International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-1 Number-2 Issue-2 December 2011

A Modern Non Candidate Approach for sequential pattern mining with Dynamic Minimum Support

Kumudbala Saxena[1], C.S. Satsangi[2]
M-Tech Research Scholar, Medicaps, College Indore[1]
Head and Professor, CSE/IT, Medicaps, College Indore[2]

Abstract

Finding frequent patterns in data mining plays a significant role for finding the relational patterns. Data mining is also called knowledge discovery in several database including mobile databases and for heterogeneous environment. In this paper we proposed a modern non candidate approach for sequential pattern mining with dynamic minimum support. Our modern approach is divided into six parts. 1) Accept the dataset from the heterogeneous input set. 2) Generate Token Based on the character, we only generate posterior tokens. 3) Minimum support is entering by the user according to the need and place. 4) Find the frequent pattern which is according to the dynamic minimum support 5) Find associated member according to the token value 6) Find useful pattern after applying pruning. Our approach is not based on candidate key so it takes less time and memory in comparison to the previous algorithm. Second and main thing is the dynamic minimum support which gives us the flexibility to find the frequent pattern based on location and user requirement.

Keywords

Data Mining, KDD, Dynamic Minimum Support, Frequent Pattern

1. Introduction

Mining data streams is a very important research topic and has recently attracted a lot of attention, because in many cases data is generated by external sources so rapidly that it may become impossible to store it and analyze it offline. Moreover, in some cases streams of data must be analyzed in real time to provide information about trends, outlier values or regularities that must be signaled as soon as possible. The need for online computation is a notable challenge with respect to classical data mining algorithms [1], [2].

With the explosive growth of digital data in every field of life, amount of data is increment at a very high rate. To extract or mine knowledge from these large amounts of data, data mining come forward. The main reason that data mining attracted a great attention of researchers in the information industry in recent years is the availability of huge amounts of data and the need of turning this data into useful information and to extract hidden knowledge. Data mining can be performed on all kinds of information repository. This includes relational databases, data warehouses, transactional databases, advanced database systems, protein and gene sequences data base, social networks, flat files and World Wide Web.

Knowledge discovery or also known as data mining is the processes involve penetration into tremendous amount of data with the help from computer technology for analyzing the data. Data mining is a process of discovering interesting knowledge by extracting or mining from large amount of data and the process of finding correlations or patterns among dozens of fields in large relational databases [3, 4]. Association mining is one of the data mining tasks. The main task is to identify the relationship or correlation between items in dataset. Extensive surveys on the association mining and also frequent pattern mining have been conducted by [5, 6]. Almost a decade numbers of issues related to improve the capability of the algorithm including searching strategy, pruning techniques and data structure involved. The improvements are toward producing more meaningful rules by satisfying minimal support and also confidence constraint. There are also researches related to improvements of the algorithm to meet the domain needs.

Association rule mining is one of the most prominent research topics in data mining. It can be used in discovering relationships among items or events in various application domains. By given a user-specified threshold, also known as minimum support, the mining of association rules can discover the complete set of frequent patterns. That is, once the minimum support is given, the complete set of frequent patterns is determined. In order to retrieve more correlations among items, users may specify a relatively lower minimum support. Such a lower support often generates a huge amount of frequent patterns; but most of the patterns are already known or not interested to users. It is a tedious task for users to filter out these valueless patterns.

We provide here an overview of executing data mining services. The rest of this paper is arranged as follows: Section 2 introduces Data Mining and

Knowledge Discovery; Section 3 describes about frequent pattern; Section 4 shows the recent scenario; Section 5 describes the Proposed Work. Section 6 describes Conclusion.

2. Data Mining and Knowledge Discovery

This process model provides a simple overview of the life cycle of a data mining project. Corresponding phases of a data mining project are clearly identified throughout tasks and relationships between these tasks. Even if the model doesn't indicate it, there possibly exist relationships between all data mining tasks mainly depending on analysis goals and on the data to be analyzed. Six main phases can be distinguished in this process model which is shown in fig 1:

- Business understanding - concerns the definition of the data mining problem based on the business objectives.
- Data understanding - this phase aims at getting a precise idea about data available, identifying possible data quality issues, etc.
- Data preparation - covers all activities meant to build the dataset to analyze from the initial raw data. This includes cleaning, feature selection, sampling, etc.
- Modeling - is the phase where several data mining techniques are parameter and tested with the objective of optimizing the obtained data model or knowledge.
- Evaluation - aims at verifying that the obtained model properly answers the initially formulated business objectives and contributes to deciding whether the model will be deployed or, on the contrary, will be rebuilt.
- Deployment - is the final step of the cyclic data mooning process model. Its target is to take the obtained knowledge, put it in a convenient form and integrate it in the business decision process. It can go, upon the objectives, from generating a report describing the obtained knowledge to creating an specific application that will use the obtained model to predict unknown values of a desired parameter.

