IJCSET |September 2012 | Vol 2, Issue 9, 1402-1405 www.ijcset.net | ISSN:2231-0711

A Process to Comprehend Different Patterns of Data Mining Techniques for Selected Domains

Reshma Sultana, Vani, Managina Deepti, PV Bhaskhar, Pedada Satish, Koppala KVP Sekhar

Abstract: **One of the important problems in data mining is the Classification rule learning which involves finding rules that partition given data into predefined classes. In the data mining domain where millions of records and a large number of attributes are involved, the execution time of existing algorithms can become prohibitive, particularly in interactive applications. Data mining is a field at the intersection of computer science and statistics, is the process that attempts to discover patterns in large data sets. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.**

Keywords: FP-Growth, Apriori Algorithm, SS-BE, SS-MB

INTRODUCTION:

An enormous amount of data stored in databases and data warehouses, it is increasingly important to develop [1] powerful tools for analysis of such data and mining interesting knowledge from it. Data mining [4] is a process of inferring knowledge from such huge data. It has five major components:

- Association rules
- Classification or clustering
- Characterization & Comparison
- Sequential Pattern Analysis.
- Trend Analysis

An *association rule [5]* is a rule which implies certain association relationships among a set of objects in a database. In this process we discover a set of association rules at multiple levels of abstraction from the relevant set(s) of data in a database. For example, one may discover a set of symptoms [2] often occurring together with certain kinds of diseases and further study the reasons behind them. Since finding interesting association rules in databases may disclose some useful patterns for decision support, selective marketing, financial forecast, medical diagnosis and many other applications, it has attracted [3] a lot of attention in recent data mining research. Mining association rules may require iterative scanning of large transaction or relational databases which is quite costly in processing.

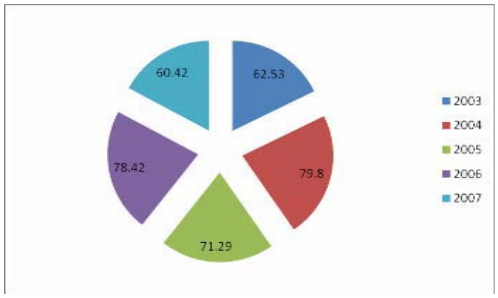
A BRIEF REVIEW OF THE WORK ALREADY DONE IN THE FIELD:

Sequential pattern mining is an interesting data mining problem with many real-world applications. This problem has been studied extensively in static databases. However, in recent years, emerging applications have introduced a new form of data called data stream. In a data stream [6], new elements are generated continuously. This poses additional constraints on the methods used for mining such data: memory usage is restricted, the infinitely [8] flowing original dataset cannot be scanned multiple times, and current results should be available on demand. Mendes, L.F. Bolin Ding, Jiawei Han [9] introduces two effective methods for mining sequential patterns from data streams: the SS-BE method and the SS-MB method. The proposed methods break the stream into batches and only process each batch once. Since there are many temporal pattern discovery algorithms that are modeled along the same lines as the Apriori algorithm, it is useful to first understand how Apriori works before discussing extensions to the case of temporal patterns. Let *D* be a database of customer transactions at a supermarket. A transaction is simply an unordered collection of items purchased by a customer in one visit to the supermarket. The Apriori algorithm systematically unearths all patterns in the form of (unordered) sets of items that appear in a sizable number of transactions. We introduce some notation to precisely define this framework. A non-empty set of items is called an itemset. An itemset i is denoted by (i1, i2, i3, · · · im), where ij is an item. Since i has m items, it is sometimes called an m-itemset. Trivially, each transaction in the database is an itemset. However, given an arbitrary itemset i, it may or may not be contained in a given transaction T. The fraction of all transactions in the database in which an itemset is contained in is called the support of that itemset. An itemset whose support exceeds a user-defined threshold is referred to as a frequent itemset. These itemsets are the patterns of interest in this problem. The brute force method of determining supports for all possible itemsets (of size m for various m) is a combinatorial explosive exercise and is not feasible in large databases (which is typically the case in data mining). The problem therefore is to find an efficient algorithm to discover all frequent itemsets in the database *D* given a User-defined minimum support threshold.

