IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-10, Oct 2018 E-ISSN: 2347-2693

Unbalanced Data Classification using Feature Selection through BitApriori Algorithm.

Pratik A Barot[1*], H B Jethva[2]

[1]GEC Gandhinagar, India
[2]GEC Patan, India

*Corresponding Author: pratikabarot@gmail.com

Available online at www.ijcseonline.org

Accepted: 20/Oct/2018, Published: 31/Oct/2018

Abstract- Frequent pattern mining is used to derive association rules. Association rules specify relativity of target class with rest of the feature(s). The Apriori and FP-growth algorithms are the most famous algorithms used for frequent pattern mining. Classification with feature selection approach is also widely used. This paper provides a detailed study of frequent pattern mining using BitApriori algorithm and use mined association rules for performance improvement of unbalanced data classification. We present a model called FPCM which first mine association rules. Mined association rules are than used for features selection. In final phase, selected features are used in unbalanced data classification using decision tree classifier. Our model shows improved accuracy as compare to the past studies.

Keywords- Frequent pattern mining, Apriori, BitApriori, Unbalanced data classification, machine learning.

I. INTRODUCTION

Data mining is the process of analyzing data from different perspectives and summarizing it into knowledgeable information - information that can be used to increase revenue, cuts costs, or both [1]. In data mining, frequent pattern mining (FPM) and classification (Supervised and unsupervised) are most important data mining functionality. Over the last two decades, numerous algorithms have been proposed for frequent pattern mining [2, 3, 4, 5]. In the era of rapidly increasing data FPM becomes more important [6]. Frequent pattern mining play an important role in domains like market basket analysis, crime pattern mining, chemical analysis etc. [9].

From mined frequent pattern, association rules can be generated. Features derived from association rules can be used for classification. There are classification algorithms which are based on association rules. CBA, CMAR and CPAR are association rule based classifier. CBA uses Apriori algorithm for association rule generation while CMAR uses FP-Growth algorithm [1]. These algorithms classify the given tuple based on the rule matching procedure.

For effective machine learning system feature subset selection is an important research issue. Particularly in classification it helps the learning algorithm to focus on those aspects of the data that are most useful for analysis [15]. In classification it is very important to select effective feature for given class label. Feature subset selection saves time and increases the classification accuracy.

In imbalance data classification, selection of class related features and weighting to those features is important task. Class imbalance and within-class imbalance are major challenges in the area of classification.

Unbalance dataset makes classification even more complicated and time consuming which decreases overall accuracy of classifier. K. Rajeswari et al. [14] proposed the use of Apriori in feature selection in reduced dataset and shows improvement in classifier accuracy. But dataset reduction in areas like medical and judicial system result into loss of vital information and it is not tolerable as in this area each evidence has some importance in final decision making process.

The focus of this paper is to provide an overview of the Apriori algorithm and BitApriori algorithm [2, 3] with their differences and to show important of frequent pattern mining in determination of attribute weight.

Chang Wan [13], proposed the use of FP-Tree in discovery of association rule in order to develop cost sensitive Naïve Bayesian algorithm. Although FP-growth has good performance as compare to Apriori, improvement in Apriori algorithm is important [2, 9].

In this paper, Section II includes description of Apriori algorithm and its limitation followed by detailed explanation

of working of BitApriori algorithm and then performance comparison of both the algorithms. Section III devoted to show the use of association rule mining in determination of attribute weight in unbalanced data classification. Section IV presents a novel model to detect strong causal effect relationship between target class attribute and non-class attribute(s). Our model detects strong association rule which is used in deciding weight of attributes for classification. Section V discusses result analysis. This paper ends with Section VI of conclusion followed by references.

II. BACKGROUND STUDY AND RELATED WORK

II.I Apriori algorithm:

Apriori algorithm is based on prior knowledge. It adopts breadth search and find all frequent itemsets. When the database is sparse and the frequent itemsets are short, Apriori and its improvements give high performance [3]. FP-growth is also an important algorithm for frequent pattern mining but it is complex to implement and suffer in incremental approach [8].