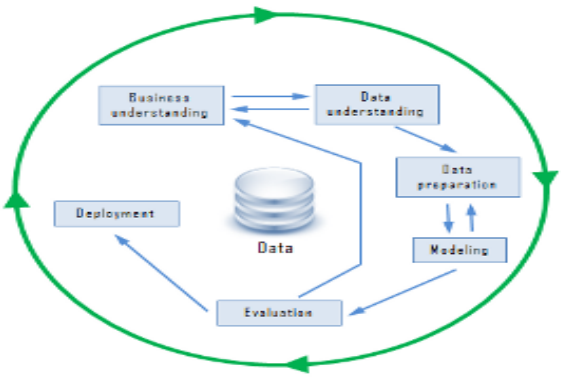

Fig 1: Data Mining Phases

3. Frequent Patterns

There is some fundamental terminology which is important to understand before to understand the whole data mining process. Describing association relationships among the attributes in the set of relevant data is called association rule mining. Find all frequent patterns in a database is called frequent pattern mining. Patterns (set of items, sequence, etc.) that occur frequently in a database are called frequent patterns.

Finding regularities in data is called frequent pattern mining.
Rule Definition:
Body ==> Consequent [Support, Confidence]
(IF <> THEN <>)
Body: represents the examined data.
Consequent: represents a discovered property for the examined data.
Support: represents the percentage of the records satisfying the body or the consequent.
Confidence: represents the percentage of the records satisfying both the body and the consequent to those satisfying only the body
itemset: a set of items
=>E.g., acm={a, c, m}
Support of itemsets
=>Sup(acm)=3
Given min_sup=3, acm is a frequent pattern

Frequent pattern mining: find all frequent patterns in a database currently; frequent pattern mining is performed in batch mode of two sequential steps: enumerating a set of frequent patterns as candidate features, followed by feature selection. Although many methods have been proposed in the past few years on how to perform each separate step efficiently, there is still limited success in eventually finding those highly compact and discriminative patterns. The culprit is due to the inherent nature of this widely adopted batch approach.

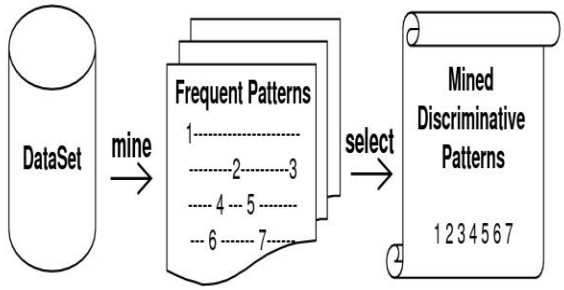

Fig 2: Frequent Pattern Mining

We propose a new and different approach to mine frequent patterns as discriminative features. It builds a Hierarchical structure that sorts or partitions the data onto nodes from the whole list. Then at each node, it directly discovers a discriminative pattern to further divide its examples into purer subsets that previously chosen patterns during the same run cannot separate. Since the number of examples towards leaf level is relatively small, the new approach is able to examine patterns with extremely low global support that could not be enumerated on the whole dataset by the batch method. This approach is shown in fig 3.

