ISSN: 2321-7782 (Online)

Volume 4, Issue 4, April 2016

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

Summarizing the Concept of Data Mining, Frequent Pattern Mining and Actionable Pattern Mining Techniques

Dr. K. Kavitha
Assistant Professor, Department of Computer Science
Mother Teresa Women's University
Kodaikanal – India

Abstract: Data mining is an activity that offers business advantages, as well as solutions to some mounting problems associated with exploiting knowledge embedded within corporate databases such as growing disk space capabilities, Improvements over the relational Database Management engines and Enhancements to online Analytical Processing. Data Mining is particularly valuable for organizations that collect large quantities of information. Banks, Insurance Companies, Credit card companies and even medicinal field use this technology to derive critical information from large unwieldy used data samples. The techniques are also widely used in the retailing industry to determine the best arrangements for the products. It is also often associated with other data storage and data manipulation techniques such as the data warehousing and online transaction processing (OLTP). In this paper, detailed overview of Data Mining and its Techniques are discussed as well as reviewed recently implemented techniques.

Keywords: Actionable Pattern Mining, Association Rule Mining, Itemset, Antecedent, Consequent.

I. Introduction

Data Mining is the process of discovering meaningful patterns and relationships that lies hidden within the very large database. Browsing through the tables and the records rarely leads to discover the useful patterns. Data is typically analyzed by an automated process, commonly referred to in data mining as "Knowledge Discovery". Knowledge Discovery is a component of data mining that uses the power of computer combined with the human operator's innate ability to zero in on visually apparent patterns.

By automating the data mining process, computer discovers the patterns and trends present in the data while the person in charge of making use of these discoveries decides which patterns are truly relevant. Data Mining can find descriptive and predictive information. When predictive information is sought, the goal is to derive information that offers clues about the future event. The terms that are related to data mining include the following:

II. Data Mining Algorithms

A Data Mining Algorithm is the mathematical and statistical algorithm that transforms the cases in the original data source into the data mining models. The primary motivation for the field of data mining is to provide support for decision making by detecting useful patterns in large volumes of data. The decisions that are made based on the mined patterns are often crucial to the function of an organization or an enterprise. Such patterns that support the decision making are called "Actionable patterns"[1]. Most data mining algorithms and tools stop with mining and delivery of patterns satisfying intrinsic measures and ignore the decision making with respect to that pattern.

Data Mining is a major step in the knowledge discovery in databases process. It consists of applying computational techniques that, under acceptable computational efficiency limitations produce a particular enumeration of patterns or models over the data [2]. Data mining involves the use of sophisticated data analysis tools to discover the valid data from large set of databases. The tools can include statistical methods, mathematical algorithm and machine learning tools and methods.The progress in data mining research has made it possible to implement several data mining operations efficiently on the datasets [3]. The mined information is typically represented as a model of the semantic structure of the dataset, where the model may be used on new data for prediction or classification. Data mining techniques have been successfully applied in many different fields including marketing, manufacturing, process control, fraud detection and network management [4].

In general, data mining tasks can be classified into two categories: Descriptive mining and Predictive Mining. Descriptive mining is the process of drawing the essential characteristics or general properties of the data in the database. Clustering, association and sequential mining are some of the descriptive mining techniques. Predictive mining is the process of inferring patterns from the data to make the predictions [5]. The predictive mining involves the tasks like the classification, regression and deviation detection [6]. Mining frequent itemsets from transaction databases is a fundamental task for several forms of knowledge discovery such as the association rules, sequential patterns and classification [7]. One of the popular descriptive data mining techniques is Association Rule Mining (ARM) technique. Mining association rules is particularly useful for discovering relationships among the large databases [8]. To improve the efficiency of the Association Rule mining the following techniques are employed such as sampling, reducing the number of passes, Hash – based itemset counting, transaction reduction and partitioning.

