Finding Frequent Pattern with Transaction and Occurrences based on Density Minimum Support Distribution

Preeti Khare[1], Hitesh Gupta[2]
Department of Computer Science & Engineering, PCST, Bhopal[1,2]

Abstract

The importance of data mining is increasing exponentially since last decade and in recent time where there is very tough competition in the market where the quality of information and information on time play a very crucial role in decision making of policy has attracted a great deal of attention in the information industry and in society as a whole. In this approach we also use density minimum support so that we reduce the execution time. A frequent superset means it contains more transactions then the minimum support. It utilize the concept that if the item set is not frequent but the superset may be frequent which is consider for the further data mining task. By this approach we can store the transaction on the daily basis, then we provide three different density zone based on the transaction and minimum support which is low (L), Medium (M), High (H). Based on this approach we categorize the item set for pruning. Our approach is based on apriori algorithm but provides better reduction in time because of the prior separation in the data, which is useful for selecting according to the density wise distribution in India. Our algorithm provides the flexibility for improved association and dynamic support. Comparative result shows the effectiveness of our algorithm.

Keywords

Data Mining, Density wise Distribution, Minimum Support, Frequent Pattern

1. Introduction

Data mining is a treatment process to extract useful and interesting knowledge from large amount of data. The knowledge modes data mining discovered have a variety of different types. The common patterns are: association mode, classification model, class model, sequence pattern and so on.

Frequent item set mining algorithms have been investigated for a long time, such as Apriori [1], FP-growth [2], Eclat [3], Tree-Projection [4], H-Mine [5], DHP [6], and so on. The essence of frequent item set mining is to discover frequent subsets from a set of item set. In our study, we propose an efficient approach to discover frequent superset, improved association and dynamic Minimum support.Data mining is a process to extract potentially useful information and knowledge from a large, noise, vague and random data that people don’t identify [7]. In light of such a goal, applying the appropriate data mining and approaches like artificial smart, set theory and statistics, we analysis the data, describe the model and the regulation through visual tools or regulations. With this approach, the policy maker can use the related knowledge to support the decision-making process; domain experts can use it to amend the existing knowledge system; meanwhile, we can also put this as a new knowledge and transfer it into the appropriate knowledge storage organization.

According to Yawei Wang [8] data mining has characteristics as follows: (1) discover models that can reflect system local characteristics and laws; (2) forecast the trends automatically and discover the "new" knowledge; (3) win a lot of rules and be updated timely.

In the today’s scenario in day to day life the database is growing very faster. So that to minimizing information probability for prune is the best option in data mining. Our proposed approach provides a direction in this way, so that we can reduce the transactions and also we operate with dynamic minimum support. The data mining (DM) and knowledge discovery in databases (KDD) movements have been based on the fact that the true value is not in storing the data, but rather in our ability to extract useful reports and to find interesting trends and correlations. The set of DM processes used to extract and verify patterns in data is the core of the knowledge discovery process. These processes involve data selection, data preprocessing, data transformation, DM, and interpretation and evaluation of patterns. Various researchers have made suggestions that domain knowledge should lead the DM process [9].

We provide here an overview of executing data mining services. The rest of this paper is arranged as follows: Section 2 introduces Knowledge Discovery; Section 3 describes about problem domain; Section 4 shows the recent scenario; Section 5 describes the Proposed Work. Section 6 discuss about the result. Section 7 describes Conclusion and outlook.

2. 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.

- 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 a specific application that will use the obtained model to predict unknown values of a desired parameter.

3. Problem Domain

In today’s era data mining is used in very wide sense. There are several researches in this area. We concentrate mainly on the problem of minimum support. If we apply multiple minimum supports in the data mining service it will be helpful in various application area. For example if we want to compare four different minimum support of four different location. If you enter only one minimum support at a time, then the comparison you perform is manual. If your application supports multiple minimum supports with zonal distribution then the above work is very easy and comparative study can be done. Then a problem of subset and superset also arises in previous research. Some algorithm considers the upper limit by which we may lose those items which are not considered because of the lower rank. But if we concentrate on sub-set then all those values which are of lower rank also considered.

4. Recent Scenario

In 2010 Ashutosh Dubey et al. [10] 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. [11] 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. [12] 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 an 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. [13] 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. [14] 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 computing environment or the applications which support the DAM miner on the accepted dataset. It is achieved by CLDC and MIDP component of J2ME.

In 2011, Ashwin C S et al. [15] 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. [16] 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.

In 2011, Reeta Budhani et al. [17] study and proposed about extracting useful insights from large and detailed collections of data. With the increased possibilities in modern society for companies and institutions to gather data cheaply and efficiently, this subject has become of increasing importance. This interest has inspired a rapidly maturing research field with developments both on a theoretical, as well as on a practical level with the availability of a range of commercial tools. They proposed a novel Reconfigurable Character based Token Set Pruner (RCBTSP) for heterogeneous environment

5. Proposed Work

Our proposed approach is based on apriori algorithm which is shown below. The procedure is better understood by the flowchart which is shown in figure1. For explaining the algorithm we consider an example which is shown in table 1. Then we calculate minimum support which is shown in table 2. We consider the example with 50% minimum support and achieve those results which satisfy the criteria of 50 % minimum support which is shown in table 2.

