International Journal of Engineering Sciences Paradigms and Researches (IJESPR)
(Vol. 16, Issue 01) and (Publishing Month: August 2014)
(An Indexed, Referred and Impact Factor Journal)
ISSN (Online): 2319-6564
www.ijesonline.com

Apriori Based Classification using WEKA

Sonia[1] and Dr. Rajesh Gargi[2]

[1]M.Tech Scholar, Geeta Engineering College, Noltha, Panipat, Haryana (India)
rythmhooda37@gmail.com

[2] Professor, CSE Deptt., Geeta Engineering College, Noltha, Panipat, Haryana (India)
directorengg@geetainstitutes.com

Abstract

Multivariate time series analysis has attracted considerable attention in the computational biology due to the ability to measure mRNA expression level of thousands of genes simultaneously with high-throughput microarray technology. Various computational techniques including association rule mining, clustering, Boolean networks, and Bayesian networks, have been applied to elucidate gene regulatory relationships from time series microarray data. The research implements the classification with association and apply Apriori algorithm to generate the rules of association. Then these rules are used for the classification. For analyzed the process it can use the supermarket dataset available with WEKA. The supermarket dataset is a time series dataset that keep record of the sales at a supermarket. The sales are used to classify the class of goods. The proposed work firstly uses the association to relate various items then classify them by the classification algorithm.

Keywords: *Data Mining, Classification, Association, APRORI.*

I. Introduction

Data mining is a process of analyzing data from different perspectives and summarizing it into useful information. It uses sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms and machine learning methods. The algorithms used in data mining are used in such a way that it can improve their performance automatically through experience such as neural networks or decision trees. This data mining in simple terms can be said as knowledge mining, since data mining deals with extracting or mining of knowledge. It attempts to discover hidden rules underlying the data and for this reason it is also called as data surfing. It is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses [1]. Data mining tools predict future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that traditionally were time consuming to resolve. Databases with hidden patterns, finding predictive information that expert may miss because it lies outside their expectations. Most companies already collect and refine massive quantities of data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to improve the value of existing information resources, and can be integrated with new products and systems. Examples of profitable applications illustrate its relevance to today's business environment as well as a basic description of how data warehouse architectures can evolve to deliver the value of data mining to end users. Mining frequent patterns in transaction databases, time-series databases, and many other kinds of databases has been studied popularly in data mining research. Most of the previous studies adopt an Apriori-like candidate set generation-and-test approach. However, candidate set generation is still costly, especially when there are prolific patterns and/or long patterns [1]

Classification produces a function that maps a data item into one of several predefined classes, by inputting a training data set and building a model of the class attribute based on the rest of the attributes[2]. Decision tree classification has an intuitive nature that matches the user's conceptual model without loss of accuracy [3]. However no clear winner exists [4] amongst decision tree classifiers when taking into account tree size, classification and generalization accuracy.

Association rule mining is the efficient method which is used in finding the association rules [5]. These association rules describe the associations between the attribute values of any item set. They can be found by means of various methods among which support and confidence [6],[7] will be considered as the optimized methods in finding them. The key to find the association rules is to find all the frequent item sets present in the given transactional record by means of the minimum support threshold. An association rule is best expressed by means of the expression X -> Y. it means that for any occurrence of item X present in the database there is relatively high probability of occurring the item Y. here X is called as antecedent and Y is called as the consequent. The strength of such rule can only be calculated by means of its support and confidence.

