A Non Candidate Subset-Superset Dynamic Minimum Support Approach for sequential pattern Mining

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 Subset-Superset Dynamic minimum support approach for sequential pattern mining. Our Whole procedure is further subdivided in four parts1) Superset with minimum support 2) Subset with minimum support 3) Superset mining with Dynamic Support 4) Subset Mining with Dynamic Support. The entire above block includes 1) Accept the dataset from the 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 this approach we also find improved association, which shows that which item set is most acceptable association with others. Here we also provide the flexibility to find multiple minimum supports which is useful for comparison with associated items and dynamic support range. Our algorithm provides the flexibility for improved association and dynamic support. Comparative result shows the effectiveness of our algorithm.

Keywords

Data Mining, KDD, Dynamic Minimum Support, Frequent Pattern, and Non Candidate approach

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 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. 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.

- 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.

3. Research Objective

In today’s era data mining is used in very wide sense. There is several researches in this area. We concentrate mainly on the problem of minimum support. If we apply multiple minimum support 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 then the above work is very easy and comparative study can be done.

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 n 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.

In 2012, Ashutosh Dubey et al. [14] Proposes an efficient method for knowledge discovery which is based on subset and superset approach. In this approach we also use dynamic 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 also find improved association, which shows that which item set is most acceptable association with others. A frequent subset means it contains less transactions then the minimum support. It utilizes the behavior that the less count may be frequent if we attached the less count with the higher order set. Here we also provide the flexibility to find multiple minimum supports which is useful for comparison with associated items and dynamic support range. Our algorithm provides the flexibility for improved association and dynamic support. Comparative result shows the effectiveness of our algorithm.

5. Proposed Methodology

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.

For the above consideration we proposed an algorithm which is shown below.

Assumption:
SPMS-Superset Minimum Support
SMMS-Subset Minimum Support
SPDMS- Superset Dynamic Minimum Support
SMDMS- Subset Dynamic Minimum Support
DS- Data Set

```
Algorithm:
Input 1 {N1,---------Nn}
Output  {Np1,-------Npn}
1.   [Choose from the list]
  1.1 SPMS [DS]
  1.2 SMMS [DS]
  1.3 SPDMS [DS]
  1.4 SMDMS [DS]
     1.1  SPMS [DS]
Read (DS)
{
   for each itemset I1{NSk+1
   if (length(NSk+1)>0) then
    {
    find pattern(Sk+1)
    {
     for(i=0; i<=k-1; i++)
        {
         if(a[i]length!=null)
          {
             read[i];
           }
Sring tokenizer ab[i]=new String
Tokenizer(string,"ch")
      list=ab[i];
      }
      }
  append list;
      }}

  Freq(list)
      {
while (str=object.read line())!=null)
     {
     sting tokenizer=new sting
tokenizer(str,token);
while(st2.has more element)

     {
     data [index]=(string) st2.next element();
          index++;
     }
if (data[i].index of(data1[p])>=0)
{
   c+++;}}

Association(list)
{
for(int j=0;j<Sixth.index;j++)
         {
             p1=0;p2=0;
             String str=Sixth.data[j];
               if(str.indexOf(find[i])>=0)
               {

                    while(true)
                    {
                       p2=str.indexOf(tonk,p2+1);
                       if(p2>=0)
                       {
                          s=str.substring(p1,p2);

if(find[i].equalsIgnoreCase(s)==false)
                          {
                             if(jstr.indexOf(s)==-1)
                                jstr=jstr+" , "+s;
                          }
                          p1=p2+1;
                       }
                       else
                       {

s=str.substring(p1,str.length());

if(find[i].equalsIgnoreCase(s)==false)
                           {
                             if(jstr.indexOf(s)==-1)
                                jstr=jstr+" , "+s;
                          }
                          break;}}}}
            if(jstr.startsWith(" , "))
              jstr=jstr.substring(1);
           data[i]=find[i]+" => "+jstr;

          jTextArea1.append(data[i]+"\n");
        }}}
Prune(data)
{
[Enter the minimum Support]
for(int i=0;i<Sixth.p;i++)
```

```
    {
        int n=Integer.parseInt(Sixth.data1[i][1]);
        System.out.println(Sixth.data1[i][1]);
        if(n>=a)
        {
           data[ind]=Sixth.data1[i][0];
          // jTextArea1.append(data[ind]+"\n");
           lm.addElement(data[ind]);
           ind++;}}}
   if (Second.choice==1)
      jButton2.setText("Super Set");
   else if(Second.choice==2)
      jButton2.setText("Sub Set");
```