The Apriori algorithm exploits the following very simple (but amazingly useful) principle: if i and j are itemsets such that j is a subset of i then the support of j is greater than or equal to the support of i. Thus, for an itemset to be frequent all its subsets must in turn be frequent as well. This gives rise to an efficient level wise construction of frequent itemsets in *D*. The algorithm makes multiple passes over the data. Starting with itemsets of size 1 (i.e. 1-itemsets), every pass discovers frequent itemsets of the next bigger size. The first pass over the data discovers all the frequent 1-itemsets. These are then combined to generate candidate 2-itemsets and by determining their supports (using a second pass over the data) the frequent 2-itemsets are found. Similarly, these frequent 2-itemsets are used to first obtain candidate 3-itemsets and then (using a third database pass) the frequent 3-itemsets are found, and so on. The candidate generation before the mth passes uses the Apriori principle described above as follows: an m-itemset is

considered a candidate only if all (m−1)-itemsets contained in it have already been declared frequent in the previous step. As m increases, while the number of all possible m-itemsets grows exponentially, the number of frequent m-itemsets grows much slower, and as a matter of fact, starts decreasing after some m. Thus the candidate generation method in Apriori makes the algorithm efficient. This process of progressively building itemsets of the next bigger size is continued till a stage is reached when (for some size of itemsets) there are no frequent itemsets left to continue. This marks the end of the frequent itemset discovery process.

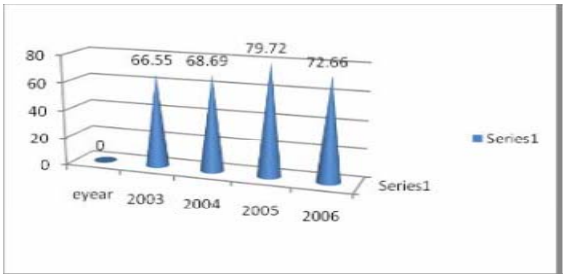
We have done study about pattern of different Result Analysis of Our University Result Data of Different Semesters as shown below.



1.1 Result Graph for BE-101

Year wise Data for Subject Code BE-101

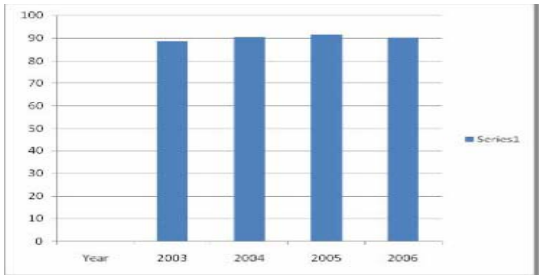
Year	*Rst_Per*
2003	62.5
2004	79.8
2005	71.3
2006	78.4
2007	60.4



1.2 Result Graph for BE-102

1.2 Year wise Data for Subject Code BE-101

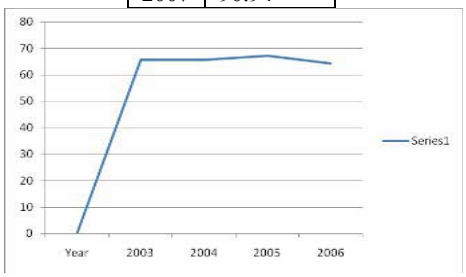
Year	*R_pst*
2003	66.55
2004	68.69
2005	79.72
2006	72.66



