Association Rule Mining and Medical Application: A Detailed Survey

K. Pazhanikumar
Research Scholar
Department of Computer Science
S.T.Hindu College, Nagercoil
Affiliated to Manonmaniam Sundaranar University
Abishekapatti, Tirunelveli-627 012, Tamil Nadu, India

S. Arumugaperumal
Head
Dept of Computer Science
S.T. Hindu College

ABSTRACT
Association rule mining is one of the well established fields in data mining. This paper has surveyed the research papers in this field from 1993 to 2013. This paper gives detailed account of fundamental algorithms and its advantages and disadvantages. This also provides brief overview of current trends of association and frequent pattern mining and medical applications.

Keywords
Association Rules, Frequent pattern, Data Mining

1. INTRODUCTION
A decade of work in [1] Association Rule Mining (ARM) has become a mature field of research. So many research papers, articles are surveyed in the field of ARM. This paper details some fundamental about frequent itemset generation which helps to develop new algorithm for that process. The field of ARM is divided into the following areas: Positive rule mining, Negative rule mining and Interestingness measures. Major area of work in ARM is coming under these three categories. The classical rules are called positive rules which are showed in the section 3.1. The positive rules are mined from a set of frequent itemsets. Due to the deficiency of frequent itemset mining, the frequent itemsets are extended to various formats like closed, maximum, sequential, complex frequent itemset. The frequent itemset mining is detailed in the section 4. The above types of frequent itemset are supported to constraint based rule mining. The constraint based rule mining is described in the section 3.3. The negative relationships between itemset are mined by rule mining process using infrequent itemset. The rules are mined from these kinds of infrequent itemset that is called negative rule mining which is explained in section 3.2. The interestingness measures play an essential role in the field of ARM similar to the data mining process. These measures are discussed in section 7.

2. ASSOCIATION RULE
Initially it was largely motivated to understand the market basket data, the results of which allowed companies to understand purchasing behavior and, as a result, better target market audiences. ARM is user centric as the objective is the elicitation of interesting rules from which new knowledge can be derived. ARM is to facilitate the discovery, heuristically filter, and enable the presentation of these inferences or rules for subsequent interpretation by the user to determine their usefulness. ARM has been divided into two phase of process as follows:

Phase 1: Identify the sets of frequent items or itemsets or pattern within the set of transaction using user-specified support threshold.

Phase 2: Generate inferences or rules from these above patterns using user-specified confidence threshold.

The above two phases are generated strong association rules from dataset. The first phase is called frequent itemset construction or mining. That is extremely computational expensive than phase 2. The second phase is called association rule generation. That is, straight forward process. This phase computational complexity is negotiable to compare with first phase. There are two major problems in second phase. The first problem is rule quantity means that algorithms can produce large number of rules. The second problem is rule quality means that, all the rules are not interesting. The support and confidence measures play a vital role to filter unwanted itemsets and rules from the mining process. These measures are discussed in the section 7.3.

3. TYPES OF ASSOCIATION MINING
3.1 Positive Association Rule Mining
The classical association rules consider only items enumerated in transactions of the dataset. The positive relationship can be found between the set of items. The rules are generated from the positive related items. These rules are referred to as positive association rules. Most of the algorithms were developed for generating positive associations between items. These are useful to decision making [69].

The positive rules are classified as follows:
1. Boolean association rule
    a. Quantitative [62]
    b. Constrained rules [44]
    c. Sequential rules [5]
2. Qualitative association rule [62]
3. Spatial association rule
4. Temporal association rule

3.2 Negative Association Rule Mining
Negative association rules also consider the same items, but in addition the item also considers which were absent from transactions. The negative rules are generated from infrequent itemsets. These rules play some important role in decision-making [69]. These are useful in market basket analysis to identify products that conflict with each other or products that complement each other. This is a difficult task, due to the fact

that there are essential differences between positive and negative rule mining.

Brin et al [13] mentioned for the first time in the literature the notion of negative relationships. The authors have used statistical chi-square test to verify the independence between two variables. The authors have also used correlation measure to determine the nature (positive or negative) of the relationship. The strong negative rules are mined by Savasere et al [55]. They combined positive frequent itemsets with domain knowledge in the form of taxonomy.

3.3 Constraint based Association Rule Mining

The constraints were applied during the mining process to generate only those association rules that are interesting to users instead of all the rules. By doing this lots of cost of mining those rules that turned out to be not interesting can be saved. Usually constraints are provided by users. The constraints are classified as follows:

1. Knowledge based constraints [41]
2. Data constraints [10]

4. FREQUENT PATTERN MINING

Patterns are set of item, sequences, graph or structures that appear in a dataset. The frequency of pattern is no less than a user-specified threshold that is called frequent pattern or itemset. Finding frequent patterns plays a fundamental role in association rule mining, classification, clustering, and other data mining tasks. Frequent pattern mining was first proposed by Agarwal et al [1] for market basket analysis in the form of association rule mining. The fundamental frequent pattern algorithms are classified into three ways as follows:

1. Candidate generation approach (E.g. Apriori algorithm)
2. Without candidate generation approach (E.g. FP-growth algorithm)
3. Vertical layout approach (E.g. Eclat algorithm)

4.1 Candidate Generation Approach

4.1.1 Apriori Algorithm

The first frequent itemset mining algorithm was denoted as AIS [1]. Later, the algorithm was improved and called Apriori. The main improvement has developed the monotonicity property of the support of sets [4]. After the improvement, the monotonicity further got better by Mannila et al [39] and Agarwal et al [2]. The Apriori algorithm is based on candidate generation approach. The Apriori algorithm is implemented with various data structures in more detail [12].

4.1.2 Extension of Apriori

Since the Apriori algorithm was proposed, there have been extensive studies on the improvements or extensions of Apriori. The extended algorithms are classified into following nine ways:

4. Transaction reduction [35] and mapping technique [60]
5. Hashing technique [46]
6. Partitioning technique [54]
7. Sampling approach [66]
8. Incremental mining [16]
9. Parallel and distributed mining [3]
10. Integrating mining with relational database systems [53]
11. Level-wise mining approach [23]

4.1.3 Advantages and Disadvantages of Candidate Generation Approach

Advantages

1. It significantly reduces the size of candidate sets using the Apriori principle.
2. It uses large itemset property.
3. It is easily parallelized.
4. It is easy to implement with all kind of real datasets.