A brief overview of data mining as it pertains to the analysis of the event data. It begins with describing the Market – basket analysis, the context in which data mining was first proposed. Market- Basket analysis originates from the analyzing data from the supermarkets, in which each supermarket customer has a basket of purchased goods. The main goal is to find association rules, according to which purchasing a set of items indicates that another set of items also is likely to be purchased. A key problem in data mining problem is to find a set of items, typically referred to as itemsets or patterns, with occurrences above a predefined threshold called minimum support(minsup). A second and closely related problem is prediction, in which patterns that have a high probability of prediction that a given item will be in the same basket. The metric used is “Confidence” and it is expressed as a conditional probability. Fortunately the search for frequent patterns can be made more efficient. The support of patterns can be no greater than the support for its subset. A level wise search algorithm called Apriori was developed to efficiently discover the frequent patterns from large databases. Clearly, such a level – wise algorithm can be generalized to discover any type of pattern satisfying the downward closure property.

Data mining is a mixture of statistical, machine learning and data – management techniques that provides a way to mine categorical data so as to find interesting combinations. The first sequential data mining [9] takes into account the sequence of events rather than just their occurrence. The second is the temporal mining [9] that considers the time between the event occurrences. Data mining has been applied to numerous domains. The more commonly occurring patterns in system management tasks are the Event burst analysis, periodic pattern analysis and mutually dependent patterns. Data Mining has become an increasingly pervasive activity in all areas of research. It is the computer – intensive activity of exploring large datasets in the hope of discovering, within the subset of the data, some relationship or pattern or hypothesis that might be worthy [10].

III. Association Rule Discovery

Association rules are used to show the relationships between data items. The uncovered relationships are not inherent in the data, as with functional dependencies, and they do not represent any sort of correlation. Instead, association rules detect common usage of items. A database in which an association rule is to be found is viewed as a set of tuples, where each tuple contains a set of items. The “Support” of an item (or set of items) is the percentage of transactions in which the data item occurs. There may be an exponential growth in the set of items. This explosive growth in potential sets of items is an issue that most association rule algorithms must contend with, as the conventional approach to generating association rules is in actuality counting the occurrences of sets of items in the transaction database. Some definitions for the association rules are described as follows:

Definition 1: Given a set of items I= $\{I_{1,}I_{2}\ \ldots\ldots\ I_{m}\}$ and a database of transactions D=$\{t_{1,}t_{2,}\ \ldots\ t_{n}\}$ where $t_i=\{I_{1i,}I_{2i}\ \ldots\ldots\ I_{mk}\}$ and I_{ij} Σ I, an association rule is an implication of the form X=> Y where X,Y ⊂ I are sets of items called "Itemsets".

Definition 2: The Support (s) for an association rule X=> Y is the percentage of transaction in the database that contains XUY.

Definition 3: The confidence or strength (α) for an association rule X=> Y is the ratio of the number of transactions that contains XUY to the number of transactions that contains X.

The selection of association rules is based on these two values in which confidence measures the strength of the rule, whereas support measures how often it should occur in the database. Typically, large confidence values and a smaller support are used. The efficiency of association rule algorithms usually discussed with respect to the number of scans of the database that are required and the maximum number of itemsets that must be counted.

Large Itemsets

The most common approach to finding association rules is to break up the problem into two parts such as Find large Itemsets and Generate rules from frequent Itemsets.

An "Itemset" is any subset of the set of all items referred as I. Most association rule algorithms are based on smart way to smart ways to reduce the number of itemsets to be counted. These potentially large itemsets is the "Candidate Itemset (C)", one performance measure used for association rule algorithm is the size of C. Another problem to be solved by association rule algorithm is what data structure is to be used during the counting process. When all large itemsets are found, generating association rules is straightforward. A set of items is referred to as an itemset. An itemset that contains the K – items is a K-itemset. The occurrence frequency of an itemset is the number of transactions that contains the itemset. This is also known simply as the frequency, support count, or count of the itemset. An itemset satisfies minimum support if the occurrence frequency of the itemset is greater than or equal to the product of minimum support and total number of transactions in database D.

Apriori Algorithm

The Apriori Algorithm is the most well known association rule algorithm and is used in most commercial uses. It uses the following property which is called the "Large Itemset Property".

Any subset of a large itemset must be large.