1.2 SMMS [DS]

```
Read (DS)
     Freq(list)
     Association(list)
     Prune(data)
```

1.3 SPDMS[DS]

Same as SPMS but use dynamic minimum support

1.4 SMDMS[DS]

Same as SMS but use dynamic minimum support

Our propose algorithm is categorized in the following manner shown in Figure 1:

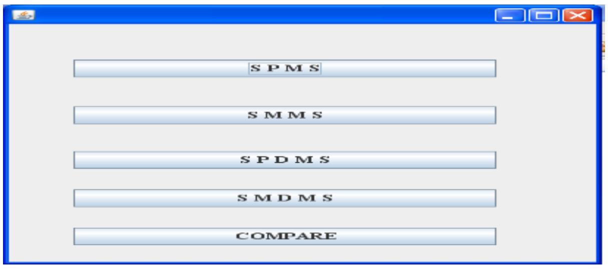


Figure 1: Application Architecture

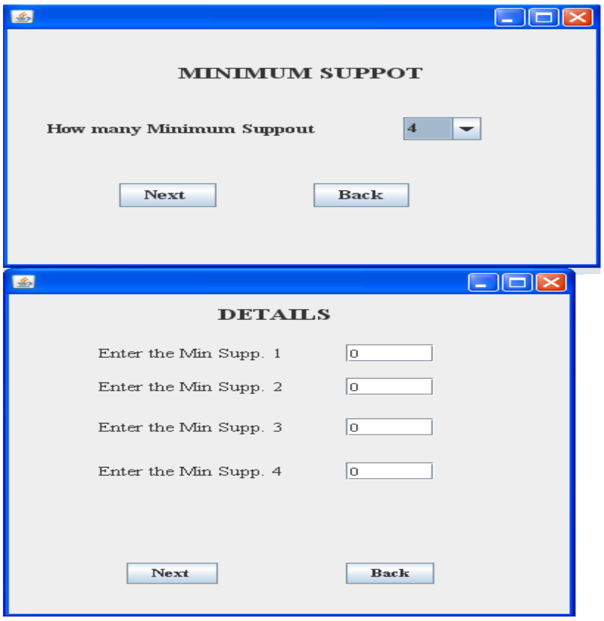
In SPMS, we can read the dataset, arrange in patterns, find the frequency, find associations, finally perform the pruning based on the minimum support and find the superset of the item.

In SMMS, we can read the dataset, arrange in patterns, find the frequency, find associations, finally perform the pruning based on the minimum support and find the subset of the item.

In SPDMS, we can read the dataset, arrange in patterns, find the frequency, find associations, finally perform the pruning based on the minimum support and find the superset of the item [Figure 2].

In SMDMS, we can read the dataset, arrange in patterns, find the frequency, find associations, finally perform the pruning based on the minimum support and find the subset of the item.

Finally perform the comparison, and according to the comparative value we show the highest and lowest frequent data. We also perform improved association, where we can show the possibility of the item set for increasing the frequency.


MINIMUM SUPPOT

How many Minimum Suppout 4

DETAILS

Enter the Min Supp. 1 0
Enter the Min Supp. 2 0
Enter the Min Supp. 3 0
Enter the Min Supp. 4 0

Figure 2: Dynamic Minimum Support

4. Conclusion and Future Work

In this paper we proposed an algorithm for data mining, in which we proposed a dynamic minimum support concept. We present the concept of dynamic support and single support, based on the result discussion, our efficient is efficient in terms of execution time.

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 and 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 Communications 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] K. Zuhtuogullari and N. Allahverdi ,"An Improved Itemset Generation Approach for Mining Medical Databases".

[14] Ashutosh Kumar Dubey, Animesh Kumar Dubey, Vipul Agarwal, Yogeshver Khandagre, "Knowledge Discovery with a Subset-Superset Approach for Mining Heterogeneous Data with Dynamic Support",Conseg-2012.