Disadvantages

1. It generates huge number of candidate sets.
2. When the longest frequent itemsets is *k*, Apriori needs *k* passes of database scans. So it will have low efficiency.
3. Repeatedly scanning the database and checking the candidates by pattern matching.
4. The computation time is very intensive at generating the candidate itemsets and computing the support values for application with very low support and vast amount of items.

4.2 Without Candidate Generation Approach

4.2.1 FP-Growth Algorithm

Han et al [26] devised an FP-growth method that mines the complete set of frequent itemsets without candidate generation. It employed in a divide-and-conquer manner. In first scan, the database derives a list of frequent items in which items are ordered by frequency descending order. The database is compressed into a frequent pattern tree (FP-tree) using frequency descending order list. The FP-tree is mined by starting from each frequent length-1 pattern, constructing its conditional pattern base, then constructing its conditional FP-tree, and performing mining recursively on such a tree. The pattern growth is achieved by the concatenation of the suffix pattern with the frequent patterns generated from a conditional FP-tree. The FP-growth algorithm transformed the problem of finding long frequent patterns to searching for shorter ones recursively and then concatenating the suffix. It uses the least frequent items as a suffix, offering good selectivity. The performance studies of FP-Growth exhibit that the method significantly reduces search time.

4.2.2 Extended Algorithms

There are many alternatives and extensions to the FP-growth approach, including

1. Depth first generation of frequent itemsets [6].
2. H-Mine (Hyper-structure Mining) algorithm [49].
3. Building alternative trees [38].

4.2.3 Advantages and Disadvantages of Without Candidate Generation Approach

Advantages

1. It does not break a long pattern of transaction.
2. It conserves complete information for frequent pattern mining.
3. It reduces irrelevant information or infrequent items are gone.

4. The frequency descending ordering is more likely to be shared.
5. It does not make transaction set larger than the original database.
6. It is much faster than Apriori algorithm.

Disadvantages

1. Frequent pattern tree may not fit in memory.
2. Frequent pattern tree is expensive to build. The time takes to build, but once it is built, frequent itemsets are read of easily.
3. If support is high, time is wasted, as the only pruning that can be done is on single items.
4. The support can only be calculated once the entire dataset is added to the FP-Tree.

4.3 Vertical Layout Approach

4.3.1 Eclat Algorithm

The first algorithm developed to generate all frequent itemsets in a depth-first manner is the Eclat (Equivalence CLAss Transformation) algorithm [73]. If the database is stored in the vertical layout, the counting of support can be much easier by simply intersecting the covers of two of its subsets that together give the set itself. The Eclat algorithm essentially used this technique inside the Apriori algorithm. Always this is not possible since the total size of all covers at a certain iteration of the local set generation procedure could exceed main memory limits. It is usually more efficient to first find the frequent items and frequent 2-sets separately and use the Eclat algorithm only for all larger sets [73].

4.3.2 Extended Algorithms

1. Diffset with Eclat (dEclat) algorithm [75].
2. J. Hipp, U.Guntzer, and G. Nakhaeizadeh combine Apriori and Eclat into a single hybrid algorithm [27].
3. Vertical Itemset Partitioning for Efficient Rule Extraction (VIPER) algorithm [58].

4.3.3 Advantages and Disadvantages of Vertical Layout Approach

Advantages

1. There is no need to scan the database to find the support of k+1 itemsets. This is because the *TID* set of each k-itemset carries the complete information required for counting such support.
2. It is possible to significantly reduce this total size by generating collections of candidate itemsets in a depth-first strategy.
3. This is not always possible since the total size of all covers at a certain iteration of the local itemset generation procedure could exceed main memory limits.

Disadvantages

1. It fails to manage main memory at time of high candidate itemsets.
2. The merge routine contains a large amount of conditional branches, which are extremely badly predictable.

5. EXTENSION OF FREQUENT PATTERN

Frequent pattern mining is challenging and essential task in data mining. So many times, this task resists with the following problems. It generates huge number of patterns satisfying user-specified threshold as low from a large dataset. Many scientific and commercial applications have more complicated items in their dataset that are difficult to mine frequent itemsets. The existing frequent pattern mining is not enough to mine interestingness and useful pattern from all kind of datasets. And also with the increase in use and development of data mining techniques and tools, much work has recently focused on finding the alternate patterns.

Those are including the following patterns:

1. Closed frequent pattern
2. Maximal frequent pattern
3. Sequential pattern
4. Complex pattern
5. Structural pattern
6. Infrequent pattern
7. Surprising pattern

5.1 Closed Frequent Pattern

Mining frequent pattern often generates a huge number of patterns satisfying the minimum support threshold which is set as low. If a pattern is frequent, each of its sub patterns is frequent as well. A large pattern will contain an exponential number of smaller, frequent sub-patterns. The closed frequent pattern mining overcame this problem. A pattern X is a closed frequent pattern [47] in a dataset D if X is frequent in D and there exists no proper super-pattern x such that x has the same support as X in D. For the same minimum support threshold, the set of closed frequent patterns contain the complete information regarding to its corresponding frequent patterns; whereas the set of max-patterns, though more compact, usually does not contain the complete support information regarding to its corresponding frequent patterns. The closed pattern algorithms are as follows:

1. Apriori based Closed itemset mining (A-Close) algorithm [47]
2. CLOsed item SET mining (CLOSET) algorithm [48]
3. CLOSET+ [68]

5.2 Maximal Frequent Pattern

A pattern X is a maximal frequent pattern [9] in set D if X is frequent, and there exists no super-pattern x such that $X \subset x$ and x is frequent in D. The maximal frequent pattern mining algorithms are given below:

1. Max-Miner algorithm [9]
2. MAximal Frequent Itemset Algorithm (MAFIA) [14]

5.3 Sequential Pattern

A sequence database consists of ordered elements or events, recorded with or without a concrete notion of time such as customer shopping sequences, web click streams, and biological sequences. Sequential pattern mining, the mining of frequently occurring ordered events or subsequences as patterns, was first introduced by Agarwal et al [5]. Some algorithms are as shown below.