- **Support**

Support of an association rule is defined as the percentage/fraction of records that contain X Y to the total number of records in the database. Support(s) is calculated by the following formula:

$$Support(XY) = \frac{SupportcountofXY}{TotalnumberoftransactioninD}$$

Support is used to find the strongest association rules in the item sets

- **Confidence**

Confidence is another approach for finding the association rules. Confidence of an association rule is defined as the percentage/fraction of the number of transactions that contain X Y to the total number of records that contain X, where if the percentage exceeds the threshold of confidence an interesting association rule X=>Y can be generated.

$$Confidence(X|Y) = \frac{Support\ (XY)}{Support(X)}$$

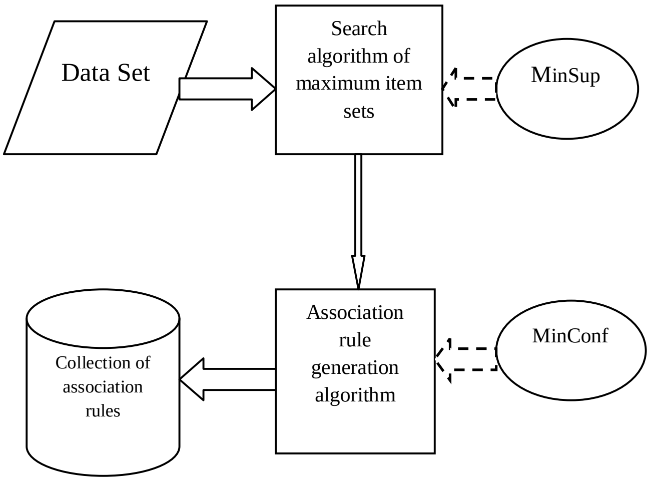
Data Set
Search algorithm of maximum item sets
MinSup
Association rule generation algorithm
MinConf
Collection of association rules

Figure 1: The Basic Model of Association Rule Mining

II. Aprori Algorithm

Apriori algorithm is one of the Data Mining algorithm which is used to find the frequent items/ itemsets from a given data repository. [9] The Apriori Algorithm is an influential algorithm for mining frequent item sets for Boolean association rules. All combination of items in a set of transactions that occurs with a specified minimum frequency. These combinations are called frequent itemsets. Apriori calculates the probability of an item being present in a frequent itemset, given that another item or items is present. Apriori discovers patterns with frequency above the minimum support threshold. Therefore, in order to find associations involving rare events, the algorithm must run with very low minimum support values. However, doing so could potentially explode the number of enumerated itemsets, especially in cases with a large number of items. This could increase the execution time significantly. Classification or anomaly detection may be more suitable for discovering rare events when the data has a high number of attributes. The algorithm mainly involves 2 steps: Pruning and joining. The Apriori property is the important factor to be considered before proceeding with the algorithm [9].

Apriori Property: if an item X is joined with item Y,

Support(X U Y) = min(Support(X), Support(Y))

Algorithm

```
    //find all frequent itemsets
    Apriori(database D of transactions, min_support) {
    F1 = {frequent 1-itemsets}
    k = 2
    while Fk-1 ≠ EmptySet
    Ck= AprioriGeneration(Fk-1)
    for each transaction t in the database D{
    Ct= subset(Ck, t)
    for each candidate c in Ct {
    count c ++
    }
    Fk = {c in Ck such that countc ≥
    min_support}
    k++
    }
    F = U k ≥ 1 Fk
}
//prune the candidate itemsets
AprioriGeneration(Fk-1) {
//Insert into Ck all combinations of elements in
Fk-1 obtained by self-joining itemsets in Fk-1
//self joining means that all but the last item
in the itemsets considered "overlaps," i.e join items p, q
from Fk-1 to make candidate k-itemsets of form p1p2 …p k-
1q1q2…q k-1 (without overlapping) such that p i =q i for
i=1,2, .., k-2 and pk-1 < qk-1.
//Delete all itemsets c in Ck such that some (k-1)
subset of c is not in Lk-1
//find all subsets of candidates contained in t
Subset(Ck, t)
```

III. Proposed Work

The proposed work implements the classification with association. The Apriori algorithm is used to generate the rules of association. Then these rules are used for the classification. The process is analyzed over the supermarket dataset available with WEKA. The supermarket dataset is a time series dataset that keep record of the sales at a supermarket. The sales are used to classify the class of goods. The proposed work firstly uses the association to relate various items then classify them by the classification algorithm. The whole process can be easily understood by following algorithm:

1. Input dataset
2. Generate association rules using the Apriori Algorithm
3. The item will high correlation will form a cluster
4. For each cluster
5. Calculate the information gain
6. Classify the items on the basis of information gain by using J48 algorithm
7. End
8. Compare the classification results with original results to evaluate the classification accuracy.

