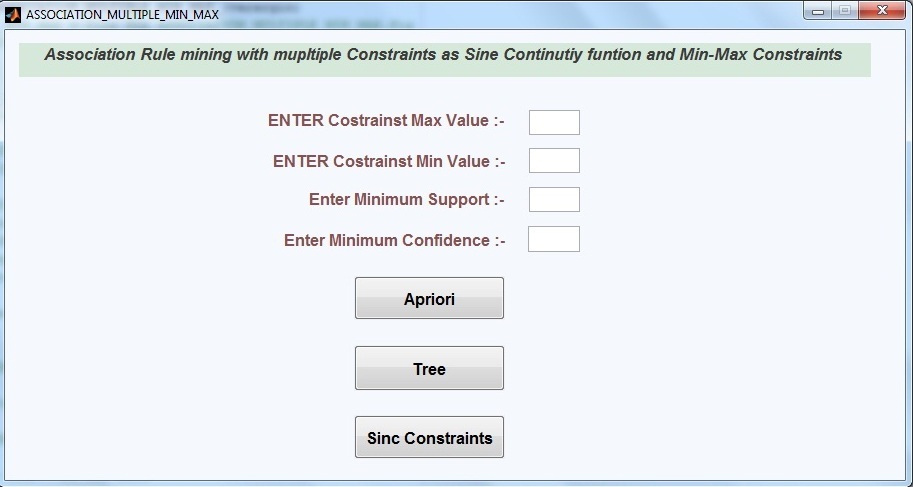
|  |
| --- |
| DGN2,2.88,2.16,PRZ1,F,F,F,T,T,OC14,F,F,F,T,F,60,F  DGN3,3.4,1.88,PRZ0,F,F,F,F,F,OC12,F,F,F,T,F,51,F  DGN3,2.76,2.08,PRZ1,F,F,F,T,F,OC11,F,F,F,T,F,59,F  DGN3,3.68,3.04,PRZ0,F,F,F,F,F,OC11,F,F,F,F,F,54,F  DGN3,2.44,0.96,PRZ2,F,T,F,T,T,OC11,F,F,F,T,F,73,T  DGN3,2.48,1.88,PRZ1,F,F,F,T,F,OC11,F,F,F,F,F,51,F  DGN3,4.36,3.28,PRZ1,F,F,F,T,F,OC12,T,F,F,T,F,59,T  DGN2,3.19,2.5,PRZ1,F,F,F,T,F,OC11,F,F,T,T,F,66,T  DGN3,3.16,2.64,PRZ2,F,F,F,T,T,OC11,F,F,F,T,F,68,F  DGN3,2.32,2.16,PRZ1,F,F,F,T,F,OC11,F,F,F,T,F,54,F  DGN3,2.56,2.32,PRZ0,F,T,F,T,F,OC12,F,F,F,F,F,60,F  DGN3,4.28,4.44,PRZ1,F,F,F,F,F,OC12,F,F,F,T,F,58,F  DGN3,3,2.36,PRZ1,F,F,F,T,T,OC11,F,F,F,T,F,68,F  DGN2,3.98,3.06,PRZ2,F,F,F,T,T,OC14,F,F,F,T,F,80,T  DGN3,1.96,1.4,PRZ1,F,F,F,T,F,OC11,F,F,F,T,F,77,F  DGN3,4.68,4.16,PRZ1,F,F,F,T,F,OC12,F,F,F,T,F,62,F  DGN2,2.21,1.88,PRZ0,F,T,F,F,F,OC12,F,F,F,T,F,56,F  DGN2,2.96,1.67,PRZ0,F,F,F,F,F,OC12,F,F,F,T,F,61,F  DGN3,2.6,1.68,PRZ1,F,F,F,T,F,OC12,F,F,F,T,F,70,F  DGN3,2.88,2.48,PRZ0,F,F,F,F,F,OC11,F,F,F,T,F,71,F  DGN3,4.48,3.48,PRZ0,F,F,F,F,F,OC12,F,F,F,T,F,51,F  DGN4,3.32,2.84,PRZ0,F,F,F,F,F,OC12,F,F,F,T,F,62,F |

**Table 4.3: Sample of Thoracic Surgery data set.**

* 1. **Experimental Results**

The System Execution details are shown with the help of screenshots,

First run the matlab R2012a from the run command, then browse to the project folder from the matlab or directly open the .m file from the file menu of matlab and then execute the ASSOCIATION\_MULTIPLE\_MIN\_MAX\_.m file of desertation.