The large itemsets are also said to be downward closed because if an itemset satisfies the minimum support requirements, so do all of its subsets. The basic idea of the Apriori algorithm is to generate candidate itemsets of a particular size and then scan the database to count these to see if they are large. During Scan the itemset I, candidates of size itemset I, C, are counted. Only those candidates that are large are used to generate candidates for the next pass. An itemset is considered as a candidate only if all its subsets also are large. To generate candidates of size i+1, joins are made of large itemsets found in the previous pass.

Generalized Association Rules

Association rules could be generated for any and all levels in the hierarchy. A "Generalized association rule", X=>Y, is defined like a regular association rule with the restriction that no item in Y may be above any item in X. When generating generalized association rules, all possible rules are generated using one or more given hierarchies.

Quantitative Association Rules

A Quantitative association rule is one that involves categorical and quantitative data. In these types of datasets the itemsets are not simple literals. Efficiency has been concerned in the research of association rule mining. The proposed work by Sheng Chai et.al., presents an improved method called Direct - Find - Remove (DFR) algorithm to mine a database consisting of remove and direct steps. When pruning the candidate itemsets, the algorithm eliminates the non – frequent subsets of candidate in the removal step. In the direct steps, the algorithm directly generates the frequent itemsets by computing and comparing the frequency of frequent k - itemset with K - mean time.

The work focuses an algorithm to raise the probability of containing information in scanning database and reduce the potential scale of the itemsets. In classical Apriori algorithm, when candidate generations fare generated, the algorithm needs to test their occurrence frequencies. The manipulation with redundancy results in high frequency querying, so tremendous amount of resources will be expended whether in time or in space.

Increasing the Efficiency of Association Rules Algorithms

The computational cost of association rules mining can be reduced in four ways:

- Reducing the number of passes over the database
- Sampling the database
- Adding extra constraints on the structure of patterns through parallelization.

In recent years much progress has been made in all these directions.

Reducing the number of passes over the database

FP-Tree, frequent pattern mining, is another milestone in the development of association rule mining, which breaks the main bottlenecks of the Apriori. The frequent itemsets are generated with only two passes over the database and without any candidate generation process. FP-tree is an extended prefix-tree structure storing crucial, quantitative information about frequent patterns. Only frequent length-1 items will have nodes in the tree, and the tree nodes are arranged in such a way that more frequently occurring nodes will have better chances of sharing nodes than less frequently occurring ones.

FP-Tree scales much better than Apriori because as the support threshold goes down, the number as well as the length of frequent itemsets increase dramatically. The candidate sets that Apriori must handle become extremely large, and the pattern matching with a lot of candidates by searching through the transactions becomes very expensive. The frequent patterns generation process includes two sub processes: constructing the FP-Tree and generating frequent patterns from the FP-Tree. The mining result is the same with Apriori series algorithms. To sum up, the efficiency of FP-Tree algorithm accounts for three reasons. First, the FP-Tree is a compressed representation of the original database because only those frequent items are used to construct the tree, other irrelevant information are pruned. Secondly this algorithm only scans the database twice. Thirdly, FP-Tree uses a divide and conquer method that considerably reduces the size of the subsequent conditional FP-Tree. Every algorithm has its limitations, For FP-Tree it is difficult to be used in an interactive mining system. During the interactive mining process, users may change the threshold of support according to the rules. However for FP-Tree, the changing of support may lead to repetition of the whole mining process. Another limitation of FP-Tree is that it is not suitable for incremental mining. Since, as time goes on, databases keep changing. New datasets may be inserted into the database. Those insertions may also lead to a repetition of the whole process.

Tree Projection is another efficient algorithm recently proposed in [11]. The general idea of Tree Projection is that it constructs a lexicographical tree and projects a large database into a set of reduced, item-based sub-databases based on the frequent patterns mined so far. The number of nodes in its lexicographic tree is exactly that of the frequent itemsets. The efficiency of Tree Projection can be explained by two main factors: (1) the transaction projection limits the support counting in a relatively small space; and (2) the lexicographical tree facilitates the management and counting of candidates and provides the flexibility of picking efficient strategy during the tree generation and transaction projection phases. Wang and Tjortjis presented an efficient algorithm for mining association rules. Their approach reduces large itemset generation time, known to be the most time-consuming step, by scanning the database only once and using logical operations in the process. Another algorithm for efficient generating large frequent candidate sets is proposed by Matrix Algorithm [12]. The algorithm generates a matrix which entries 1 or 0 by passing over the cruel database only once, and then the frequent candidate sets are obtained from the resulting matrix. Finally association rules are mined from the frequent candidate sets. Experiments results confirm that the matrix algorithm is more effective than Apriori Algorithm.