The existing rule based classification algorithm can be compared with the proposed algorithm by using the WEKA.

The table 1 shows the parameters comparison of the J48, simple CART and the proposed algorithm. The parameters are TP i.e. true positive rate and FP i.e. false positive rate, classification accuracy, precision, recall and the F-measure.

Table 1: Parameter Analysis of Various Algorithms

Algorithm Name	Classification accuracy	Tp rate	Fp rate	Precision	Recall	F-measure
Existing	63.713	0.637	0.637	0.637	0.496	0.5
Proposed	76.6587	0.767	0.304	0.767	0.762	0.744

The comparison can also be done graphically as shown in the following figures.

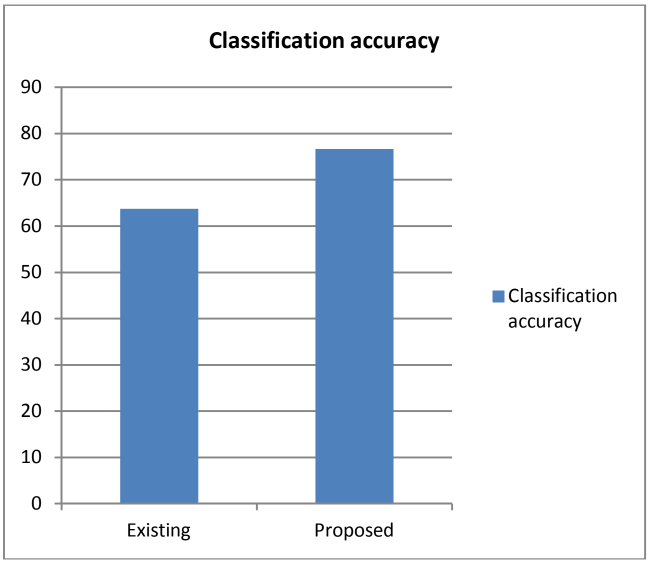

Figure 2: Classification Accuracy Comparison between Existing and Proposed Algorithm