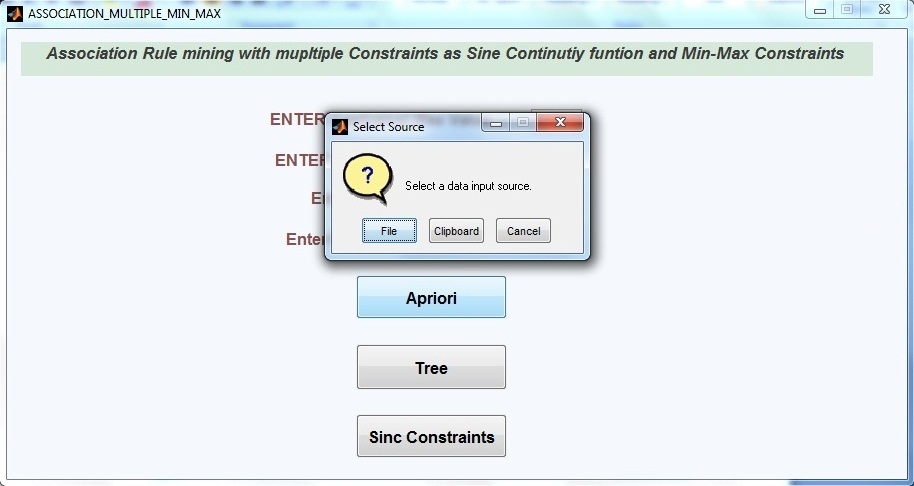
**Screenshot 4.1: The Main Window Initially Empty For Experimental Results.**

Screenshot shows the values given for Max Constraints = 0.4, Min Constraints = 0.3, Minimum Support = 0.5, & Minimum Confidence = 0.6



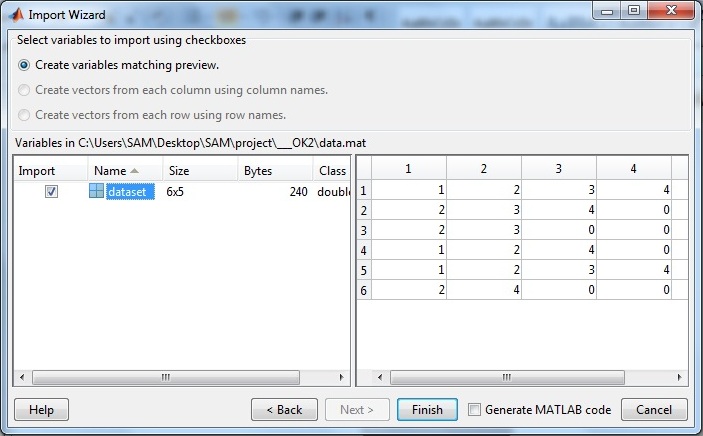
**Screenshot 4.2: The Main Window With 4 parameters.(1)**

Screenshot shows the popup message asking to select the source dataset by clicking on the file button.



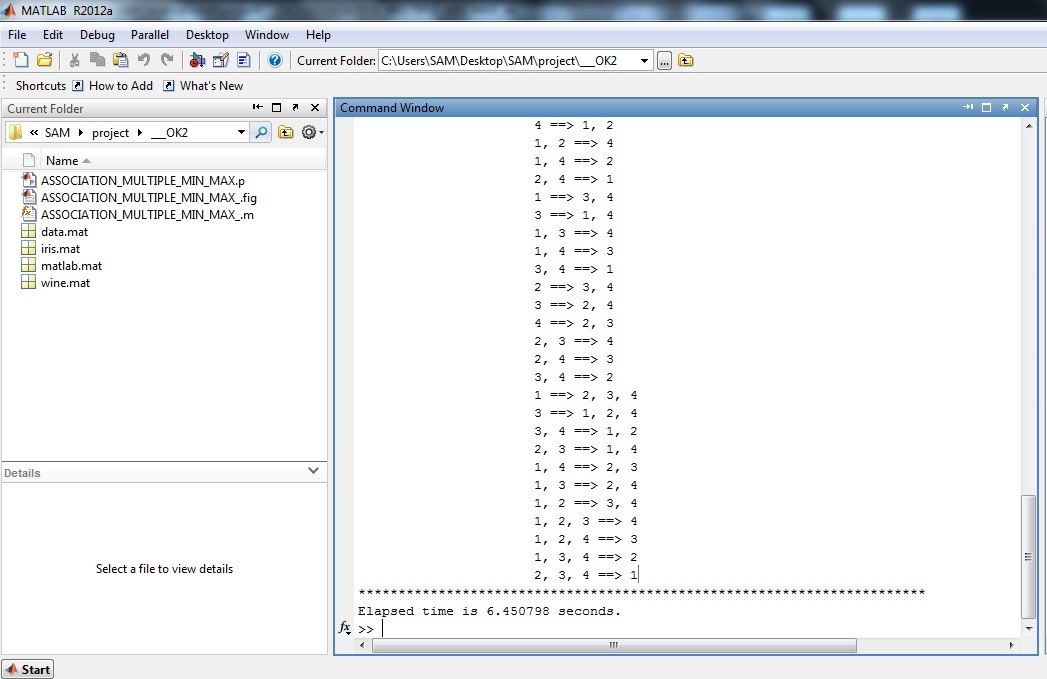
**Screenshot 4.3: The Popup Window With Load The Data For Experimental Results.**