1. Generalized Sequential Patterns (GSP) [63]

2. Indexing sequences by sequential pattern analysis (SeqIndex) algorithm [20]

3. Sequential PAttern Discovery using Equivalent Class (SPADE) algorithm [74]

5.4 Complex Pattern

Due to the large volume of complex objects such as transaction sequence, event logs, proteins and images, it is inefficient to perform a sequential scan on the whole database and examine objects one by one. High performance indexing mechanisms thus are, in heavy demand in filtering objects that obviously violate the query requirement. The algorithms are as follows:

1. Graph indexing based pattern mining (gIndex) algorithm [71]

2. Partition-based Graph Index and Search (PIS) algorithm [72]

5.5 Structural Pattern

Frequent substructures are the very basic patterns that can be discovered in a collection of graphs. Recent studies have developed several frequent substructure mining methods. Some structural pattern mining algorithms are as follows.

1. Apriori based Graph Mining (AGM) algorithm [29]

2. Frequent Sub Graph mining (FSG) algorithm [32]

3. MoFa [11]

5.6 Infrequent Pattern

Infrequent pattern [13] defines an itemset (small itemset) that does not meet the user-specified minimum support. The negation of an itemset A is indicated by ¬A. The support of ¬A, Support (¬A) = 1 – Support (A). In particular, for an itemset i_1, $\neg i_2$, i_3, its support is Support (i_1, $\neg i_2$, i_3) = Support (i_1, i_3) – Support (i_1, i_2, i_3). The forms ($A \Rightarrow \neg B$, $\neg A \Rightarrow B$ and $\neg A \Rightarrow \neg B$) are called negative rules.

5.7 Surprising Pattern

The unexpected patterns and exceptional patterns are called surprising patterns. It produces exception rules. An exception is defined as a deviational pattern to a well known fact, and exhibits unexpectedness. For example, while "bird(x) ⇒ flies(x)" is a well known fact, an exceptional rule is "bird(x), penguin(x) ⇒ ¬ flies(x)". This exception indicates that unexpected patterns and exceptional patterns can involve negative terms and therefore can be treated as a special case of negative rules.

Some of the existing works are given below:

1. Unexpected patterns [41]

2. Exceptional patterns [5]

6. RULE GENERATION ALGORITHMS

Most researches have said that the rule generation procedure is straight forward. Initially this process was implemented by Agarwal et al [1]. The procedure is to generate association rules from those large itemsets with the constraints of minimal confidence. Suppose one of the large itemsets is l_k, $l_k = \{i_1, i_2, \ldots, i_k\}$, association rules with this itemsets are generated in the following way: the first rule is $\{i_1, i_2, \ldots, i_{k-1}\} \Rightarrow \{i_k\}$, by checking the confidence this rule can be determined as interesting or not. Then other rules are generated by deleting the last items in the antecedent and inserting it to the consequent, further the confidences of the new rules are checked to determine the interestingness of them. Those processes iterated until the antecedent becomes empty.

Agarwal et al [4] improved the above procedure by generating subsets of a large itemset in recursive depth first strategy. If a subset x of a large itemset l_k does not generate a rule, the subsets of x need not to be considered for generating rules using l_k. For example $l_k=\{i_1, i_2, i_3, i_4\}$, consider the subset $\{i_1,i_2,i_3\}$, $\{i_1,i_2\}$. If $\{i_1, i_2, i_3\} \Rightarrow \{i_4\}$ does not satisfy user specified confidence, $\{i_1, i_2\} \Rightarrow \{i_3, i_4\}$ need not to check a rule or not. This algorithm is shown in section 5.4.2. The researchers have added and improved some other techniques with the above rule generation procedure.

6.1. Using user-defined templates or item constraints

Elena baralis and Giuseppe psaila [7] developed a general framework for the design of association rule extraction applications from dataset. It has identified several classes of relevant association rules that can be extracted from the dataset. The classes are allowed to pinpoint several extraction criteria which are used to research relevant rules. Based on the above criteria, this framework defined a template language that allows the specification of a predefined format for different extraction conditions, in which only the target database and attributes must be instantiated. Hence, association rule templates provide a simplified interface for defining rule extraction criteria. Templates can be used by inexpert users to extract interesting rules with a predefined structure

6.2. Using interestingness measures

There are three types of measures used to mine interesting rules. Those are subjective, objective and semantic measures that are explained in section 2.7. In the literature, there are thirty two measure used in data mining to mine interestingness rules [28].

PangNing Tan, Vipin Kumar and Jaideep Srivastava [65] described several key properties one should examine in order to select the right measure for a given application domain. These properties are used with twenty one of the existing measures. It showed that each measure has different properties which make them useful for some application domains, but not for others. The authors also presented two scenarios in which most of the existing measures agree with each other, namely, support-based pruning and table standardization. The authors presented an algorithm to select a small set of tables such that an expert can select a desirable measure by looking at just this small set of tables.

6.3. Using inference systems to prune redundant rules

Yves bastide, Nicolas pasquier, Rafik Touil, Gerd Stumme and Lotif Lakhal [8] defined new bases for association rule which union is a generating set for all valid association rules with support and confidence. These bases are characterized using frequent closed itemsets and their generators. It carried non redundant exact and approximated rules having minimal antecedents and maximal consequents. This new basis is much suitable to real life databases.

6.4. Using new framework with different formats or properties

Bart Goethals, Juho Muhonen and Hannu Toivonen [24] presented a technique for computing upper and lower bounds of the confidence of an association rule. When the upper and

lower bounds are equal or almost equal, the association rules are called derivable. It considered being redundant with respect to its sub rules. This technique is based on the inclusion–exclusion principle. The method is simple which gives absolute bounds, and it does not assume any specific inference system. The bounds and derivability follow from the definitions of support and confidence: when a rule is pruned as exactly derivable, then there exists only one value for the confidence that is consistent with all the sub rules. They do not actually have all sub rules of an association rule as some of them might not be confident. They never used the confidence threshold for pruning.