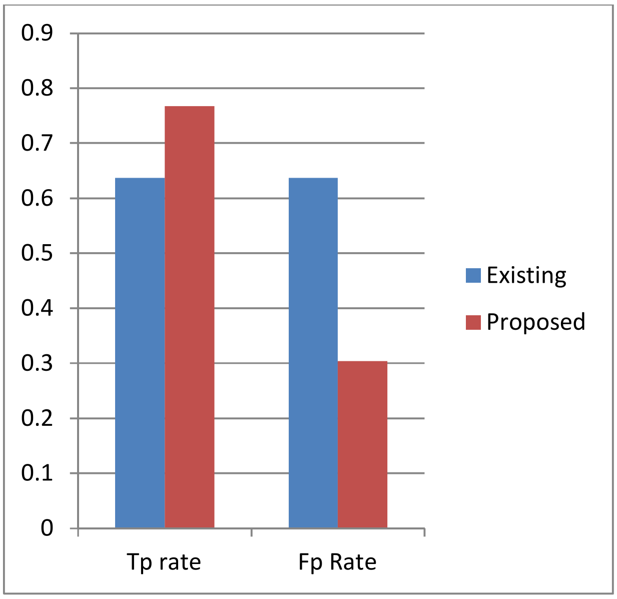

Figure 3: Tp Rate and Fp Rate Comparison between Existing and Proposed Algorithm

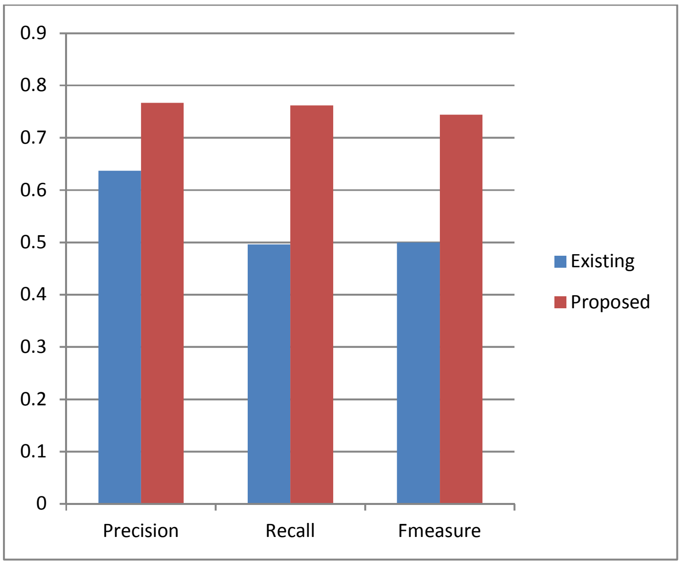

Figure 4: Precision, Recall and F-Measure Comparison between Existing and Proposed Algorithm

The figure 2, 3 and 4 shows the comparison of the various parameters between the J48, simple cart and the proposed algorithm. A significant increase of almost 7% in the classification accuracy can be analyzed using the above the figures. The Fprate get decreased and the true positive rate get increased. The Precision, Recall as well as the F-measure of the proposed algorithm are better than the existing algorithms. It means the performance of the proposed algorithm is better than the existing algorithms.

IV. Conclusion

The Paper implements the classification with association. The Apriori algorithm is used to generate the rules of association. Then these rules are used for the classification. The process is analyzed over the supermarket dataset available with WEKA. The supermarket dataset is a time series dataset that keep record of the sales at a supermarket. The sales are used to classify the class of goods. The proposed work firstly uses the association to relate various items then classify them by the classification algorithm. The implementation is done by using the WEKA tool. The simulation result shows the comparison of the various parameters between the existing and the proposed algorithm. A significant increase of almost 13% in the classification accuracy can be analyzed using the above the figures. The Fp rate gets decreased and the true positive rate get increased. The Precision, Recall as well as the F-measure of the proposed algorithm are better than the existing algorithms. It means the performance of the proposed algorithm is better than the existing algorithms.

References

[1] Surendiran R, Rajan. K.P, Sathish Kumar.M, “Study on the Customer targeting using Association Rule Mining”,(IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 07, 2010, 2483-2484.

[2] Tjortjis, C., & Keane, J. T3: “A Classification Algorithm for Data Mining” (pp. 50-55). Springer Berlin Heidelberg. (2002).

[3] Ganti, V., Gehrke, J., & Ramakrishnan, R. “Mining Very Large Databases”. Computer, 32(8), 38-45, (1999).

[4] Kohavi, R., Sommerfield, D., & Dougherty, J. “Data Mining Using a Machine Learning Library in C++”. International Journal on Artificial Intelligence Tools, 6(04), 537-566. (1997)

[5] Priyanka Asthana, “A Survey on Association Rule Mining Using Apriori Based Algorithm and Hash Based Methods”, Volume 3, Issue 7, July 2013.

[6] J.S. Park, M. Chen, and P.S. Yu, "An Effective Hash Based Algorithm for Mining Association Rules," Proc. 1995 ACM SIGMOD Int'l Conf. Management of Data, ACM Press, 1995.

[7] J.W. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, San Francisco, 2001.

[8] Lei Chen, “The Research of Data Mining Algorithm Based on Association Rules”,The 2nd International Conference on Computer Application and System Modeling ,2012.

[9] Phani Prasad J, Murlidher Mourya, “A Study on Market Basket Analysis Using a Data Mining Algorithm”, Volume 3, Issue 6, June 2013.