Screenshots shows import wizard after clicking file button, click perticular dataset and click on the finish button.



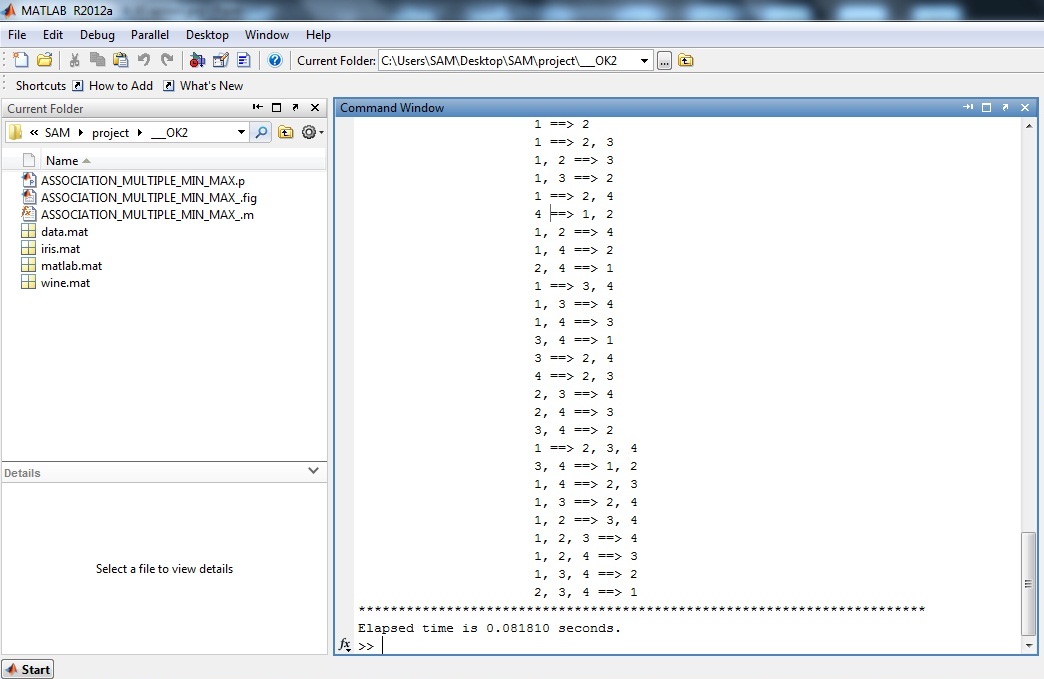
**Screenshot 4.4: Import Wizard Window**

Screenshots shows the total rules generated are 48 with elapsed time 6.45 using Apriory Algorithm.