1.3 Result Graph for BE-103

1.3 Year wise Data for Subject Code BE-103

Year	*R_percent*
2003	88.62
2004	90.54
2005	91.57
2006	90.28
2007	90.94



1.4 Result Graph for BE-104

LIMITATIONS OF APRIORI ALGORITHM

Apriori algorithm, in spite of being simple and clear, has some limitation. It is costly to handle a huge number of candidate sets. For example, if there are104 frequent 1-item sets, the Apriori algorithm will need to generate more than107 length-2 candidates and accumulate and test their occurrence frequencies. Moreover, to discover a frequent pattern of size 100, such as {a1, a2. . . , a100},it must generate 2100 -2 ˜ 1010 candidates in total. This is the inherent cost of candidate generation, no matter what implementation technique is applied. It is tedious to repeatedly scan the database and check a large set of candidates by pattern matching, which is especially true for mining long patterns. Apriori Algorithm Scans the database too many times, When the database storing a large number of data services, the limited memory capacity, the system I/O load, considerable time scanning the database will be a very long time, so efficiency is very low. In order to overcome the drawback inherited in Apriori, an efficient FP-tree based mining method, FP-growth, which contains two phases, where the first phase constructs an FP tree, and the second phase recursively Researches the FP tree and outputs all frequent patterns.

SOLUTIONS TO OVERCOME DRAWBACKS IN APRIORI

In the association rule mining area, most of the research efforts went in the first place to improving the algorithmic performance and in the second place into reducing the output set by allowing the possibility to express constraints on the desired results. Over the past decade a variety of algorithms that address these issues through the refinement of search strategies, pruning techniques and data structures have been developed. While most algorithms focus on the explicit discovery of all rules that satisfy minimal support and confidence constraints for a given dataset, increasing consideration is being given to specialized algorithms that attempt to improve processing time or facilitate user interpretation by reducing the result set size and by incorporating domain knowledge. There are also other specific problems related to the application of association rule mining from e-learning data. When trying to solve these problems, one should consider the purpose of the association models and the data they come from. Now a

day, normally, data mining tools are designed more for power and flexibility than for simplicity. Most of the current data mining tools are too complex for educators to use and their features go well beyond the scope of what an educator might require. As a result, the courses administrator is more likely to apply data mining techniques in order to produce reports for instructors who then use these reports to make decisions about how to improve the student's learning and the online courses. However, it is most desirable that teachers participate directly in the iterative mining process in order to obtain more valuable rules. But normally, teachers only use the feedback provided by the obtained rules to make decisions about modification to improve the course, detect activities or students with problems, etc.Some of the main drawbacks of association rule algorithms in e-learning are: the used algorithms have too many parameters for somebody non expert in data mining and the obtained rules are far too many, most of them non-interesting and with low comprehensibility. In the following subsections, we will tackle these problems.

- **Finding the appropriate parameter settings of the mining algorithm**
- **Discovering too many rules**
- **Discovery of poorly understandable rules**

Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases.

PROPOSED WORKED TO IMPROVE EFFICIENCY OVER PREVIOUS DATA ANALYSIS.

Methodology & Results

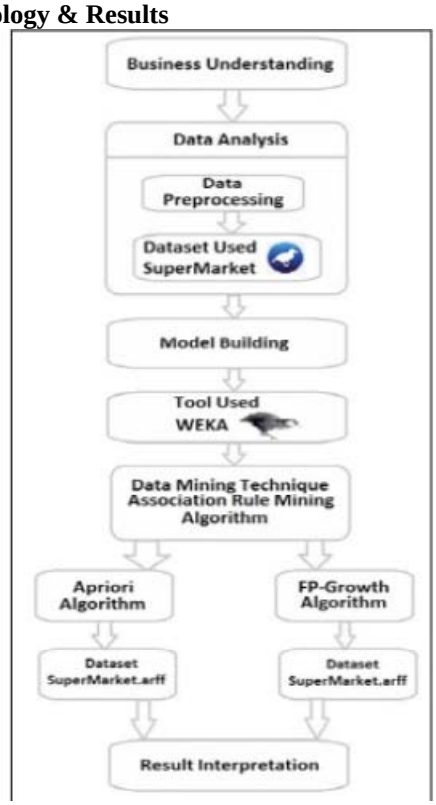
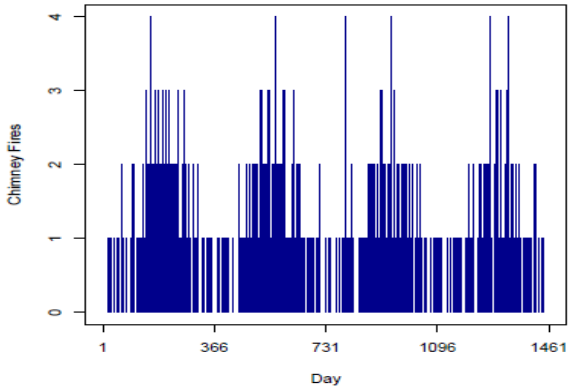


