IJCSC 0973-7391 Volume 9 • Issue 1 Sept 2017 - March 2018 pp. 42-52 available online at www.csjournals.com

Optimized Predictive Model using Artificial Neural Network for Market Basket Analysis

Roshan Gangurde[1], Dr. Binod Kumar[2], Dr. S. D. Gore[3]
[1]Research Scholar, Dept of Comp. Sc., Savitribai Phule Pune University Pune, Maharashtra, India
[2]Director, Jayawant Institute of Computer Applications, Pune, Maharashtra, India
[3]Rtd. Professor, Department of Statistics, Savitribai Phule Pune University, Pune, Maharashtra, India
[1]roshanant@gmail.com, [2]binod.istar.1970@gmail.com, [3]sharaddgore@gmail.com

Abstract: Market Basket Analysis (MBA) is a modeling technique in view of the theory that in the event that you purchase a specific items, you are progressively likely to purchase another items. The changing requests of the consumer with estimation of seasons are the main task against the market basket analysis. MBA is nothing but predictive model which is used to predict the buyer's behaviour with goal of finding the relationship among various products from their market basket. The optimization in finding of such relationships can help the retailers and merchants to design a sales strategy by considering the items frequently purchased together by customers. Regardless of benefits of using MBA, there are some major research challenges associated with the MBA designing in previous methods. As there is significant growth of online shopping portals and product purchase now days, the current predictive models are ineffective and inefficient over large sales datasets. In this paper, we are attempting to design optimized predictive model to overcome the current research problems. We proposed novel predictive model for MBA by using data cleaning and neural network approach. Our designed data cleaning method helps to improve the quality of input dataset and hence MBA results by removing the all types of errors from it. Secondly unsupervised machine learning based MBA model based on artificial neural network designed. The existing Apriori algorithm is modified by using neural network method in order to optimize the prediction results. To the best of our knowledge, this is the first attempt in MBA. The practical results showing that proposed predictive model for MBA outperforming the previous method.

Keywords: Apriori, Data Mining, Data Cleaning, Market Basket Analysis, MBA, Neural Network.

Introduction

The working of Market Basket Analysis (MBA) framework is purely based on prediction of buyer's paradigm. The buying items patterns of the buyer are recognizing, which product are obtained together which product is purchasing with another product means finding right combinaton of product, which product are obtain in specific seasons. Such approach inclines to yield exceptionally valuable data. Assume that buyer can buy set of items in particular group, then how likely buyer can buy some other similar groups of items [1]. For example, if buyer is buying the raincoats in rain season, then how likely will they buy other products of rainy season? Such information helps to increase the sale of such products sets and hence re-designing the products bundles with aim of increasing business profit i.e. assume that shopping store owner or manager needs the prediction system to recognize the paradigm of their customers [2]. From many individual order, shopping store owner or management required to find group of products the customer will purchase in present visit to the shop. This can be achieved through the process of MBA. MBA is performed on large database which containing the previous records of transactions at shopping store [3]. Such recorded transactions are fetched with respect to year, month, week and day.

There are number of benefits of using MBA. However, there are few designing issues related to MBA which needs to be addressed. The first main issue which needs to be addressed is to modify the customers' requests based on time and seasons. This can be modifying based on customers buying paradigm. Based on such approach, the efficient MBA is performed over large transaction datasets which are composed of different buying transactions [4]. Let's consider that size of input dataset is big, this leads to extra burden of cost and time to perform the MBA over complete dataset, and this may reduce the profit using MBA. Hence there is challenge of improving the profit ratio of MBA using large datasets. This can be done by using data cleaning methods of data mining domain. So our first contribution in this paper is to propose novel data cleaning algorithm with the goal of reducing the size of original dataset which is not actually utilized in process of MBA [5]. This process is also called as data pre-processing. This solves the problem of data cleaning in MBA domain.

In MBA, another important research problem is finding the frequent item sets [6]. This is the major problem as the working of MBA completely depends on process of discovering the bundle of item sets or product sets. As MBA is nothing but the process of finding association rules, the Apriori algorithm is widely used to discover the association rules. For example, the shopping store ABC needs to conduct the study over their buyers in order get the associations as well as relationship among different items purchased in shopping store. MBA considers predicting what products are regularly bought together by customers [6].

The designing of framework is done in order to perform the multi-dimensional data mining process in which every variable is represented in one specific dimension. The dimension include is the items bought and time [7] [8]. The main principle of this research paper is to propose a novel approach to predict what products are generally bought together by consumers. MBA is done by mining the association rules among different sets of products bundles that are bought by customers together. The proposed MBA approach is decent which used to perform the logical decision support for any kind of retail market. Another quality of proposed MBA method is that association rule mining is performed over the internal features of products as well. This can be done by automatic approach of words segmentation [9] [10].

Additionally, to optimize the performance of proposed MBA method, we used neural network classification approach. We adopted FFNN (feed forward neural network) to train the already known and desired datasets. The FFNN with single output layer is nothing but single layer FFNN. The preparation of FFNN is done by altering the network. This can be done by changing the weights through the backward propagation with objective of achieving the designed result for every input training set. Only just altering the neural network weights, the process of performing MBA repeatedly is prevented. This helps in the minimization of time and cost associated in performing frequently MBA over the large database of consumer transaction. Basically, the results of MBA method are based on times and seasons. The purchase of products of customers in one specific season is not similar as the other season and hence it is must to perform the task of market basket analysis repeatedly. In previous methods, this process leads to extra consumption of resources and costs, however in our proposed approach single layer FFNN is used with Apriori algorithm. In section II, we are discussing the related works to solve the problem of MBA. In section III, the design and working of proposed single layer FFNN is discussed. In section IV, the algorithms are discussed. In section V, the practical results and analysis is presented. Finally in section VI, the conclusion and future work is presented.