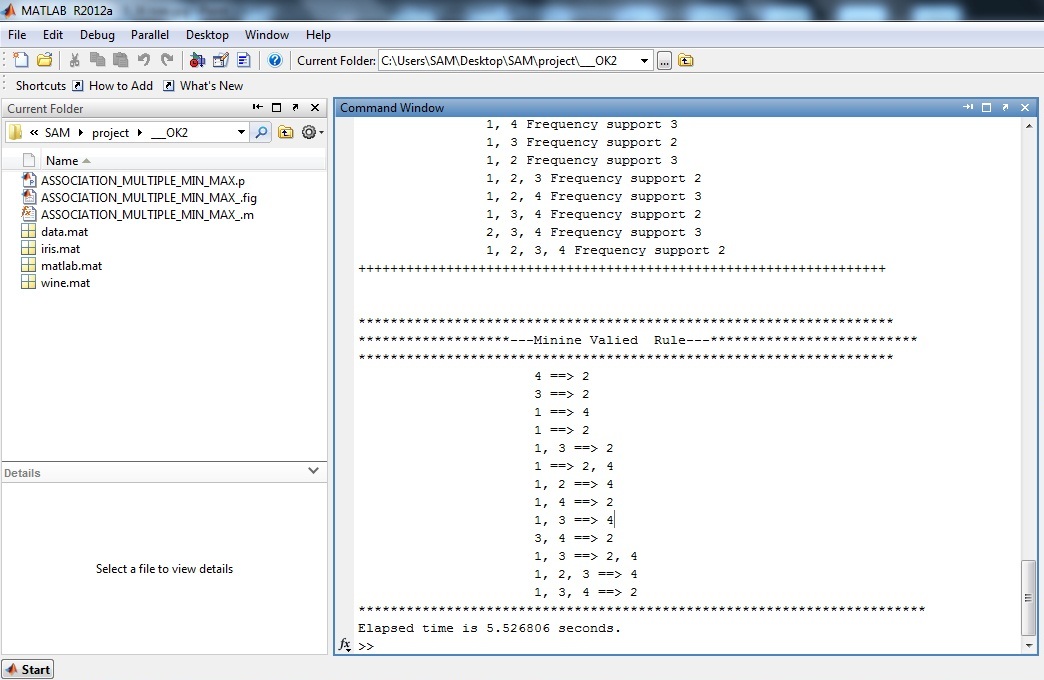
**Screenshot 4.5: Rule generation by Apriory method(1)**

Screenshots shows the total rules generated are 36 with elapsed time 0.081 using Tree Method.



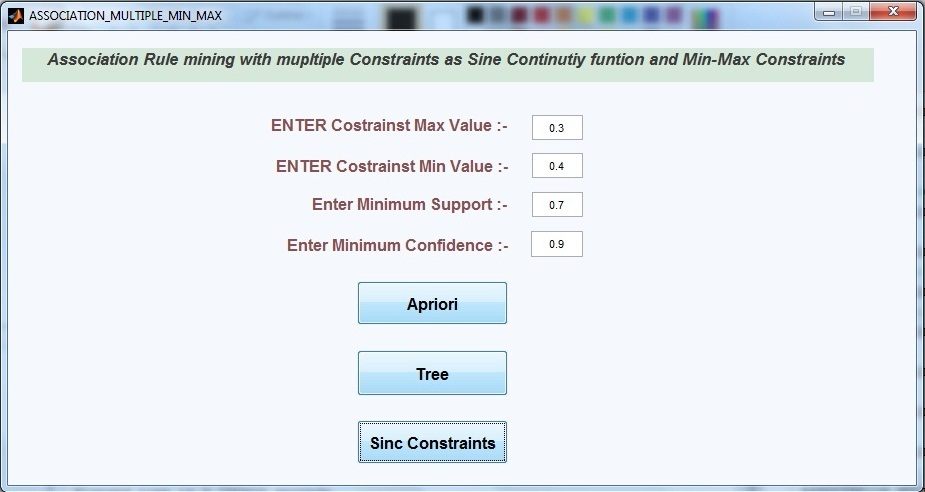
**Screenshot 4.6: Rule generation by Tree method(1)**

Screenshots shows the total rules generated are 13 with elapsed time 5.52 using Sin-Cosine Algorithm.