Sampling

Toivonen [13] presented an association rule mining algorithm using sampling. The approach can be divided into two phases. During phase 1 a sample of the database is obtained and all associations in the sample are found. These results are then validated against the entire database. To maximize the effectiveness of the overall approach, the author makes use of lowered minimum support on the sample. Since the approach is probabilistic (i.e. dependent on the sample containing all the relevant associations) not all the rules may be found in this first pass. Those associations that were deemed not frequent in the sample but were actually frequent in the entire dataset are used to construct the complete set of associations in phase 2. Parthasarathy presented an efficient method to progressively sample for association rules. This approach relies on a novel measure of model accuracy (self similarity of associations across progressive samples), the identification of a representative class of frequent itemsets that mimic (extremely accurately) the self-similarity values across the entire set of associations, and an efficient sampling methodology that hides the overhead of obtaining progressive samples by overlapping it with useful computation.

Chuang et al. [14] explore another progressive sampling algorithm, called Sampling Error Estimation (SEE), which aims to identify an appropriate sample size for mining association rules. SEE has two advantages. First, SEE is highly efficient because an appropriate sample size can be determined without the need of executing association rules. Second, the identified sample size of SEE is very accurate, meaning that association rules can be highly efficiently executed on a sample of this size to obtain a sufficiently accurate result. Especially, if data comes as a stream flowing at a faster rate than can be processed, sampling seems to be the only choice. How to sample the data and how big the sample size should be for a given error bound and confidence levels are key issues for particular data mining tasks. Li and Gopalan derive the sufficient sample size based on central limit theorem for sampling large datasets with replacement.

Parallelization

Association rule discovery techniques have gradually been adapted to parallel systems in order to take advantage of the higher speed and greater storage capacity by [15]. The transition to a distributed memory system requires the partitioning of the database among the processors, a procedure that is generally carried out indiscriminately. Cheung et al. [16] presented an algorithm called FDM. FDM is a parallelization of Apriori to shared nothing machines, each with its own partition of the database. At every level and on each machine, the database scan is performed independently on the local partition. Then a distributed pruning technique is employed. Schuster and Wolff described another Apriori based D-ARM algorithm - DDM. As in FDM, candidates in DDM are generated level wise and are then counted by each node in its local database. The nodes then perform a distributed decision protocol in order to find out which of the candidates are frequent and which are not.

Another efficient parallel algorithm FPM (Fast Parallel Mining) for mining association rules on a shared-nothing parallel system has been proposed by [16]. It adopts the count distribution approach and has incorporated two powerful candidate pruning techniques, i.e., distributed pruning and global pruning. It has a simple communication scheme which performs only one round of message exchange in each iteration. A new algorithm, Data Allocation Algorithm (DAA), is presented in [17] that uses Principal Component Analysis to improve the data distribution prior to FPM. Parthasarathy et al. [18] have written an excellent recent survey on parallel association rule mining with sharedmemory architecture covering most trends, challenges and approaches adopted for parallel data mining. All approaches spelled out and compared in this extensive survey are apriori-based. More recently, Tang and Turkia [19] proposed a parallelization scheme which can be used to parallelize the efficient and fast frequent itemset mining algorithms based on FP-trees.

Constraints based association rule mining

Many data mining techniques consist in discovering patterns frequently occurring in the source dataset. Typically, the goal is to discover all the patterns whose frequency in the dataset exceeds a user-specified threshold. However, very often, users want to restrict the set of patterns to be discovered by adding extra constraints on the structure of patterns. Data mining systems should be able to exploit such constraints to speedup the mining process. Techniques applicable to constraint-driven pattern discovery can be classified into the following groups:

- Post-processing (filtering out patterns that do not satisfy user-specified pattern constraints after the actual discovery process);
- pattern filtering (integration of pattern constraints into the actual mining process in order to generate only patterns satisfying the constraints);
- Dataset filtering (restricting the source dataset to objects that can possibly contain patterns that satisfy pattern constraints).