Brin. S, Motwani. R, Ullman, J.D and Tsur. D presented a new algorithm for frequent itemset mining based on sampling. By using item reordering, it improves the efficiency of the algorithm. They also presented an approach to generate rules which are normalized based on both the antecedent and consequent. Suzuki et al presented frameworks for exception based rule mining algorithms [64].

7. MEASURES

ARM can be viewed as an algorithmic process that takes data as input and discover patterns. Interestingness measures play an essential role, reducing the number of discovered rules and retaining only the best ones, in a post-processing step. The nine specific criteria are used to determine whether a pattern is interesting or not. The criteria are conciseness, Generality / Coverage, Reliability, Peculiarity, Diversity, Novelty, Surprisingness, Utility, and Actionability/Applicability.

7.1 Types of Measures

Based on the above nine criteria, the measures are classified as the following three types:

7.1.1 Subjective Measures

A subjective measure takes into account both the data and user of the dataset. A subjective measure is to access to the user's domain or background knowledge about the data required. Subjective techniques generally operate by comparing the user's beliefs against the patterns discovered by the algorithm [59]. Novelty and surprisingness depend on the user of the patterns, as well as the data and patterns themselves, and hence can be considered subjective. The main disadvantage of the subjective or user driven approach is that it constrains the discovery process to seek only what the user can anticipate or hypothesize i.e. it cannot discover unexpected or unforeseen patterns because it is entirely goal driven [31].

7.1.2 Objective Measures

An objective measure is based only on the raw data. The users do not require about any knowledge regarding application or domain. Most objective measures are based on theories in probability, statistics, or information theory. Conciseness, generality, reliability, peculiarity, and diversity depend only on the data and patterns, and thus can be considered objective. Objective measure or data driven measures tend to concentrate on finding patterns through statistical strength or correlations [31].

7.1.3 Semantic Measures

A semantic measure considers the semantics and explanations of the patterns. Utility and actionability depend on the semantics of the data, and thus can be considered semantic.

7.2 Role of Measures

The measures are used in the following three ways:

1. They helped to classify each pattern as either interesting or uninteresting.
2. The measures are used to determine that one pattern is more interesting than another.
3. Also the measure helps to rank the interesting or useful patterns.

7.3 List of Measures

7.3.1 Support

It is a basic measure related to probability and set theory. It is defined as the fraction of transactions in the database which contain all items in a specific rule [1]. This can be written as:

$$Support(x \Rightarrow y) = Support(x \cup y) = \frac{|xy|}{|D|}$$

Where $|xy|$ is the number transactions (itemset) which contain both x and y and $|D|$ represents the total number of transactions (itemset) in the database. It is usually specified in generating the association rules which select only the most frequent items in the database.

7.3.2 Confidence

Another measure of the association rules is confidence [1]. This is the strength of the implication of a rule and can be represented as a ratio between the transaction numbers, including x and y and those including X, and X means that $x \cup y$.

$$Confidence(x \Rightarrow y) = \frac{Support(x \Rightarrow y)}{Support(x)} = \frac{|xy|}{|x|}$$

Where $|x|$ is the number of transactions (itemset) containing X. It is specified to generate association rules.

7.3.3 Representativity

It is used to obtain good sample of the dataset. It is introduced by Ragel et al [50]. It is needed to influence itemsets that all transactions in Dataset D have missing values for some of the attributes on confidence and support.

$$representativity(X) = \frac{|D| - |Disabled(X)|}{|D|}$$

Disabled (X) : Transaction t is disabled for X in D, if t contains missing values for at least one item i of X.

7.3.4 Other Measures

The survey [37] analyzed the thirty eight interestingness measures for association rules, classification and summaries. Another study [34] reviewed twenty interestingness measures by using 10 datasets. The authors were compared to an analysis of formal properties of the measures which make a best choice of user's needs. The reviews [40] discussed seventeen interestingness objective measures for association rule mining.

8. APPLICATIONS

ARM applications have since been applied to many different domains including market basket and risk analysis in commercial environments, epidemiology, clinical medicine, fluid dynamics, astrophysics, crime prevention, and counter-terrorism - all areas in which the relationship between objects can provide useful knowledge.

8.1. Biological and Medical

Ansaf Salleb, Teddy Turmeaux, Christel Vrain and Cyril Nortet [52] have developed a new tool QuantMiner, genetic-based algorithm software for mining quantitative association rules on atherosclerosis dataset. The authors have mined quantitative rules from the atherosclerosis dataset. These rules are had nice feature to handle both categorical and numeric attributes. QuantMiner is an interesting tool for mining descriptive rules of medical and other datasets.

Carlos Ordonez, Norberto Ezquerra and Cesar A. Santana [42] have used association rule in high dimensional medical domain. The authors applied greedy algorithm to compute rule covers in order to summarize rules having the same consequent. The significance of association rules is evaluated using through support, confidence and lift. They are focused association rules on a real dataset to predict absence or existence of heart disease. The constraints are reduced the number of discovered rules and improved running time. Rule covers summarized a large number of rules by producing a brief set of rules with high-quality metrics.

Carlos Ordonez, Cesar A.Santana and Levien de Braal [43] explored the idea of discovering association rule in medical data. They are improved ARM algorithm which incorporated several important constraints. The constraints were incorporated to find relevant rules and avoid the redundant rules. They validated the mined rule using an expert system to aid in perfex heart disease diagnosis. The relevant rules were enriched the expert system knowledge.