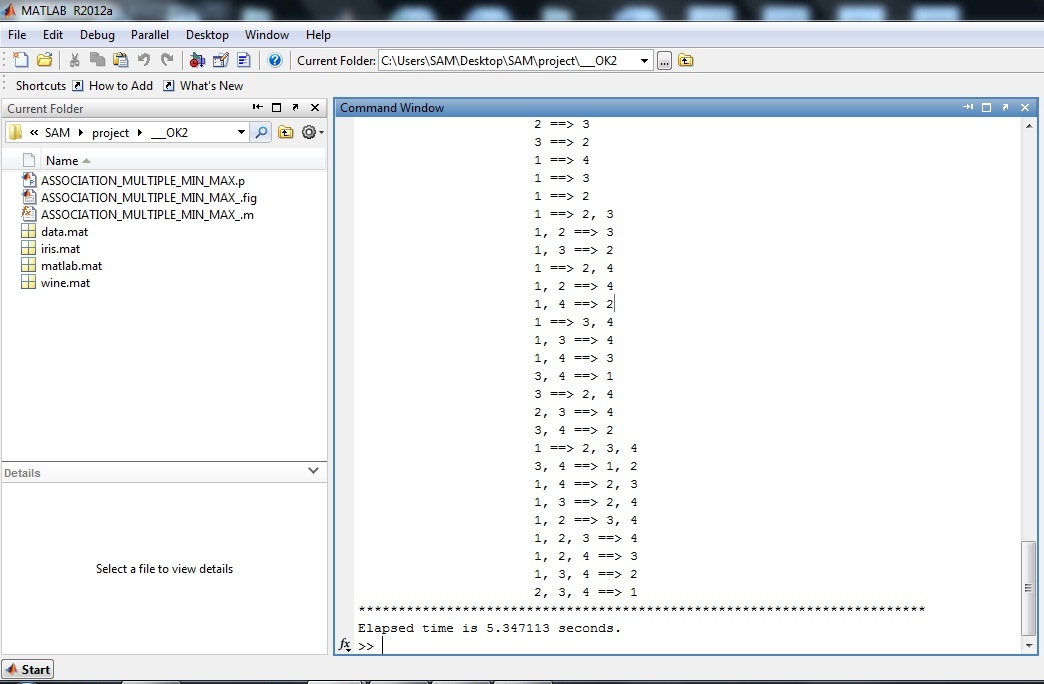
**Screenshot 4.7: Rule generation by Sin Cosine method(1)**

For the second experimental result we have change the values as Max Constraints = 0.4, Min Constraints = 0.3, Minimum Support = 0.7, & Minimum Confidence = 0.9



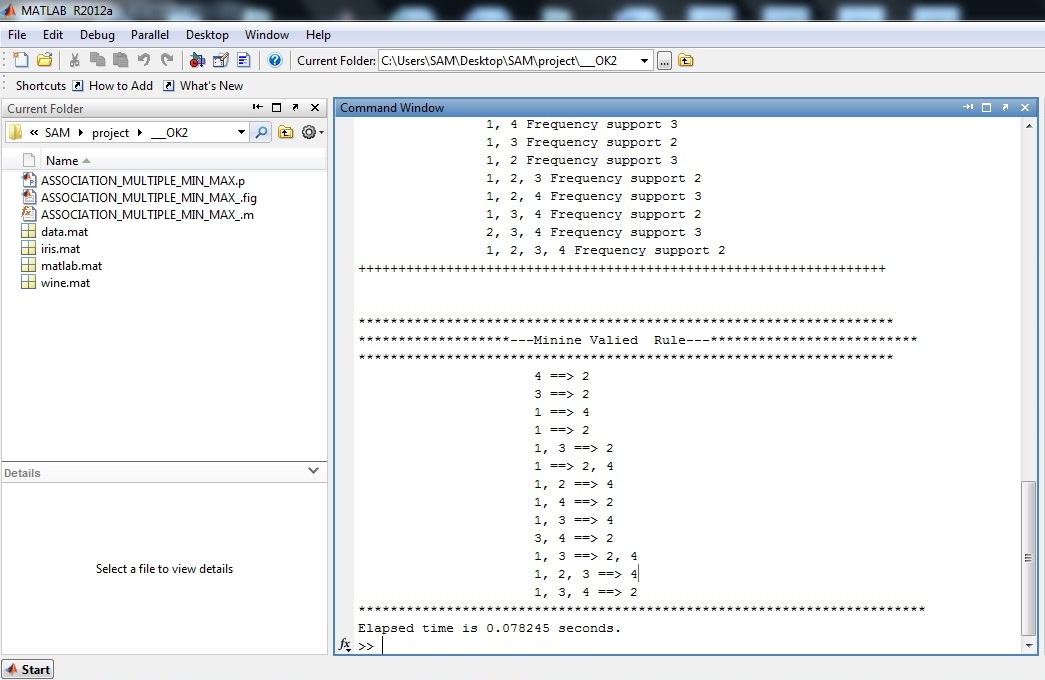
**Screenshot 4.8: The Main Window With 4 parameters(2)**

Screenshots shows the total rules generated are 45 with elapsed time 5.34 using Apriory Algorithm.