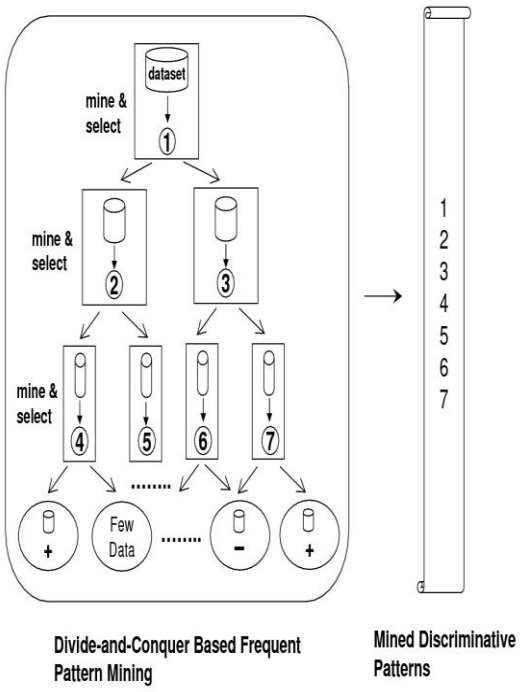

Fig 3: Hierarchical approach for frequent pattern Finding

4. Recent Scenario

In 2010 Ashutosh Dubey et al. [7] proposed a novel data mining algorithm named J2ME-based Mobile Progressive Pattern Mine (J2MPP-Mine) for effective mobile computing. In J2MPP-Mine, they first propose a subset finder strategy named Subset-Finder (S-Finder) to find the possible subsets for prune. Then, they propose a Subset pruner algorithm (SB-Pruner) for determining the frequent pattern. Furthermore, they proposed the novel prediction strategy to determine the superset and remove the subset which generates a less number of sets due to different filtering pruning strategy. Finally, through the simulation their proposed methods were shown to deliver excellent performance in terms of efficiency, accuracy and applicability under various system conditions.

In 2011, Avrilia Floratou et al. [8] proposed a new algorithm called FLexible and Accurate Motif Detector (FLAME). FLAME is a flexible suffix-tree-based algorithm that can be used to find frequent patterns with a variety of definitions of motif (pattern) models. It is also accurate, as it always finds the pattern if it exists. Using both real and synthetic data sets, we demonstrate that FLAME is fast, scalable, and outperforms existing algorithms on a variety of performance metrics.

In 2011, Shawana Jamil et al. [9] focus on focus on investigation of mining frequent sub-graph patterns in DBLP uncertain graph data using an approximation based method. The frequent sub-graph pattern mining problem is formalized by using the expected support measure. Here approximate mining algorithm based Weighted MUSE, is proposed to discover possible frequent sub-graph patterns from uncertain graph data.

In 2011, Ashutosh Dubey et al. [10] proposed a novel algorithm named Wireless Heterogeneous Data Mining (WHDM). The entire system architecture consists of three phases: 1) Reading the Database. 2) Stores the value in Tbuf with different patterns. 3) Add the superset in the list and remove the related subset from the list. Finally we find the frequent pattern patterns or knowledge from huge amount of data. They also analyze the better method or rule of data mining services which is more suitable for mobile devices.

In 2011, Ashutosh Dubey et al. [11] propose a novel DAM (Define Analyze Miner) Based data mining approach for mobile computing environments. In DAM approach, we first propose about the environment according to the requirement and need of the user where we define several different data sets, then DAM analyzer accept and analyze the data set and finally apply the appropriate mining by the DAM miner on the accepted dataset. It is achieved by CLDC and MIDP component of J2ME.

In 2011, Ashwin C S et al. [12] proposed an apriori-based method to include the concept of multiple minimum supports (MMS in short) on association rule mining. It allows user to specify MMS to reflect the different natures of items. Since the mining of sequential pattern may face the same problem, we extend the traditional definition of sequential patterns to include the concept of MMS in this study. For efficiently discovering sequential patterns with MMS, we develop a data structure, named PLMS-tree, to store all necessary information from database.

In 2011, K. Zuhtuogullari et al. [13] observe that an extendable and improved item set generation approach has been constructed and developed for mining the relationships of the symptoms and disorders in the medical databases. The algorithm of the developed software finds the frequent illnesses and generates association rules using Apriori algorithm. The developed software can be usable for large medical and health databases for constructing association rules for disorders frequently seen in the patient and determining the correlation of the health disorders and symptoms observed simultaneously.

5. Proposed Methodology

Our proposed methodology is shown in fig 4. In this paper we proposed a modern non candidate approach for sequential pattern mining with dynamic minimum support. Our modern approach is divided into six parts.

[1] Accept the dataset from the heterogeneous input set: In this phase we accept the data set from the source.

[2] Generate Token Based on the character, we only generate posterior tokens. Posterior Tokens means if item set is a,b,c then the posterior is a,ab,abc but we not include ba,bca combination.

[3] Minimum support is entering by the user according to the need and place: In this paper we use dynamic minimum support which is enter by the authorized person at the run time. For example if we want to find frequent pattern of "ABC Shop" which is situated in Mumbai , then the probability of customer is more in comparison to small cities like Bhopal. So minimum support is always dynamic because it is changed according to the place and also by the need of the user. Minimum support is also change day by day. For example in festival times the customers are more probable to visit shop in comparison to normal days, so the probability of customers in festival days is higher than in the normal days.

[4] Find the frequent pattern which is according to the dynamic minimum support. We get those values from the data set which full fills the minimum support decides by the user.

[5] Find associated member according to the token value:
In this phase we can determine which item set is associated by the user input item set. For example the probability of purchasing milk with bread and sugar is more, so the association is Milk, Bread, and Sugar. It depends on the Market Basket Analysis. This process analyzes customer buying habits by finding associations between the different items that customer buying habits by finding associations between the different items that customer place in their shopping baskets.