Gasmi. G, T. Hamrouni1, S. Abdelhak, S. Ben Yahia1, and E. Mephu Nguifo [22] extracted association rule in sage dataset. They have to stress on the extraction of generic basis of association rules from the sage data generated, in different biological situations. Generic basis of association rules is a subset of all association rules, from which the remaining association rules are generated. They avoided the extraction of an overwhelming knowledge, which is of primary importance as it guaranties extra value knowledge usefulness and reliability. This is reinforced while handling highly dense sage data. They are, compared and assessed frequent closed itemset algorithm performances on sage data. Also they extracted the *IGB* generic basis of association rules. *IGB* are informative and more compact than other generic basis.

Hung-Wen Chiu and Fei-Hung Hung [19] have applied association rule mining in Protein-Protein Interaction (PPI) database. It mined the associations of functional regions of two interacting proteins to help PPI prediction. The data are collected from Database of Interaction Proteins (DIP) and Interaction Proteins (IntAct), and downloaded the information for functional regions of proteins from Uniprot. A web-based system was designed to integrate process and mine these data to create some rules, based on functional region association. PPIs of other species were used to evaluate these rules. In result, over 80% association rules were produced from yeast PPI data in other species. This indicated that the rules learning from known PPI provide good references for PPI prediction.

Nitin Gupta, Nitin Mangal, Kamal Tiwari, and Pabitra Mitra [25] applied quantitative association rule mining to decipher the nature of associations between different amino acids that are present in a protein. The association rules are enhanced their understanding of protein composition and hold the potential to give clues regarding the global interactions amongst some particular sets of amino acids occurring in proteins. It has discovered rules based not only on the presence of amino acids, but also on absence. This is the first systematic study to discover global associations between amino acids.

Parameshvyas Laxminarayan, Carolina Ruiz, Sergio A. Alvarez and Majaz Moonis [33] introduced an association rule mining technique for complex datasets described by both static and time-dependent attributes. They applied ARM technique to find associations among sleep questionnaire responses, clinical summary information, and all-night polysomno graphic recordings of sleeping human subjects. The Apriori algorithm designed to deal with time-varying sequences using time windows was developed and employed to uncover statistically significant and clinically meaningful associations among summary and polysomno graphic time series variables.

Qingfeng Chen and Yi-Ping Phoebe Chen [15] mined frequent patterns for Adenosine Mono Phosphate (AMP)-activated protein kinase regulation on skeletal muscle. They intended a framework that can identify the potential correlation, either between the state of isoforms of α, β and γ subunits of AMP-activated protein kinase or between stimulus factors and the state of isoforms. Their approach is applied item constraints in the closed interpretation to the itemset generation so that a threshold is specified in terms of the amount of results, rather than a fixed threshold value for all itemsets of all sizes. It is found that most of the extracted association rules have biological meaning and some of them were previously unknown.

Ronaldo Cristiano Prati, Maria Carolina Monard and André C.P.L.F. de Carvalho [51] presented a new approach to induce knowledge rules from HIV cleavage dataset. Its main characteristic is to incorporate exceptions into the representation used by machine learning algorithms. That approach has used the following two steps: induction of common sense rules and looking for exceptions. They wanted a real world dataset related to where a viral protease cleaves HIV viral poly protein amino acid residues. That approach is to find exceptions out of general rules, especially suitable for such analysis. It allows a more compact and easy to understand model description, helping the domain expert to understand the underlying process.

Sengul Dogan and Ibrahim Turkoglu [56] presented a new approach to find association rules an effective method for discovering Hyperlipidemia. The presented system projected from the biochemistry blood parameters which are very helpful to make everything easier for the physicians in the diagnosis of Hyperlipidemia. The basic characteristic of the lipide parameters that is total cholesterol, Low Density Lipoprotein (LDL), Triglyceride, High Density Lipoprotein (HDL) and Very Low Density Lipoprotein (VLDL) parameters are used in the process of entering the system and finally results evaluated at the end of this process. The results

of the decision support system have completely matched with those of the physicians' decisions.

Shantakumar B.Patil and Y.S.Kumaraswamy [57] are constructed an efficient approach for the extraction of significant patterns from the heart disease warehouses for heart attack prediction. They preprocessed dataset to make it appropriate for the mining process. After preprocessing, the heart disease dataset was clustered using the K-means clustering algorithm, which extracted the data relevant to heart attack from the warehouse. The frequent patterns were mined from the extracted data, relevant to heart disease, using the Maximum Frequent itemset algorithm (MAFIA). Then the significant weightage of the frequent patterns are calculated. The patterns significant to heart attack prediction are chosen based on the calculated significant weightage. These significant patterns are used in the development of heart attack prediction system.

Stephen. M, Downs, Michael. Y, Wallace [21] applied association rule mining algorithm with child health improvement program dataset. They are used a pattern discovery algorithm to extract second and third order association rules from the data. The algorithm discovered 16 second order associations and 103 third order associations. The third order associations contained no new information. The second order associations showed a covariance among a range of health risk behaviors. The algorithm discovered that both tobacco smoke exposure and chronic cardiopulmonary disease are associated with failure on developmental screens. These relationships have been described before and have been attributed to underlying poverty. The algorithm demonstrated the ability of association rule mining on sparse clinical data to discover clinically important associations.

Susan Jensen [30] used SPSS Clementine data mining tool with medical dataset. That dataset is related to patient information and medical exams connected with thrombosis attacks were analysed. The ability to predict the onset and successful diagnosis of thrombosis is a key to the intervention of the disease, and sequential patterns of symptoms and laboratory examinations may indicate a trending from a pre-thrombosis to active thrombosis condition. The predictive modelling, association rules and sequence detection were used to investigate these patterns.

9. SUMMARY

Initially this paper started the discussion with association rule and its sub problems. The problem of frequent itemset construction is the major and cost effective task of association analysis that is deeply discussed with their advantages and disadvantages. The Apriori algorithm is shown and detailed in this paper. The rule construction procedure is argued in effective manner. Measures play an important role in data mining and association analysis. This paper is also talked about usage of measures, types and some survey regarding measures in association analysis. The association analysis is applied in various domains. The association rule mining with medical domain is discussed with some existing works. Some of the other interesting domains are discussed. A deep literature review is also done about association rule mining, frequent itemset mining and its application.