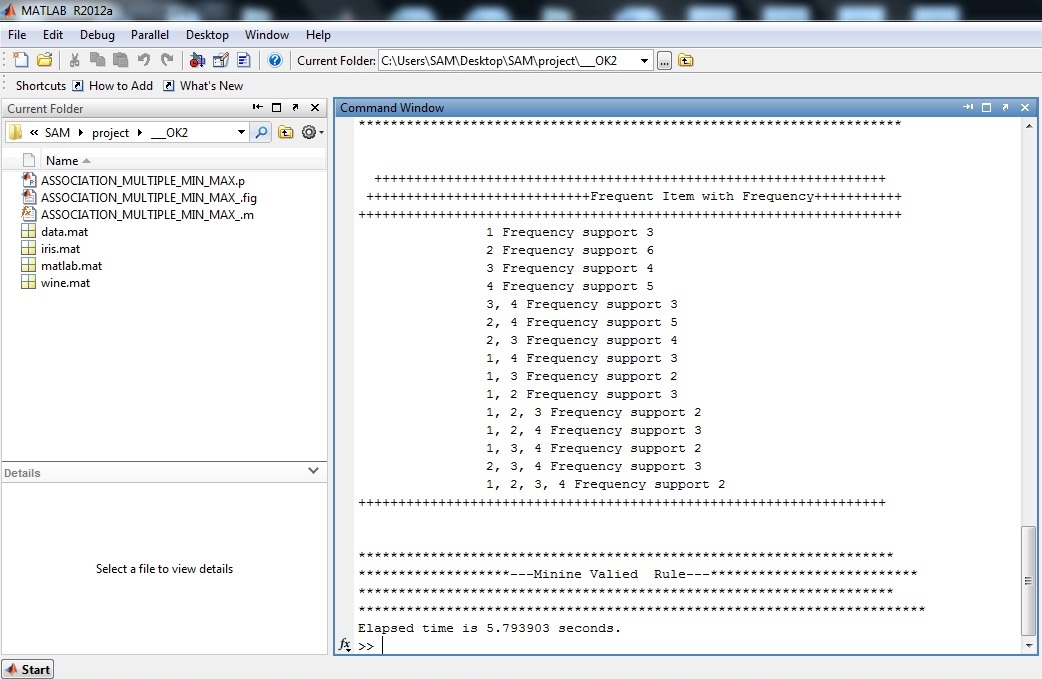
**Screenshot 4.9: Rule generated by Apriory method(2)**

Screenshots shows the total rules generated are 13 with elapsed time 0.078 using Tree method.

****

**Screenshot 4.10: Rule generated by Tree method (2)**

Screenshots shows the total rules generated are 0 with elapsed time 5.79 using Sin Cosine method.



**Screenshot 4.11: Rule generated by Sin Cosine method (2)**

* 1. **Comparative Result Analysis**
     1. **Comparison Table**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Method Name** | **Min value** | **Max value** | **Minimum support** | **Minimum Confidence** | **No. of Rule generation** | **Elapsed time** |
| Apriori | 0.3 | 0.4 | 0.5 | 0.6 | 48 | 6.458 |
| Tree | 0.3 | 0.4 | 0.5 | 0.6 | 36 | 0.081 |
| Sin Cosine  (MLMS-GA) | 0.3 | 0.4 | 0.5 | 0.6 | 13 | 5.52 |

**Table 4.4: Shows that the Comparative result analysis of different methods and the number of generated rule are also different, for our proposed experimental method.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Method Name** | **Min value** | **Max value** | **Minimum support** | **Minimum Confidence** | **No. of Rule generation** | **Elapsed time** |
| Apriori | 0.3 | 0.4 | 0.7 | 0.9 | 45 | 5.345 |
| Tree | 0.3 | 0.4 | 0.7 | 0.9 | 13 | 0.078 |
| Sin Cosine  (MLMS-GA) | 0.3 | 0.4 | 0.7 | 0.9 | 0 | 5.793 |

**Table 4.5: Shows that the Comparative result analysis of different methods and the number of generated rule are also different, for our proposed experimental method for different value.**

* + 1. **Comparison Graph**

**Figure 4.1: Comparative result for different methods and shows that the proposed method generated the number of rule are less than other methods.(1)**

**Figure 4.2: Comparative result for different methods and shows that our proposed method generated the number of rule are less than other methods.(2)**

1. **Conclusions and Future Scope**

**Conclusions:**

The proposed algorithm is combination of min-max function and genetic algorithm. The min-max function and genetic algorithm work together and perform condition based rule generation process. The min-max condition based function operates in sine and cosine based trigonometric function for the processing of genetic fitness function.