Wojciechowski and Zakrzewicz [20] focus on improving the efficiency of constraint- based frequent pattern mining by using dataset filtering techniques. Dataset filtering conceptually transforms a given data mining task into an equivalent one operating on a smaller dataset. Tien Dung Do et al [21] proposed a specific type of constraints called category-based as well as the associated algorithm for constrained rule mining based on Apriori. The Category-based Apriori algorithm reduces the computational complexity of the mining process by passing most of the subsets of the final itemsets. An experiment has been conducted to show the efficiency of the proposed technique. Rapid association Rule Mining (RARM) [22] is an association rule mining method that uses the tree structure to represent the original database and avoids candidate generation process. In order to improve the efficiency of existing mining algorithms, constraints were applied during the mining process to generate only those association rules that are interesting to users instead of all the association rules.

IV. Frequent Pattern Mining

Frequent Pattern mining in databases plays an indispensable role in many data mining tasks namely, classification, clustering and association rules analysis, when a large number of item sets are processed by the database, that needs to be scanned multiple times. Consecutively multiple scanning of the database increases the number of rules generation, which then consumes more system resources.

Association rules are valuable patterns because they offer useful insight into the types of dependencies that exist between attributes of data sets. Due to the complete nature of algorithms such as Apriori, the numbers of pattern extracted are often very large. Frequent Pattern Mining is a most powerful problem in association mining. Most of the algorithms are based on association rule mining. Lots of algorithms for mining association rules and their mutations are proposed on the basis of Apriori Algorithm. Frequent Itemset Mining came from efforts to discover useful patterns in customer's transaction databases. A customer's transaction database is a sequence of transactions, where each transaction is an itemset. An itemset is frequent if its support is greater than a support threshold, originally denoted by their minimum support values. Frequent Itemset mining problem is to find all frequent itemset in a given transaction database.

Determining the most frequent patterns in a target database is the primary issue in frequent pattern mining, which is related to the threshold value specified by the user. Many of the traditional frequent pattern mining approaches identify the interesting frequent patterns by deploying a parameter called support. In this approach, the assumption is made by the users to define a suitable threshold value for the minimum support. Combined mining is a technique for analyzing the object relations and pattern relations and for extracting and constructing actionable patterns or exceptions. Although combined patterns can be built within a single method, such as combined sequential patterns by aggregating relevant frequent sequences.The concept of combined pattern mining is introduced by [23] and proposed for handling the complexity of employing multi-feature sets, multi-information sources, constraints, multi-methods and multi models in data mining and analyzing complex relations between the objects or descriptors.

Combined patterns may form through the analysis of the internal relations between the objects or pattern constituents obtained by a single dataset [23]. The main contribution of the combined mining is that it enables the extraction, discovery, construction and induction of knowledge which consists not simply of discriminate objects but also of interactions and relations between objects as well as their impact. This is called “Actionable Complex Patterns”. Combined mining provides an overall solution for meeting the challenge of mining complex knowledge in complex data. It also substantially builds upon other individual approaches such as conceptual inductive learning [24], and inference, generalization, aggregation and summarization, in order to integrate them with the data - driven knowledge discovery from complex environments. Pattern relations analysis augments the following areas such as knowledge representation, reasoning, inductive learning, semantic and ontological engineering, pattern theory and pattern language [25].

In combined mining, the word ‘combined’ principally refers to either one or more of the following aspects on demand.

1) Combination of Multiple data sources

2) Combination of multiple features and

3) Combination of multiple methods.

The outcomes of combined mining are combined patterns, which are actually the patterns evaluated from the heterogeneous sources. Such patterns reflect characteristics of the every source from which they are extracted as they contain features from various sources. More methods can be used for the purpose such as association rules, clustering, classification and prediction, etc., for mining the same data. Resultant combined patterns surely have a complete essence of data by taking advantage of different methods. In order to satisfy the need of the particular business application these patterns needs to be treated by various interestingness measures so that they reveal the importance as well as concerns to the required perspective. Such patterns are known as Actionable Patterns. Interestingness measures need to be developed by taking into account many aspects such as the technical performance, domain knowledge, end user experience as well as the social and organizational factors.