9. REFERENCES

[1] Agarwal, R., Imielinski, T., Swami,A,N. 1993. "*Mining association rules between sets of items in large databases.*" Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, D.C., United States, May 26-28, pp.207-216.

[2] Agarwal, R., Mannila,H., Srikant,R., Toivonen,H., Verkamo,A. 1996. "*Fast discovery of association rules.*" In Advances in Knowledge Discovery and Data Mining, eds. Fayyad,U., Piatetsky Shapiro,G., Smyth,P., and Uthurusamy,R. MIT Press 1996, pp.307-328.

[3] Agarwal, R., Shafer,J.C. 1996. "*Parallel mining of association rules: design, implementation, and experience.*" IEEE Transaction Knowledge and Data Engineering, Vol.8, No.6, pp.962-969.

[4] Agarwal, R., Srikant, R. 1994. "*Fast algorithm for mining association rules.*" Proceedings of Twentieth International Conference Very Large Data Bases, Santiago, Chile, September 12-15, pp.487-499.

[5] Agarwal, R., Srikant,R. 1995. "*Mining sequential patterns.*" Proceedings of Eleventh International Conference on Data Engineering, Taipei, Taiwan, March 6-10, pp.3-14.

[6] Agarwal, R.C., Aggarwal, C. C., Prasad, V.V. 2001. "*A tree projection algorithm for generation of frequent itemsets.*" Journal of Parallel Distributed Computing, Vol.61, Issue.3, pp.350-371.

[7] Baralis, E. and Psaila, G. 1997. "*Designing Templates for Mining Association Rules.*" Journal of Intelligent Information Systems, Vol.9, Issue.1, pp.7-32.

[8] Bastide, Y., Pasquier, N., Taouil, R., Stumme, G., and Lakhal, L. 2000. "*Mining Minimal Non-redundant Association Rules Using Frequent Closed Itemsets.*" Proceedings of the First international Conference on Computational Logic, London, UK, July 24-28, pp.972-986.

[9] Bayardo, J.R. 1998. "*Efficiently mining long patterns from databases.*" Proceedings of the ACM SIGMOD International Conference on Management of Data, Seattle, Washington, United States, June 01 - 04, pp.85-93.

[10] Bayardo, J.R., Agarwal, R., Gunopulos, D. 1999. "*Constraint-based rule mining in large, dense databases.*" Data Mining and Knowledge Discovery Journal, Vol.4, Issue:2-3, pp.217-240.

[11] Borgelt, C and Berthold, M.R. 2002. "*Mining molecular fragments: finding relevant substructures of molecules.*" Proceedings of the International Conference on Data Mining, Maebashi, Japan, December 9-12, 2002, pp.211–218.

[12] Borgelt, C., Kruse, R. 2002. "*Induction of association rules: Apriori implementation.*" Proceedings of the Fifteenth Conference on Computational Statistics, Berlin, Germany, August 24-28, pp.395–400.

[13] Brin, S., Motwani, R., Silverstein, C. 1997. "*Beyond Market Baskets: Generalizing Association Rules to Correlations.*" Proceedings of the ACM SIGMOD International Conference on Management of Data, Tucson, Arizona, USA, May 13-15, pp.265-276.

[14] Burdick, D., Calimlim, M., Gehrke, J. 2001. "*MAFIA: a maximal frequent itemset algorithm for transactional databases.*" Proceedings of the seventeenth International

Conference on Data Engineering, Heidelberg, Germany, April 02-06, pp.443-452.

[15] Chen, Q., Chen, Y. 2006. "*Mined frequent patterns for AMP-activated protein kinase regulation on skeletal muscle*." BMC Bioinformatics, Vol.7, No.394, pp.1-14.

[16] Cheng, H., Yan, X., Han, J. 2004. "*IncSpan: Incremental mining of sequential patterns in large.*" Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery in Databases, Seattle, WA, August 22- 25, pp.527-532.

[17] Cheung D.W, Han, J., Ng, V.T., Wong C.Y. 1996. "*Maintenance of discovered association rules in large an incremental updating technique.*" Proceedings of the International Conference on Data Engineering, New Orleans, Louisiana, February 26 – March 1, pp.106-114.

[18] Cheung, D.W., Han, J., Ng, V.T., Fu, A.W., Fu, Y. 1996. "*A fast distributed algorithm for mining association rules.*" Proceedings of the Fourth International Conference on Parallel and Distributed Information Systems, Miami Beach, Florida, United States, December 18 - 20, pp.31-43.

[19] Chiu, H.W., Hung, F.H. 2008. "*Association Rule Mining from Yeast Protein Interaction to Assist Protein-Protein Interaction Prediction*" Biomedical Soft Computing and Human Sciences, Vol.13, No.1, pp.3-6.

[20] Chung, H., Yan, X., Han, J. 2005, "*SeqIndex: Indexing Sequences by Sequential Pattern Analysis*" Proceedings of the Fifth SIAM International Conference on Data Mining, Newport Beach, USA, April 21-23, pp. 601-605.

[21] Downs, S., Wallace, M. 2000. "*Mining Association Rules from a Pediatric Primary Care Decision Support System*." Proceeding of the 2000 Annual Symposium of American Medical Informatics Association, Los Angels, CA, USA, November 4-8, pp.200-204.

[22] Gasmi, G., Hamrouni, T., Abdelhak, S., Ben Yahia, S., Mephu Nguifo, E. 2005. *Extracting generic basis of association rules from SAGE data*, Proceedings of the Eighth International ECML/PKDD Workshop Discovery Challenge, Porto, Portugal, October 7, pp.1-6.

[23] Geerts, F., Goethals, B., Bussche, J. 2001. "*A tight upper bound on the number of candidate patterns.*" Proceedings of the International Conference on Data Mining, San Jose, California, November 29- December 2, pp.155-162.

[24] Goethals, B., Muhonen, J., Toivonen, H. 2005. "*Mining non-derivable association rules.*" Proceedings of the Second SIAM International Conference on Data Mining, Newport Beach, CA, April 21-23, pp.239-249.