This has been observed that when condition is modified the new rules in large quantity are found. This implies that when min-max is solely determined through support and confidence, there is a high chance of eliminating interesting rules. With more rules emerging it implies there should be a mechanism for managing their large numbers. The large generated rule is optimized with genetic algorithm.

Theoretically it is proved that a relation between locally large and globally large patterns is used for local pruning at each site to reduce the searched candidates. A locally large threshold has been derived using a globally set minimum recall threshold.

It has been observed that when modify the scan process of transaction, generation of rule is become fast. With more rules emerging it implies there should be a mechanism for managing their large numbers. The large generated rule is optimized with min-max algorithm.

The proposed algorithm is performed better optimization in comparison of Apriori, Tree and monotonic condition based association rule mining.

* 1. **Future Scope:**

Condition based association rule mining is great advantage over conventional rule generation technique. The conventional rule generation technique used some standard algorithm. The proposed algorithm is very promising in the field of association rule mining. The proposed algorithm has multiple constraints such as genetic algorithm and sine and cosine function. The value of sine and cosine increases the process of algorithm work as normal Apriori and tree based technique.

Due to the value range limitation of sine cosine i.e., 0 to 1, if the constraint value is change in data mining with change in the method then the large range of the input value in the algorithm is possible.

**References**

[1] Gavin Shaw, Yue Xu, Shlomo Geva “Interestingness Measures for Multi-Level Association Rules” Proceedings of the 14th Australasian Document Computing Symposium, 2009. Pp 2-9.

[2] Xiufeng Piao, Zhanlong Wang, Gang Liu “Research on Mining Positive and Negative Association Rules Based on Dual Confidence” IEEE Fifth International Conference on Internet Computing for Science and Engineering, 2010, Pp 102-105.

[3] Li Guang-yuan, Cao Dan yang, Guo Jian-wei “Association Rules Mining with Multiple Constraints” Elsevier ltd. 2011, Pp 1678-1683.

[4] Xiaobing Liu, Kun Zhai, Witold Pedrycz “An improved association rules mining method” Expert Systems with Applications 2012, Pp 1362–1374.

[5] Ying-Ho Liu “Mining frequent patterns from univariate uncertain data” Data & Knowledge Engineering, 2012. Pp 47-68.

[6] Idheb Mohamad Ali O. Swesi, Azuraliza Abu Bakar, Anis Suhailis Abdul Kadir “Mining Positive and Negative Association Rules from Interesting Frequent and Infrequent Item sets” IEEE 9th International Conference on Fuzzy Systems and Knowledge Discovery, 2013. Pp 650-655.

[7] Huang QingLan, Duan LongZhen “Multi-level association rule mining based on clustering partition” IEEE Third International Conference on Intelligent System Design and Engineering Applications, 2013. Pp 982-986.

[8] Vidhu Singhal, Gopal Pandey “A Web Based Recommendation Using Association Rule and Clustering”, International Journal of Computer & Communication Engineering Research, 2013. Pp 1-5.

[9] Noha Negm , Passent Elkafrawy , Mohamed Amin, Abdel Badeeh M. Salem “Investigate the Performance of Document Clustering Approach Based on Association Rules Mining” International Journal of Advanced Computer Science and Applications, vol-4, 2013. Pp 142-151.

[10] Huang QingLan, Duan LongZhen “Multi-level association rule mining based on clustering partition” Third International Conference on Intelligent System Design and Engineering Applications, IEEE 2013. Pp 982-985.

[11] Aritra Roy, Rajdeep Chatterjee “Introducing New Hybrid Rough Fuzzy Association Rule Mining Algorithm” Proc. of Int. Conf. on Recent Trends in Information, Telecommunication and Computing, ITC, 2014. Pp 167-175.

[12] Deepak A. Vidhate, Dr. Parag Kulkarni “improvement in association rule mining by multilevel relationship algorithm” International Journal of Research in Advent Technology, Vol-2, 2014. Pp 366-373.

[13] Yin-Fu Huang, San-Des Lin “Applying Multidimensional Association Rule Mining to Feedback-based Recommendation Systems” International Conference on Advances in Social Networks Analysis and Mining, 2011. Pp 412-417.

[14] XING Xue, CHEN Yao, WANG Yan-en “Study on Mining Theories of Association Rules and Its Application” IEEE, 2010. Pp 2-10.