Fig: Methodology to generate frequent itemset using Associations Rules

PROBLEM AND DATA USING REGRESSION

In this scenario, we study chimney fires, in particular the extent to which the frequency of chimney fires fluctuates with changes in temperature and weather conditions, and whether there is a temperature threshold below which the frequency increases markedly.



It is possible that the relationship between the log-rate of chimney fires and other variables is not linear. To find such a nonlinear relationship, generalized additive models are fitted to the data; see Appendix A.4. By using spline functions with 3 degrees of freedom, humid is only very weakly significant to have a nonlinear relationship but it is extremely significant that T has a nonlinear effect. After taking these into account, the value of AIC is further reduced to 5047 from 5053 of the Poisson linear regression model.

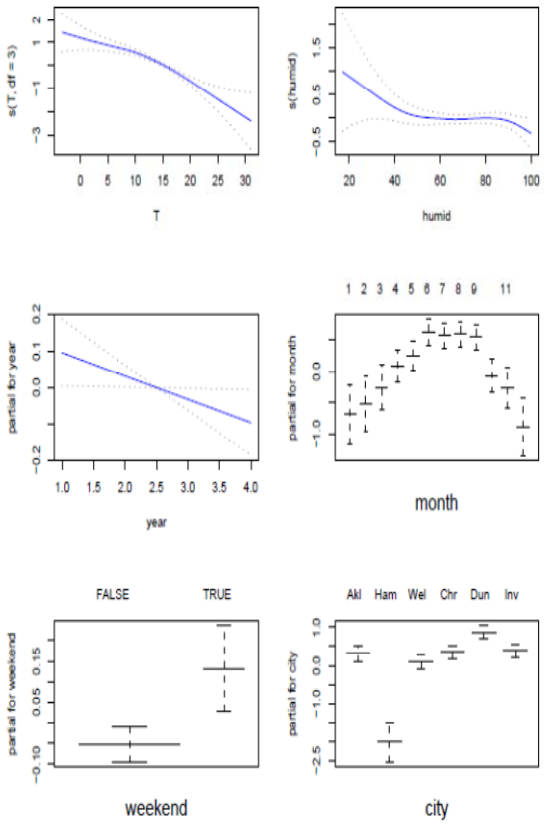


Figure: Partial effects of the variables used in the generalized additive model on the log-rate of chimney fires.

The partial effects of the explanatory variables and their twice-standard-error bands are plotted in Figure .From the plot for month, seasonal effects on the rate of chimney fires are quite obvious—there are simply more chimney fires during the winters, which is not surprising. In addition, at a temperature of about 13 degrees Celsius there is clearly a change in the trend of the partial effect of temperature on the log-rate of chimney fires.

CONCLUSIONS AND FUTURE TRENDS

We have outlined some of the main drawbacks for the application of association rule mining in learning management systems and we have described some possible solutions for each problem. Our goal of research is to find a new scheme for finding the rules out of the transactional dataset which outperforms in terms of running time, number of database scan, memory consumption and the interestingness of the rules over the classical APRIORI ALGORITHM. We believe that some future research lines will focus on: developing association rule mining tools that can more easily be used by educators; proposing new specific measures of interest with the inclusion of domain knowledge and semantic; embedding and integrating mining tools.