Data Mining has already been widely used in many areas such as the public services, telecom, share market, health care and many more. Now a day the data sources involve heterogeneous data for example transactional data, XML data, text files, etc. [26]. The combined mining process does the different approaches as elaborated below:

1) The transactional data sources involved in data mining applications may have multiple selects the features from all sources which have more importance and incorporate them into resultant patterns. Such patterns are known as combined Patterns [26].

2) Sometimes the data to be mined can be distributed or volume of data can be so large that it is impossible to scan the whole data. Combined mining scans each data source separately and combine the generated patterns.

3) As known, there are many methods of data mining for example association rule mining, classification, clustering, summarization, prediction, etc. But many times, outcome of a single method may not be useful in required perspective. Combined mining takes multiple methods to generate patterns which reveal the real meaning of data.

Multiple data sources with homogenous patterns are easy to handle, but data sources with multiple fearture - set needs to be handled with special management. In multiple featured combined mining, atomic patterns (generated from single source) are merged together to form combined patterns which are often more informative. In many situations the patterns discovered by a particular method do not serve to the user's perspective. Multiple methods can be used parallel, serial or in closed loop fashion [23].

V. Actionable Pattern Mining

Actionability of patterns is measured in subjective as well as objective perspectives. Also the discovered patterns should be both of technical and business interest [27]. The technically interested patterns are said to be independent on certain technical measures defined for a particular mining technique. Technical interestingness is measured in terms of technical objective and technical subjective measures.

1) Technical objective measures are set of criteria which decide whether the given pattern is interested or not. For example in case of association rule mining a pattern is said to be interested if it satisfies minimum support and minimum confidence measures.

2) Technical subjective measures help to rank up to what extent the discovered pattern is useful to a specific user's need.

In business, interested patterns it needs to be decided using various aspects via economical, social, analytical and personal perspectives. Just like the technical interestingness, the business interestingness is also measured in terms of business objective and subjective measures.

1) Business objective measures are criteria depending upon the economical, social or a particular business person's perspectives.

2) Business subjective measures are said to be psychoanalytical measures. A pattern is said to be actionable if it satisfies both the technical as well as the business interestingness measures.

In many database applications, information stored in a database has a built – in hierarchy consisting of multiple levels of concepts. In such database users may want to find out association rules among items only at the same levels. This task is called multiple - level association rule mining. However, mining frequent patterns at multiple levels may lead to the discovery of more specific and concrete knowledge of data.

Initial step is to find the frequent pattern to preprocess the multiple dataset to find the large 1-itemset frequent pattern for all levels. The work proposed by B.Jayanthi et.al [28] introduced a new algorithm called CCB tree that is Category - Content - Brand tree. It is developed to mine Large 1-itemset frequent itemsets for all levels of abstraction. The proposed algorithm is a tree based structure and it first constructed the tree in CCB order for entire database and secondly, it searches for frequent pattern in CCB order. A data mining should provide efficient methods for mining multiple – level association rules. To explore multiple – levels association rule mining, one need to provide data at multiple levels of abstraction, efficient methods for effective association rule mining. The CCB tree has following advantages. It generates the frequent patterns at all levels, it follows top – down searching process so that the searching time is reduced for lower level tree if ancestors are not at minimum support. It also reduces execution time.

A pattern is a combination of relevant *descriptors* associated with certain *relations* (for instance, frequency, classifier or probabilistic distribution) and *constraints*. In the existing pattern discovery and exception mining, a resultant pattern is an individual outcome that has one of the pattern structures detailed below:

Type I:

{Antecedent}, a combination of attributes, in which a pattern is composed of a collection of internal elements in the underlying problem. Typical examples include frequent pattern mining, association rules and clustering. Unsupervised learning usually delivers different combinations of underlying variables.