[15] Yingqin Gu, Hongyan Liu, Jun He, Bo Hu and Xiaoyong Du “A Multi-relational Classification Algorithm based on Association Rules” IEEE, 2009. Pp 235-242.

[16] W. Li, J. Han, and J. Pei “CMAR: Accurate and efficient Classification Based on Multiple Class-Association Rules” Proceedings of the ICDM, IEEE Computer Society, San Jose California, 2001, Pp 369-376.

[17] X. Yin, and J. Han “CPAR: Classification based on Predictive Association Rules” Proceedings of the SDM, SIAM, Francisco California, 2003.

[18] Zhen Peng, Lifeng and Wu Xiaoju Wang “Research on Multi-Relational Classification Approaches” IEEE, 2009. Pp 3-9.

[19] He J, Liu HY, Du XY. “Mining of multi-relational association rules” Journal of Software, 2007, Pp 2752-2765.

[20] Baralis, E., Cerquitelli, T, & Chiusano, S. “IMine: Index support for item set mining” IEEE Transactions on Knowledge and Data Engineering, 2009, Pp 493–506.

[21] Chen, L., & Wang, C. L. “Continuous sub-graph pattern search over certain and uncertain graph streams” IEEE Transactions on Knowledge and Data Engineering, 2010, Pp 1093–1109.

[22] Chen, S. M., & Bai, S. M. “Using data mining techniques to automatically construct concept maps for adaptive learning systems” Expert Systems with Applications, 2010, Pp 4496–4503.

[23] Chu, C. J., Tseng, V. S., & Liang, T. “Efficient mining of temporal emerging itemsets from data streams” Expert Systems with Applications, 2009 Pp 885–893.

[24] Ding, J. F., & Yau, S. S. T. “TCOM, an innovative data structure for mining association rules among infrequent items” Computers and Mathematics with Applications, 2009, Pp 290–301.

[25] Du, Y. J., & Li, H. M. “Strategy for mining association rules for web pages based on formal concept analysis” Applied Soft Computing, 2010 Pp 772–783.

[26] Eichhorn, A., Girimonte, D., Klose, A., & Kruse, R. “Soft computing for automated surface quality analysis of exterior car body panels” Applied Soft Computing, 2005 Pp 301–313.

[27] Gharib, T. F., Nassar, H., Taha, M., & Abraham, A. “An efficient algorithm for incremental mining of temporal association rules” Data and Knowledge Engineering, 2005, Pp 800–815.

[28] Kuo, R. J., Chao, C. M., & Chiu, Y. T. “Application of particle swarm optimization to association rule mining” Applied Soft Computing, 2011 Pp 326–336.

[29] Lee, A. J. T., Wang, C. S., Weng, W. Y., Chen, Y. A., & Wu, H. W. “An efficient algorithm for mining closed inter-transaction item sets” Data and Knowledge Engineering, 2008, Pp 68–91.

[30] C. K. Leung, D. A. Brajczuk “Efficient algorithms for mining constrained frequent patterns from uncertain data” Proc. ACM SIGKDD Workshop on Knowledge Discovery from Uncertain Data, 2009, Pp 9–18.

[31] C.K. Leung, D.A. Brajczuk “Mining uncertain data for constrained frequent sets” Proc. International Database Engineering & Applications Symposium, 2009, Pp 109–120.

[32] R. Agrawal, Tomasz Imielinski, Arun Swami "Mining association rules between sets of items in large databases". Proc. ACM SIGMOD International conference on Management of Data -SIGMOD '93.1993 p. 207

[33] C. C. Agrawal, Y. Li, J. Wang, J. Wang “Frequent pattern mining with uncertain data” Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009, Pp 29–37.

[34] Peter P. Wakabi–Waiswa, Venansius Baryamureeba “Optimized Association Rule Mining with Genetic Algorithms “ Seventh International Conference on Natural Computation, 2011,Pp 1116-1120.

[35] S. Zhang, Q. Chen, Q. Yang “Editorial: acquiring knowledge from inconsistent data sources through weighting” Data & Knowledge Engineering, 2010. Pp 779–799.

[36] L.A. Abd-Elmegid, M.E. El-Sharkawi, L.M. El-Fangary, Y.K. Helmy “Vertical mining of frequent patterns from uncertain data” Computer and Information Science, 2010. Pp 171–179.

[37] L. Sun, R. Cheng, D.W. Cheung, J. Cheng “Mining uncertain data with probabilistic guarantees” Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2010, Pp 273–282.