[25] Gupta, N., Mangal, N., Tiwari, K., Mitra, P. 2006. "*Mining Quantitative Association Rules in Protein Sequences*." Data Mining, Lecture Notes on Artificial intelligence 3755, Springer-Verlag, Berlin, pp.273-281.

[26] Han, J., Pei, J., Yin, Y. 2000. "*Mining frequent patterns without candidate generation.*" Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Dallas, Texas, May 16-18, pp.1-12.

[27] Hipp, J., Guntzer, U., Nakhaeizadeh, G. "*Mining association rules: Deriving a superior algorithm by analyzing today's approaches.*" Proceedings of the Fourth European Conference on Principles of Data Mining and Knowledge Discovery, Lyon, France, September 13-16, pp.159-168.

[28] Hussain, F., Liu, H., Suzuki, E., and Lu, H. 2000. "*Exception rule mining with a relative interestingness measure.*" Proceedings of the Fourth Pacific Asia Conference on Knowledge Discovery and Data Mining, Kyoto, Japan, April 18-20, pp.86-97.

[29] Inokuchi, A., Washio, T., Motoda, H. 2000. "*An apriori-based algorithm for mining frequent substructures from graph data.*" Proceedings of the Fourth European Symposium on the Principle of Data Mining and Knowledge Discovery, Lyon, France, September 13-16, pp.13-23.

[30] Jensen, S. 2001. "*Mining Medical Data for Predictive and Sequential patterns.*" Proceedings of the Fifth European Conference on Principles and Practice of Knowledge Discovery in Databases, Freiburg, Germany, September 3-5, pp. 1-10.

[31] Ken McGarry. 2005. "*A survey of interestingness measures for knowledge discovery*." The Knowledge Engineering Review, Vol.20, No.1, pp.39-61.

[32] Kuramochi, M., Karypis, G. 2001. "*Frequent subgraph discovery.*" Proceedings of the First IEEE International Conference on Data Mining, San Jose, California, USA, November 29 - December 2, pp.313-320.

[33] Laxminarayan, P., Ruiz, C., Alvarez, S.A., Moonis, M. 2005. "*Mining Associations over Human Sleep Time Series.*" Proceedings of the eighteenth IEEE Symposium on Computer-Based Medical Systems, Dublin, Ireland, June 23-24, pp.323-328.

[34] Lenca, P., Vaillant, B., Meyer, P., Lallich, S. 2007. *Quality Measures in Data Mining*, chapter "*Association rule interestingness measures: experimental and theoretical studies.*" Studies in Computational Intelligence, In F. Guillet, and H. J. Hamilton (eds.). Springer: Berlin Heidelberg New York.

[35] Li, Z., He, P., Lei, M. 2005. "*A High Efficient AprioriTID Algorithm for mining Association rule.*" Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, China, August 18-21, pp.18-21.

[36] Li-jian, H., Li-Chao, C., Shuang-Ying, L. 2003. "*Improvement of AprioriTid Algorithm for Mining Association Rules*", Journal of Yantai University, Vol.16, No.4.

[37] Liu, H., Lu, H., Feng, L., and Hussain, F. 1999. "*Efficient search of reliable exceptions.*" Proceedings of the Third Pacific Asia Conference on Knowledge Discovery and Data Mining, Beijing, China, April 26-28, pp.194-204.

[38] Liu, J., Pan, Y., Wang, K., Han, J. 2002. "*Mining frequent item sets by opportunistic projection.*" Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery in Databases, Edmonton, Canada, July 23-26, pp.239-248.

[39] Mannila, H., Toivonen, H., Verkamo, A.I. 1994. "*Efficient algorithms for discovering association rules.*" Proceedings of the AAAI'94 Workshop on Knowledge

Discovery in Databases, Seattle, Washington, July 31-August 4, pp.181-192.

[40] Michael Steinbach, Pang-Ning Tan, Hui Xiong, Vipin Kumar. 2007. "*Objective Measures for Association Pattern Analysis*" International Journal of Contemporary Mathematics, No.443, pp.205-226.

[41] Ng, R. T., Lakshmanan, L. V. S., Han, J., Pang, A. 1998. "*Exploratory mining and pruning optimizations of constrained association rules*." Proceedings of the ACM SIGMOD International Conference on Management of Data, Seattle, Washington, USA, June 2-4, pp.13-24.

[42] Ordonez, C., Ezquerra, N., Santana, C.A. 2006. "*Constraining and summarizing association rules in medical data*." International Journal of Knowledge Information System, Vol.9, Issue.3, pp.259-283.

[43] Ordonez, C., Santana, C.A., Braal, L. 2000. "*Discovering interesting association rules in medical data*." Proceedings of the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, Dallas, Texas, USA, May 14, pp.78-85.

[44] Padmanabhan, B., Tuzhilin, A. 1998. "*A belief-driven method for discovering unexpected patterns.*" Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, New York city, USA, August 27-31, pp.94-100.

[45] Park J.S., Chen, M.S., Yu, P.S. 1995. "*Efficient data parallel mining for association rules.*" Proceedings of the Fourth International Conference on Information and Knowledge Management, Baltimore, Maryland, USA, November 29- December 02, pp.31-36.

[46] Park, J.S., Chen, M.S., Yu, P.S. 1995. "*An effective hash-based algorithm for mining association rules.*" Proceedings of the ACM SIGMOD International Conference on Management of Data, San Jose, California, May 22-25, pp.175-186.

[47] Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L. 1999. "*Discovering frequent closed itemsets for association rules.*" Proceedings of the Seventh International Conference on Database Theory, Jerusalem, Israel, January 10-12, pp.398-416.

[48] Pei, J., Han, J., Mao, R. 2000. "*CLOSET: an efficient algorithm for mining frequent closed itemsets.*" Proceedings of the ACM SIGMOD International Workshop Data Mining and Knowledge Discovery, Dallas, Texas, USA, May 16-18, pp.11-20.