Type II:

{{*Antecedent*} {*connective*} {*Consequent*}}, or {{*Premise*} {*connective*} {*Conclusion*}}, in which a series of attributes are connected with one another to form the antecedent (or premise) and are then associated in terms of certain *connectives* with (or lead to) an additional *consequent* (or *conclusion*, *Impact*).

Supervised learning such as classification usually delivers outcomes associated with supervised indicators (e.g., class labels). The combination of unsupervised learning with supervised learning [129] also results in this type of deliverable, such as the frequent pattern-based classifier and classification rules. Emerging discussions on high utility pattern mining also fall into this category.

Type III:

{{*Antecedent*}| {*condition*} {*connective*} {*Consequent*}}, in which the occurrence of an antecedent connects consequent results from certain *conditions*. A condition may be an exception or exclusion of some attributes, a constraint, or a certain context. For instance, a mobile preference pattern {{*business managers*}| {*between* 20 *and* 35 *years old*} {{*more likely*} {*touse iphone rather than Blackberry*}}. It calls the source data and pattern outcomes *Type I* if they consist of internal elements only.

Type II data and patterns include additional external conclusions or impacts. If a source or pattern is context, condition or constraint dependent, then it falls into Type III. Type II and Type III patterns are clearly much more informative and actionable [133] than Type I patterns, because they consist of external information (the impact indicator and/or condition) which is in addition to the internal descriptors. While it is often costly or even impossible to obtain the external information, domain and background knowledge driven or partial label based semi-supervised learning is highly valued for learning Type II and Type III patterns on Type I Data. It will mainly discuss pattern combination aspects, pattern structures, relations and paradigms for these three types of data and patterns.

VI. Conclusion

Data mining involves the use of sophisticated data analysis tools to discover the valid data from large set of databases. The tools can include statistical methods, mathematical algorithm and machine learning tools and methods. Data mining involves the use of sophisticated data analysis tools to discover the valid data from large set of databases. The tools can include statistical methods, mathematical algorithm and machine learning tools and methods. This paper highlights the concept, tools and techniques available in Data Mining.

References

1. P.Swapna Raj and Balaraman Ravindran, “Mining Actionable Patterns”, in Proceedings of the 23rd Florida Artificial Intelligence Research Society Conference, 2010.
2. H. Mannila, H.Toivonen and A.Verkamo, :Discovery of Frequent Episodes in Event Sequences, in Data Mining and Knowledge Dsicovery 1, No.3, 259 – 289(1997).
3. Fayyad . U, “Data Mining and Knowledge Discovery in Databases : Implications from Scientific databases”, in Proc. of the 9th Int . Conf on Scientific and Statistical Databases Management, Olympia, Washington, USA, pp,2-11, 1997.
4. Klaus Julish, “Data Mining for Intrusion Detection - A Critical Review “, in Proc . of IBM Research on application of Data Mining in Computer Security , Chapter 1, 2002.
5. Yanbo J.Wang, Qin Xin, Frans Coenen, “A Novel Rule Weighting Approach in Classification Association Rule Mining”, in IEEE Transactions 2007, DOI , ICDMW 126.
6. F.Conen and P.Leng , “An Evaluation of Approaches to Classification Rule Selection “, in Proceedings of the 4th IEEE International conference on Data Mining (ICDM – 04) , IEEE Computer Society , United Kingdom, November 2004, pp.359-362.
7. Wojciechowski . M , Zakrzewicz.M, “Dataset Filtering Techniques in Constraint Based Frequent Pattern Mining “, Lecture Notes in Computer Science , Vol 2447, 2002, pp. 77-83.
8. Fayyad . U, “Data Mining and Knowledge Discovery in Databases : Implications from Scientific databases”, in Proc. of the 9th Int . Conf on Scientific and Statistical Databases Management, Olympia, Washington, USA, pp,2-11, 1997.
9. S.Brin , R.Motiwani and C.Silverstein, “Beyond Market Baskets: Generalizing Assocication Rules to Correlations, “ Data Mining and Knowledge Discovery 2, No. 1, 39 -68, (1998).
10. H. Mannila, H.Toivonen and A.Verkamo, :Discovery of Frequent Episodes in Event Sequences, in Data Mining and Knowledge Dsicovery 1, No.3, 259 – 289(1997).