Limitation of Apriori

- Apriori algorithm afflicted with a number of a weakness. The key limit is actually expensive throwing away of time to hold a massive amount of applicant pieces along with much typical itemsets, small minimal service as well as large itemsets [7].
- Apriori require large number of database scan to determine support count which result into high I/O rate. This will adversely affect the performance of Apriori algorithm.
- Candidate generation dominate the run time.
- If longest frequent itemset is of size n then n times of database scan is require.

II.II BitApriori

BitApriori uses a stream of bits to represent items within transaction and then perform logical operation between bit-stream to calculate support count [3]. In contrast to the Apriori algorithm, BitApriori uses bitmap to map items within transaction. Due to this BitApriori require less database scan as compare to the Apriori algorithm.

1st step

The database will be scanned to create the list of all items. After that, non-frequent items are deleted and the list of items is sorted by the descent of support count. This list is set of frequent items will be used as a map of bit stream for later processes. This map presents the order of bits in each bit
stream.

2nd step

BitApriori scans the database one more time again. All transactions are presented by bit streams. If transaction T contain item-i which has position p-th in item list, the p-th bit of T's bit stream will be set 1, otherwise will be set 0. We assume that the first transaction {1, 3} contains 2 frequent items; 1 and 3, which are respectively the 1st and the 4th in the list. Therefore, it will be presented by bit stream as 1001, the 1-st bit is the right most one.

3rd Step- Candidate Generation algorithm

```
For l= 0 to nL1 -2 do
        For l1 = l+1 to nLk – 1 do
        Begin
        IF (Lk-1 [1] and L[l1] have same first (k-2) items)
        Begin
                B=l or l1
                if (subset(b) is freq. itemset)
                        b→C2
                End
        End
        End
End
```

4th step - Support counting

For counting support for all candidate in Ck, BitApriori algorithms does AND operation between candidate ci and transaction bit stream bj then compares result with candidate. Because the AND operation only shows the common bits from 2 bit streams, if result bit stream equals to the candidate stream ci, the transaction bi contains ci and ci's support increases 1. Providing that ci's support >= minsup, ci will be inserted into Lk.

```
For i=0 to nck do
Begin
 For j=0 to nT:
        If ((b[j] AND CK[i])=CK[i])
          SupportCK[i]++;
        If (Supportck[i]>=minSup)
          Insert(Ck[i],Lk+1);
end
```

II.III Performance comparison

Major drawback of Apriori is its slow running time. Bit-Apriori reduces number of database scan and thus decreases running time. Table 1 shows run time comparison of both the algorithms for varying support count [2].

Table 1 Run time comparison of Apriori and Bit-Apriori (In sec.)

Support Count	Apriori	BitApriori
25	2.577	0.626
20	2.5	0.632
15	2.96	0.638
10	3.5	0.640
5	5.0	0.763

As support count is decreases running time of Apriori algorithm increases rapidly. While in the case of BitApriori, decrease in support count has little effect on its running time.

III. FREQUENT PATTERN MINING AND UNBALANCED DATA CLASSIFICATION.

As per [10], unbalanced data classification is classification of dataset which does not have equal class distribution. As per Issac Triguero et al. [11], in imbalance class distribution positive data tuples are over-flooded by negative data tuple. That means dataset has majority and minority class value and minority class gets hidden under the majority class. Traditional classification algorithm like Naïve Bayesian, Decision tree, SVM etc can classify majority class with good accuracy but these algorithms fail in classification of minority class. There are lots of areas which only interested in minority class classification. Like Medical diagnosis, chemical process reaction, Political Science, Judicial data are few names where minority class has more value then majority class.

There are some researches in the area of imbalanced data classification. As per [12], there are some techniques like oversampling, under-sampling, one class learning for imbalance data classification. There are certain weak points in use of oversampling/under-sampling and one class learning which detailed out by Astha Agarwal et al. [10]. But detailed discussion of those weak points is beyond the scope of this paper.