ACKNOWLEGMENT

Thanks to Mr.B.KiranKumar working as Asst.Prof in WellFare Institute of Science, Technology & Management in the dept of CSE, Pinagadi helping us to publishing this paper.

REFERENCES

[1] Wang, W., Weng, J., Su, J., Tseng, S.: Learning portfolio analysis and mining in scorm compliant environment. In: ASEE/IEEE Frontiers in Education Conf. (2004) 17–24.

[2] Machado, L., Becker, K.: Distance Education: A Web Usage Mining Case Study for the Evaluation of Learning Sites. In: Proc. of Int. Conf. on Advanced Learning Technologies (2003) 360-361.

[3] Kay, J., Maisonneuve, N., Yacef, K., Zaiane, O.R. : Mining Patterns of Events in Students' Teamwork Data. In: Proc. of Educational Data Mining Workshop (2006) 1-8.

[4] Castro, F., Velllido, A., Nebot, A. and Mugica, F.: Applying Data Mining Techniques to e-Learning Problems: a Survey and State of the Art. Evolution of Teaching and Learning Paradigms in Intelligent Environment. Springer (2007) 183-221.

[5] Tsai, C.J., Tseng, S.S., Lin, C.Y.: A Two-Phase Fuzzy Mining and Learning Algorithm for Adaptive Learning Environment. In: Proc. of Int. Conf. on Computational Science (2001)429-438.

[6] Hwang, G.J., Hsiao, C.L., Tseng, C.R.: A Computer-Assisted Approach to Diagnosing Student Learning Problems in Science Courses. Journal of Information Science and Engineering 19 (2003) 229-248.

[7] [Online] Available: http://www.wikipedia.com/datamining

[8] [Online] Available: http://www.en .wikipedia.org/wiki/ Association_rule_ learning

[9] Online] Available: http://en.wikibooks.org/wiki/ Data_Mining_Algorithms_In_R/Frequent_Pattern_Mining/ The_FP-Growth_Algorithm

[10] [Online] Available: http://www.philippefournier-viger.com /spmf/index.php?link=documentation.php# allassociationrules

[11] Cheng J., Ke Y., Ng W.,"Effective elimination of redundant association rules", Data Mining and Knowledge Discovery Journal, Vol. 16, pp. 221–249, 2008.

[12] J. Han, J. Chiang, S. Chee, J. Chen, Q. Chen, S. Cheng, W. Gong, M. Kamber, K. Koperski, G.Liu, Y. Lu, N. Stefanovic, L. Winstone, B. Xia, O.R. Zaiane, S. Zhang, H. Zhu, `DB Miner: A System for Data Mining in Relational Databases and Data Warehouses", Proc. CASCON'97: Meeting of Minds, Toronto, Canada, November 1997.

ABOUT AUTHORS

Miss Reshma Sultana pursuing M.Tech in Pydah College of Engineering and Technology.

Mrs. G Vani pursuing M.Tech in Pydah College of Engg and Technology, Vishakhapatnam.

Mrs. **Managina Deepti** pursuing M.Tech in Pydah College of Engg and Technology, Vishakapatnam.

PV Bhaskhar pursuing M.Tech in Avanti Institute of Engineering & Technology Cherukupally (P) Near Tagarapuvalasa, Vijayanagaram dist.

Mr.Pedada Satish pursuing M.Tech in Avanti Institute of Engineering & Technology Cherukupally (P) Near Tagarapuvalasa, Vijayanagaram dist.

Mr. Koppala Kvp Sekhar pursuing M.Tech in Avanti Institute of Engineering & Technology Cherukupally (P) Near Tagarapuvalasa, Vijayanagaram dist.