Apriori Algorithm [1]

Step 1: Scan the transaction database to get the support S of each 1-itemset, compare S with Minimum Support (MS), and get a set of frequent 1-itemsets, L_1.

Step2: Use L $_{k-1}$ join L $_{k-1}$ to generate a set of candidate k-item sets. And use Apriori property to prune the non-frequented k-item sets from this set.

Step3: Scan the transaction database to get the support S of each candidate k-item set in the final set, compare S with MS, and gets a set of frequent k-item sets, $L_{k.}$

Step5: For each frequent item set l, generate all nonempty subsets of l.

Step6: For every nonempty subset s of l, output the rule " s => (l-s)" if confidence C of the rule " s => (l-s)"

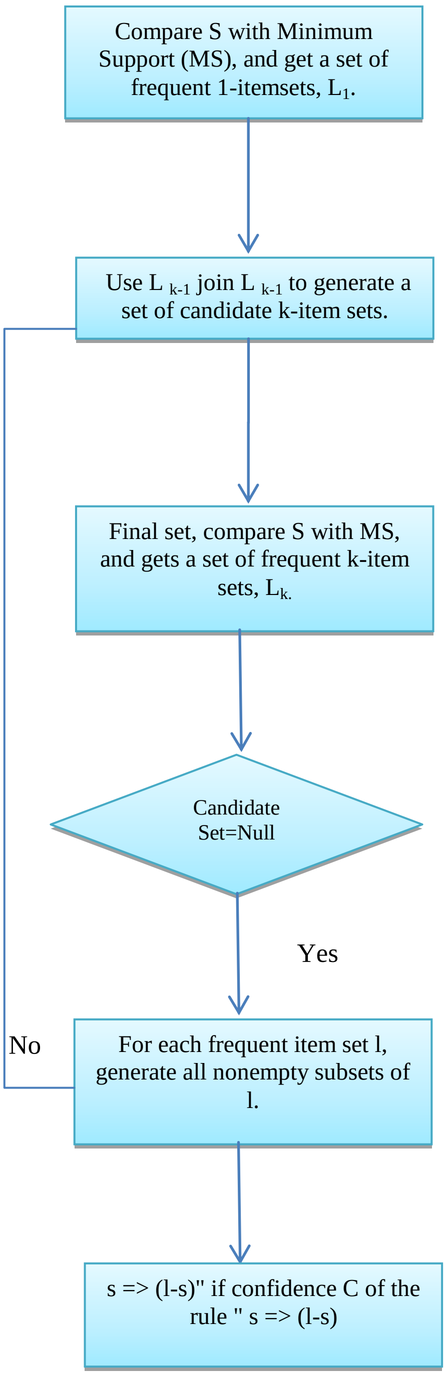
Compare S with Minimum Support (MS), and get a set of frequent 1-itemsets, L1.
Use L k-1 join L k-1 to generate a set of candidate k-item sets.
Final set, compare S with MS, and gets a set of frequent k-item sets, Lk.
Candidate Set=Null
Yes
No
For each frequent item set l, generate all nonempty subsets of l.
s => (l-s)" if confidence C of the rule " s => (l-s)

Figure 1: Flow Chart for Apriori Algorithm

Table 1: Apriori Example

Transaction ID	Items Bought
1	A,B,C
2	A,C
3	A,D
4	B,E

If the minimum support is 50%, then {A,C} is the only 2- item set that satisfies the minimum support.

Table 2: Transaction Id with Support

Frequent Item set	Support
A	75%
B	50%
C	50%
A,C	50%

If the minimum confidence is 50%, then the only two rules generated from this 2-itemset, that have confidence greater than 50%, are:
Shoes → Jacket Support=50%, Confidence=66%
Jacket → Shoes Support=50%, Confidence=100%
Support (A→ B)= Tuples Containing A and B / Total Tuples
Confidence (A→ B)= Tuples Containing both A and B / Tuples Containing A. For understanding our proposed approach we consider the example of table 3. Then we find the minimum support based on the item set. Then we categorize a one day density rule which is shown in Figure 2. According to the rule we categorize the transaction in three different parts which is based on the item set (IS).

Our approach automatically sense the zone and according to the zone minimum support is decided and final prune based calculation is achieved in different zones which is much more faster and better reduction support in time. We consider two types of example one is mixed transaction other is single transaction which is shown in Figure 3 and Figure 7 respectively.