[49] Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., Hsu M-C. 2001. "*PrefixSpan: mining sequential patterns efficiently by prefix-projected pattern growth.*" Proceedings of the International Conference on Data Engineering, Heidelberg, Germany, April 2-6, pp.215-224.

[50] Ragel., Cremilleux, B. 1998. "*Treatment of missing values for association rules.*" Proceedings of the Second Pacific-Asia Conference on Knowledge Discovery and Data Mining, Melbourne, Australia, April 15-17, pp.258-270.

[51] Ronaldo Cristiano Prati, Maria Carolina Monard and André C.P.L.F. de Carvalho. 2004. "*Looking for exceptions on knowledge rules induced from HIV cleavage data set.*" International Journal Genetics and Molecular Biology, Vol.27, Issue.4, pp.637-643.

[52] Salleb, A., Turmeaux, T., Vrain, C., and Nortet, C. 2004. "*Mining quantitative association rules in a atherosclerosis dataset*." Proceedings of the Sixth European Conference on Principles and Practice of Knowledge Discovery in Databases, Pisa, Italy, September 20-24, pp.98-103.

[53] Sarawagi, S., Thomas, S., Agrawal, R. 1998. "*Integrating association rule mining with relational database systems: alternatives and implications.*" Proceedings of the ACM SIGMOD International Conference on Management of Data, Seattle, Washington, USA, June 2-4, pp.343-354.

[54] Savasere, A., Omiecinski, E and Navathe S. 1995. "*An efficient algorithm for mining association rules in large databases.*" Proceedings of the Twenty-One International Conference on Very Large Data Bases, Zurich, Switzerland, September 11-15, pp.432-443.

[55] Savasere, A., Omiecinski, E., Navathe, S. 1998. "*Mining for strong negative associations in a large database of customer transactions.*" Proceedings of the International Conference on Data Engineering, Oralando, Florida, USA, February 23-27, pp.494-502.

[56] Sengul Dogan., Ibrahim Turkoglu. 2008. "*Diagnosing hyper lipidemia using association rules.*" Mathematical and Computational Applications, Vol. 13, No. 3, pp.193-202.

[57] Shantakumar B.Patil., Kumaraswamy, Y.S. 2009. "*Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction*." International Journal of Computer Science and Network Security, Vol.9 No.2, pp.228-235.

[58] Shenoy, P., Haritsa, J.R., Sudarshan, S., Bhalotia, G., Bawa, M., Shah, M. 2000. "*Turbo-Charging Vertical Mining of Large Databases*" Proceedings of the ACM SIGMOD International Workshop Data Mining and Knowledge Discovery, Dallas, Texas, USA, May 16-18, pp.22-23.

[59] Silberschatz, A., Tuzhilin, A. 1996. "*What makes patterns interesting in knowledge discovery systems*?" IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No.6, pp.970-974.

[60] Song, M., Rajasekaran, S. 2005. "*Finding frequent itemsets by transaction mapping."* Proceedings of the twentieth ACM Symposium on applied computing, Santa Fe, New Mexico, March 13-17, pp.488-492.

[61] Song, M., Rajasekaran, S. 2006. "*A transaction mapping algorithm for frequent itemsets mining.*" IEEE transactions on knowledge and data engineering, Vol.18, No.4, pp.472-481.

[62] Srikant, R and Agrawal, R. "*Mining quantitative association rules in large relational tables.*" Proceedings of the ACM SIGMOD International Conference on Management of Data, Montreal, Quebec, Canada, June 4-6, pp.1-12.

[63] Srikant, R., Agrawal, R. 1996. "*Mining sequential patterns: generalizations and performance improvements.*" Proceedings of the Fifth International

Conference on Extending Database Technology, Avignon, France, March 25-29, pp.3-17.

[64] Suzuki, E. 1997. "*Autonomous discovery of reliable exception rules.*" Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, Newport Beach, California, USA, August 14-17, pp.259-262.

[65] Tan, P., Kumar, V., Srivastava, J. 2002. "*Selecting the right interestingness measures for association pattern.*" Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada, July 23-26, pp.32-41.

[66] Toivonen, H. 1996. "*Sampling large databases for association rules.*" Proceedings of the Twenty-Second International Conference on Very Large Data Bases, Bombay, India, September 3-6, pp.134-145.

[67] Waleed A. Aljandal., 2009. *Itemset size-sensitive interestingness measures for association rule mining and link prediction.* Ph.D. Manhattan : Kansas State University.

[68] Wang, J., Han, J., Pei, J. 2003. "*CLOSET+: Searching for the best strategies for mining frequent closed itemsets.*" Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 24-27, pp.236-245.

[69] Wu,X., Zhang,C., Zhang,S. 2004. "*Efficient mining of both positive and negative association rules.*" ACM Transactions on Information Systems, Vol. 22, No. 3, pp.381-405.

[70] Yan, X., Han, J., Afshar, R. 2003. "*CloSpan: mining closed sequential patterns in large datasets.*" Proceedings of the SIAM International Conference on Data Mining, San Fransisco, CA, May 1-3, pp.166-177.

[71] Yan, X., Yu, P.S., Han, J. 2004. "*Graph indexing: a frequent structure-based approach.*" Proceedings of the ACM SIGMOD International Conference on Management of Data, Paris, France, June 13-18, pp.335-346.

[72] Yan, X., Zhu, F., Han, J., Yu, P.S. 2006. "*Searching substructures with superimposed distance.*" Proceedings of the twenty-second International Conference on Data Engineering, Atlanta, Georgia, April 3-8, pp.88.

[73] Zaki, M.J. 2000. "*Scalable algorithms for association mining.*" IEEE Transaction on Knowledge and Data Engineering, vol.12, No.3, pp.372-390.

[74] Zaki, M.J. 2001. "*SPADE: An Efficient Algorithm for Mining Frequent Sequences.*" Machine Learning Journal, Vol.42, No.1-2, pp.31-60.

[75] Zaki, M.J., Parthasarathy, S., Ogihara, M., Li, W. 1997. "*Parallel algorithm for discovery of association rules*." International Journal of Data mining and Knowledge Discovery, Vol.1, No.4, pp.343-374.