As per [9], frequent pattern mining play important role in crime data analysis for frequent pattern mining from criminal data. Our practical study found similar type of application of frequent pattern mining in the area of judicial system, medical diagnosis, political science etc. Based on the result of BitApriori algorithm from unbalanced data we proposed model called frequent pattern based classification model.

K. Rajeswari et al. [14] proposed use of Apriori in feature selection in reduced dataset. Feature selection from reduced dataset can work in classification of binary class dataset but it performs worst in multi-class classification.
We proposed a novel model for feature selection and classification. Our model first selects important features for a given class and than after uses those selected features as an input to the classifier algorithm for classification of unbalanced dataset.

IV. PROPOSED FREQUENT PATTERN BASED CLASSIFICATION MODEL (FPCM)

Apriori and so as BitApriori algorithms are based on candidate generation process. These algorithms iteratively generate frequent k-itemset. Based on this functionality we proposed a model which uses BitApriori algorithm to decide selection of non-class attributes based on their strong association with class attribute. Use of BitApriori instead of Apriori reduces running time of our classification model.

Our model is shown in Figure 1. In the first phase data is preprocess to make it suitable for frequent pattern mining using BitApriori. BitApriori algorithm is modified to concentrates only on frequent patterns which contains target class label. Based on the association rules discovered by BitApriori, features are selected which are later on used for the classification. This reduced set of features decrease learning time and increases classifier accuracy.

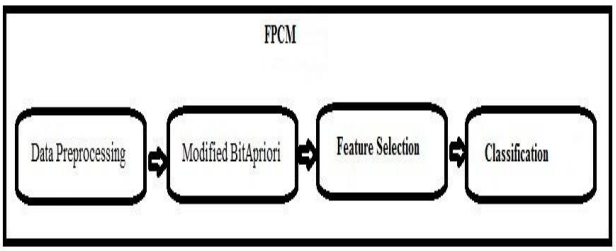

Figure 1. *FPCM Model*

V. RESULT ANALYSIS

Figure 2 shows result of FPCM. We used Heart and Breast Cancer dataset from UCI repository [16]. FPCM uses BitApriori algorithm instead of Apriori for feature subset selection and thus decreases running time and increases classifier accuracy. For classification we used Weka implementation of Decision tree classifier.

```
Time taken to build model: 0.16 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         229               84.8148 %
Incorrectly Classified Instances        41               15.1852 %
Kappa statistic                          0.6907
Mean absolute error                      0.2612
Root mean squared error                  0.3588
Relative absolute error                 52.8863 %
Root relative squared error             72.2019 %
Total Number of Instances              270

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.887     0.200     0.847      0.887     0.866      0.692    0.877     0.866     absent
               0.800     0.113     0.850      0.800     0.824      0.692    0.877     0.855     present
Weighted Avg.  0.848     0.161     0.848      0.848     0.848      0.692    0.877     0.861

=== Confusion Matrix ===

   a   b   <-- classified as
 133  17 |   a = absent
  24  96 |   b = present
```

Figure 2. *FPCM result*

Table 2 Performance Comparison (Run Time is in Sec.)

Dataset	Rule Mining With Reduced Dataset [14] Accuracy | Run Time		FPCM	
			Accuracy	Run Time
Heart	81.10	0.4	84.81	0.16
Breast Cancer	92.8	0.9	95.28	0.31

Table 2 shows that FPCM is more accurate and need less time as compare to earlier study by K. Rajeswari et al. [14]. As per K. Rajeswari et al. accuracy is decreases after feature subset selection. But in the case of FPCM, running time decreases and accuracy increases. This is because of the use of BitApriori algorithm. Figure 3 shows graphical representation of performance comparison. We used accuracy as an performance parameter.