Table 3: Example set

Pattern
21
21,22
21,22,23
21,22,23,24
25

Table 4: Example set with Minimum Support (MS)

Item set	**Minimum Support**
21	4
21,22	3
21,22,23	2
21,22,23,24	1
25	0

Table 5: Based on Transactions

Item set
21
21,22
21,22,23
21,22,23,24
25

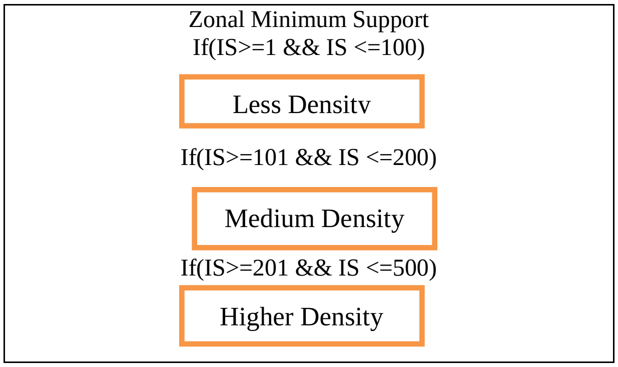

Figure 2: Day Wise Proposed Zonal Minimum Support based on Density

1) If(ms>=1 && ms <=100)
L=10
M=8
H=5
2) If(ms>=101 && ms <=200)
L=20
M=18
H=10
3) If(ms>=201 && ms <=500)
L=40
M=35
H=25

For understanding the above approach we consider our example and take the values of L=3, M=2 and H=1. We can take the value 21,22 is for the lower density area,21,22,23 is used for the middle density area and 21,22,23,24 is for the higher density area.

Table 6: Based on Association

Association
21→22,23,24
22→21,23,24
23→21,22,24
24→21,22,23
25→Null

Table 7: Pattern

Pattern
10
20
50
90
10
40
30
40
50
90
10
80

Table 8: Pattern with Minimum Support

Item set	Support
10	3
20	1
50	2
90	3
40	2
30	1
80	1

For understanding the above approach we consider our example and take the values of L=3, M=2 and H=1. We can conclude that the value 10, 90 is for the lower density area 50, 40 is used for the middle density area and 20, 30, 80 is for the higher density area.

6. Result analysis

Our first comparison is based on the overall execution time of files for single multiple support as shown in table 9. Second comparison is based on multiple minimum support and also the time for the minimum support, which can provide a better comparison in between simple apriori and the Density (D) based apriori as shown in table 10.

Table 9: Execution Time in single MS

Compare1				
Filename	file size	time MS	Time DMS	Time Zonal
ab2.txt	38	13	21	27
main.txt	4122053	370	387	392

7. Conclusion

In this paper we proposed an algorithm for data mining, in which we proposed a density based minimum support concept with subset and superset approach. Our approach provides a greater computation in a less time. For this we provide several computation results based on tables. We also show some result analysis in a comparative way, which shows that our approach is better and efficient. Future enhancement can be done based on the grid environment.

References

[1] Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. Proceedings of the 20th VLDB Conference Santiago, Chile (1994).

[2] Han, J., Pei, J., Yin, Y.: Mining Frequent Patterns without Candidate Generation. ACM SIGMOD (2000).

[3] Zaki, M.J., Parthasarathy, S., Ogihara, M., Li, W.: New Algorithms for Fast Discovery of Association Rules. ACM SIGKDD (1997).

[4] Agarwal, R.C., Aggarwal, C.C., Prasad, V.V.V.: A Tree Projection Algorithm for Generation of Frequent Itemsets. Journal on Parallel and Distributed Computing (Special Issue on High Performance Data Mining), Vol 61, Issue 3 (2000) 350-371.

[5] Pei, J., Han, J., Lu, H., Nishio, S., Tang, S., Yang, D.: H-Mine: Hyper-Structure Mining of Frequent Patterns in Large Database. International Conference on Data Mining, ICDM (2001).

[6] Park, J.S., Chen, M.-S., Yu, P.S.: An Effective Hash-Based Algorithm for Mining Association Rules. In Proc. of ACM SIGMOD (1995) 175-186.

[7] X. Su, J. Yang, N. Jiang and X. Li, Data warehouse and data mining .Beijing: Tsinghua university press, 2006, pp.116–117.

[8] Yawei Wang and Xianghui Hui ,” Application of Data Mining Techniques in Higher College Teaching”, 2011 Third International Workshop on Education Technology and Computer Science.

[9] I. Kopanas, N.M. Avouris and S. Daskalaki, “The role of domain knowledge in a large scale data mining project,” Proc of SETN 2002,LNAI 2308, Springer-Verlag, 2002, pp. 288-299.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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 Communications in Computer and Information Science, 2011, Springer LNCS.

[15] 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.

[16] K. Zuhtuogullari and N. Allahverdi,”An Improved Itemset Generation Approach for Mining Medical Databases”, IEEE 2011.

[17] Reeta Budhani, Dalima Parwani, Meenu Tahilyani, Subuhi Kashif Ansari, “A Novel Reconfigurable Character based Token Set Pruner (RCBTSP) for Heterogeneous Environment”, International Journal of Advanced Computer Research , Volume 1 Number 2 December 2011.

Table 10: Execution Time in multiple MS

Compare2									
filename	**ms1**	**ms2**	**ms3**	**ms4**	**ms5**	**MS**	**APRIORI**	**DAPRIORI**	**Time Zonal**
ab2.txt	3	2	0	0	0	2	26	21	27
main.txt	30	40	50	60	0	30	1480	387	392