11. R.Agarwal and C.Aggarwal and V.Prasad " A tree Projection Algorithm for generation of Frequent Itemsets , in International Journal of Parallel and Distributed Computing , 2000.

12. Yuan . Y, Huang . T, "A Matrix Algorithm for Mining Association Rules ", in Lecture Notes in Computer Science, Volume 3664, September 2005,pp. 370-379.

13. Toivonen. H (1996), "Sampling Large Databases for association rules", in The VLDB Journal , pp. 134-145.

14. Chuang .K, Chen .M, Yang . W., "Progressive Sampling for Association Rules based on Sampling Error Estimation, Lecture Notes in Computer Science Vol 3518, June 2005, page. 505 – 515.

15. Zaki. M.J, "Parallel and Distributed association mining: A Survey", IEEE Concurrency , Special Issue on Parallel Mechanisms for Data Mining, 7(4): 14 – 25, December 1999

16. Cheung .D, Han , Florida, . J, Ng.V, Fu. A, and Fu. Y, (1996), "A Fast Distributed Algorithm for Mining Association Rules " in Proceedings of 1996 International Conference of Parallel and Distributed Information Systems, Miami Beach, Florida, pp. 31 – 44.

17. Manning. A, Kieane.J, "Data Allocation Algorithm for Parallel Association Rule Discovery ", Lecture Notes in Computer Science, Volume 2035, Page 213-220.

18. Parthasarathy . S, Zaki.M.J.J, Ogihara.M , "A Parallel data Mining for Association ruleson Shared – memory systems, Knowledge and Information systems: An Overview ", in An International Journal,3 (1),: 1- 29, Feb 2001.

19. Tang . P, Turkia. M, "Parallelizing Frequent Itemset Mining with FP – Trees", Technical Report Titus .compsci.ular.edu/-~ptang /papers/par-fi.pdf, Department of Computer Science, University of Arkansas at Little Rock, 2005.

20. Wojciechowski . M , Zakrzewicz.M, "Dataset Filtering Techniques in Constraint Based Frequent Pattern Mining ", Lecture Notes in Computer Science , Vol 2447, 2002, pp. 77-83.

21. Tien Dung Do, Siu Chueng Hui, Alvis Fong, "Mining Frequent Itemset with Category Based Constraints ", Lecture Notes in Computer Science , Volume 2843, 2003, pp. 76- 86.

22. Das. A, Ng. W and Woon.Y, 2001, "Rapid Association rule Mining ", in Proceedings of the Tenth International Conference on Information and Knowledge Management ACM Press, 474 – 481.

23. Longbing Cao, Huaifeng Zhang, Yangchang Zhao, Dan Luo, Chengqi Zhang, "Combined Mining : Discovering Informative Knowledge in Complex Data", in IEEE Transations on Systems , man and Cybernetics, Vol.41, No.3, June 2011.

24. Mooney. R.J, "Integrating Abduction and Induction in Machine Learning " in Abduction and Induction P. Flach and A.Kakas (Eds)., pp. 181 – 191, Kluwer Academic Publishers, 2000.

25. Alexander .C, Ishikawa, S.Silverseirn, M. Jacobson, Fiksdahi-King, and Angel. S, " A Pattern Language ", Oxford University Press, 1977.

26. Mrs. Suvarna R. Bhagwat , Combined Mining and Actionable Pattern Discovery Using DDID – PD Framework : A Review, in International Journal of Engineering Research and Technology ISSN : 2278 – 0181, Vol 2 Issue 2, February – 2013.

27. Longbing Cao, Huaifeng Zhang, Yangchang Zhao, Dan Luo, Chengqi Zhang, "Combined Pattern Mining : from Learned Rules to actionable knowledge", in Proceedings of AI , pp. 393 – 403, June 2008.

28. K.Duraiswamy and B.Jayanthi , A Novel Preprocessing Algorithm for Frequent Pattern Mining in Multidatasets in International journal of Data Engineering (IJDE) Volume (2) : Issue (3) : 2011.