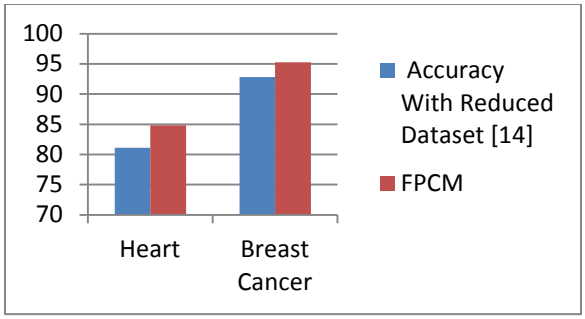

Figure 3. Performance Comparison

VI. CONCLUSION

BitApriori has very good performance improvement as compared to Apriori algorithm in terms of running time. This is because the BitApriori checks group of item at once by means of logical operation on bit-stream of transaction. BitApriori algorithm significantly reduces physical I/O operation by reducing database scan. Our proposed model FPCM uses improved version of BitApriori for attribute selection phase. With the help of selected feature classification is performed using decision tree algorithm. Our new model shows improvement in classification accuracy and decreases learning.

REFERENCES

[1] Varsha Mashoria, Anju Singh, "Literature Survey on Various Frequent Pattern Mining Algorithm", IOSRJEN, Vol-3, Jan-2013.
[2] Sumit Aggarwal, V Singal, "A Survey on Frequent Pattern Mining Algorithms", (IJERT) ISSN: 2278-0181 4, April – 2014
[3] Le Thi Thanh Nhan, Thi T T Nguyen, Tae Chong Chung, "BitApriori: An Apriori-Based Frequent Itemsets Mining Using Bit Streams", IEEE, 2010
[4] E. Ansari, G.H. Dastghaibifard, M. Keshtkaran, H.Kaabi, "Distributed Frequent Itemset Mining using Trie Data Structure", IAENG Inter. Journal of Comp. Sci., 2008
[5] Charu C. Aggarwal, Mansurul A. Bhuiyan and Mohammad Al Hasan, "Frequent Pattern Mining Algorithms: A Survey", Switzerland, Springer International Publishing Switzerland, 2014
[6] Abdul Rahaman Wahab Sait, and Dr.T.Meyappan, "Data preprocessing and Transformation Technique to Generate Pattern From the WebLog", Dubai (UAE), ICSIS'2014, Oct 17-18, 2014.
[7] Abdolrashid Rezvani, J Hosseinkhani, "Enhancing the Performance of BitApriori Algorithm in Data Mining using an Effective Data Structure", Switzerland, IJACSIT,Vol. 4, No. 3, 2015, Page: 85-92.
[8] Samiksha Kankane, V garg, "A survey paper on : Frequent Pattern Analysis Algorithm from the Web Log Data", IJCA, Vol-119, June-2015.
[9] D.Usha, Dr.K.Rameshkumar, "A Complete Survey on application of Frequent Pattern Mining and Association Rule Mining on Crime Pattern Mining", International Journal of Advances in Computer Science and Technology, vol-3, April-2014.
[10] Astha Agrawal, Herna L Victor, Eric Paquet, "SCUT: Multi-Class Imbalanced Data Classification using SMOTE and Cluster-based Undersampling", SCITEPRESS, 2015.
[11] Isaac Triguero, Sara D Rio, V Lopez, J Bacardit, J M Benitez, F Herrera, "ROSEFW-RF: The winner algorithm for the ECBDL'14 big data competition: An extremely imbalanced big data bioinformatics problem", Knowledge Based System, Elsevier, 28 May 2015.
[12] Sotiris Kotsiantis, Dimitris Kanellopoulos, Panayiotis Pintelas, "Handling imbalanced datasets: A review", GESTS International Transactions on Computer Science and Engineering, Vol.30, 2006.
[13] Chang Wan, "Test-Cost Sensitive Classification on Data with Missing Values in the Limited Time", Springer-Verlag Berlin Heidelberg 2010.
[14] K. Rajeswari, "Feature Selection by Mining Optimized Association Rules based on Apriori Algorithm", IJCA, Vol. 119, Jun-2015
[15] Guangtao Wang, Qinbao Song, "Selecting Feature Subset via Constraint Association Rules", Springer-Verlag Berlin Heidelberg 2012
[16] UCI repository, http://tunedit.org/repo/UCI/heart-statlog.arff.