[6] Find useful pattern after applying pruning:
In this phase we find the final result which is called pruning on the basis of the minimum support.

Our proposed methodology deals on different dataset or we say that it works on heterogeneous environment which is applicable to mine huge amount of data. We do not concentrate on candidate key generation it also helps to save memory space and also the execution time is increases.

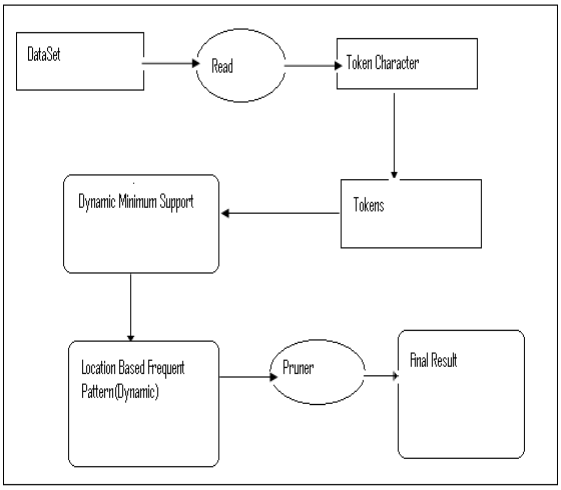

Fig 4: Proposed Methodology

6. Conclusion

In this paper we proposed a modern non candidate approach for sequential pattern mining with dynamic minimum support. Our modern approach is divided into six parts. 1) Accept the dataset from the heterogeneous input set. 2) Generate Token Based on the character, we only generate posterior tokens. 3) Minimum support is entering by the user according to the need and place. 4) Find the frequent pattern

which is according to the dynamic minimum support 5) Find associated member according to the token value 6) Find useful pattern after applying pruning.

In future we design an algorithm which applies on the above algorithm and show the real time example with the simulation comparison, which shows that our approach is better than the previous one.

References

[1] M. M. Gaber, A. Zaslavsky, and S. Krishnaswamy, "Mining data streams: A review," ACM SIGMOD Record, vol. Vol. 34, no. 1, 2005.

[2] C. C. Aggarwal, Data Streams: models and algorithms. Springer, 2007.

[3] J. Han, and M. Kamber, Data Mining: Concepts and Techniques,Morgan Kaufmann, 2001.

[4] I.H.W.E. Frank, Data Mining Practical Machine Learning Tools and Techniques, Morgan Kaufmann Publishers, 2005.

[5] A. Ceglar, and J.F. Roddick, "Association mining," ACM Computing Surveys (CSUR) 2006.

[6] J. Han, H. Cheng, D. Xin, and X. Yan, "Frequent pattern mining: current status and future directions," Data Mining and Knowledge Discovery2007, pp. 55-86.

[7] Ashutosh K. Dubey and Shishir K. Shandilya," A Novel J2ME Service for Mining Incremental Patterns in Mobile Computing", Communications in Computer and Information Science, 2010,Springer LNCS.

[8] Avrilia Floratou, Sandeep Tata, and Jignesh M. Patel," Efficient and Accurate Discovery of Patterns in Sequence Data Sets", IEEE Transactions On Knowledge and Data Engineering, VOL. 23, NO. 8, August 2011.

[9] Shawana Jamil, Azam Khan, Zahid Halim and A. Rauf Baig," Weighted MUSE for Frequent Sub-graph Pattern Finding in Uncertain DBLP Data", IEEE 2011.

[10] Smriti Pandey, Nitesh Gupta, Ashutosh K. Dubey," A Novel Wireless Heterogeneous Data Mining (WHDM) Environment Based on Mobile Computing Environments", IEEE, 2011 International Conference on Communication Systems and Network Technologies.

[11] Ashutosh K. Dubey, Ganesh Raj Kushwaha and Nishant Shrivastava," Heterogeneous Data Mining Environment Based on DAM for Mobile Computing Environments ", Information Technology and Mobile Communication in Computer and Information Science, 2011, Springer LNCS.

[12] Ashwin C S, Rishigesh.M and Shyam Shankar T M," SPAAT-A Modern Tree Based Approach for sequential pattern mining with Minimum support", IEEE 2011.

[13] Zuhtuogullari, K., and N. Allahverdi. "An improved itemset generation approach for mining medical databases." In Innovations in Intelligent Systems and Applications (INISTA), 2011 International Symposium on, pp. 39-43. IEEE, 2011.