Related work

This section presents the review of previous methods precisely.

J. Han et.al. (2006)

In [1], author presented different data mining methods. It clarifies data mining and the tools utilized as a part of finding knowledge from the gathered data or information. They referred as the knowledge discovery from data (KDD). They focused on the versatility, feasibility, usefulness and effectiveness of methods of large data sets. Author clarified the strategies for preparing, preprocessing, knowing and warehousing data. They first claimed the information about data warehouses, data cube technology and online analytical processing (OLAP). Then, the functions involved in mining frequent paradigm, correlations and associations for large data groups were described.

S. Haykin et.al. (2009)

In [2], approach proposed for predicting the relationship among the different items based on association rules technique. Authors claimed that necessity of designing association rules for discovering probabilities of items appearance in dataset. This helped to compute the frequent items prediction. They discussed the analysis among the different products in MBA based on Apriori algorithm. They presented different comparative results on this work for designing the rules and its prediction outcomes.

A. Shepherd et.al (2000)

In [3], author introduced the framework to analyze e-commerce click stream as well as focused data. They presented the forecasting over conducted shopping in online shopping basket. They collected the e-commerce dataset from the Turkey region. They designed the data mining algorithms for the analysis of online items buying behavior of customers. Based on the customers behavior, the prediction model predicts whether that particular customer willing to buy the selected products or not. Author used the classifier for prediction called multi-layer neural network along with decision tree method.

Warnia Nengsih et.al (2015)

In [4], novel approach was designed for MBA based on apriori method. They presented the comparative study among MBA using apriori technique and MBA without using apriori algorithm. The apriori technique was used to discover the association rules to predict novel bundles of products. The comparative study was conducted

based on three factors such as rule creation process, rule achieved and concept. Their comparative study claimed that both techniques had similar concept, different ways to rule creation, however discovered rule is same.

Anna Gatzioura et.al. (2015)

In [5], author introduced the new technique for addressing few challenges in MBA based on case based recommendation. Their main focus was meaningful recommendations generation by utilizing the co-occurrence patterns as well deriving detailed information in buying habits of customer. Their recommendation approach is based on ratings of users, history of users as well as case based items recommendation which helped to analyze the similarity among items. They used hierarchical model for representation of items as well as searches.

Yi Zuo et.al. (2015)

In [6], author focused on improving the conception of customer buying nature in shopping store as well as attention extensive understanding from analysis of shopping store behavior. Consequently, they try to show an extrication of obtain nature utilizing quantifiable learning thesis on arranging RFID information that can explain purchase design within a method of customer's in-store nature. Author introduced the three different elements such as behavior element, attitudinal element and personal element in order to design the model of purchase behavior. They analyzed these elements effects via data collected for several customers of their previous buy prevalence.

Anshul Bhargav et.al. (2014)

In [7], author discussed on MBA and attempted to analyze the item sets dependency. They suggested to use ANN (artificial neural network) in order overcome the current limitations of MBA method. They utilized the single layer FFNN (Feed Forward Neural Network).

Djoni Haryadi Setiabudi et.al. (2011)

In [8], author implemented yet another MBA method in this paper. They analyzed the buying habit of the shopping users using MBA. Author conducted the evaluation of implemented MBA on minimarket X. They used well know Apriori method for discovering frequent set of items which are frequently appeared in transaction history as well as database. The itemsets those are exceeded threshold of minimum support value were selected as frequent itemsets. Such selected itemsets are further utilized to generate association rules followed by decoding. Every selected frequent itemset were able to generate association rules and hence compute the confidence using hybrid dimension association rules. The experimental results claimed that their implemented MBA can able to generate knowledge about kind of items those are frequently purchased in similar time frame by the customers using the criteria of hybrid dimension association rules. Their mining process outcomes shown the correlation among association rules and confidence those can be analyzed.

Andrej Trnka et.al. (2010)

In [9], author presented the technique of MBA implementation by using the Six Sigma strategy. The MBA is one data mining based framework. They used Six Sigma methods that use a two or three measurable strategies. Accompanied by use of Market Basket Examination to Six Sigma, they can enhance outcomes as well as modify Sigma implementation stage of process. Additionally, they introduced the General Rule Induction (GRI) algorithm in order to make union regulations within items in market basket. Author used the tool called web plot to claim the dependency within products. The final algorithm in examination was CS.O. That algorithm was used for produce lead established profiles.

XIE Wen-xiu et.al (2010)

In [10], they suggest inventive market baskets examination technique through mining union orders on items' interior attributes that are required via utilizing mechanical words segmentation strategy. They conducted this method in the company of dynamic dishes suggested method as well as approved via exploratory outcomes.

Similarly, there are number of other related works [10]-[15] presented to analyze the customer behavior and prediction to increase the sales of products. Such methods are designed under different categories like product recommendations, online shopping, group products prediction etc. However, optimization is still the challenging problem in existing methods for MBA framework.

Proposed Methodology

In this section, the proposed method architecture and algorithms are presented. Figure 2 shows the overall system design with two contributions such as data cleaning and MBA with Neural Network. We used FFNN classifier products sets prediction process which elaborated in section IV.

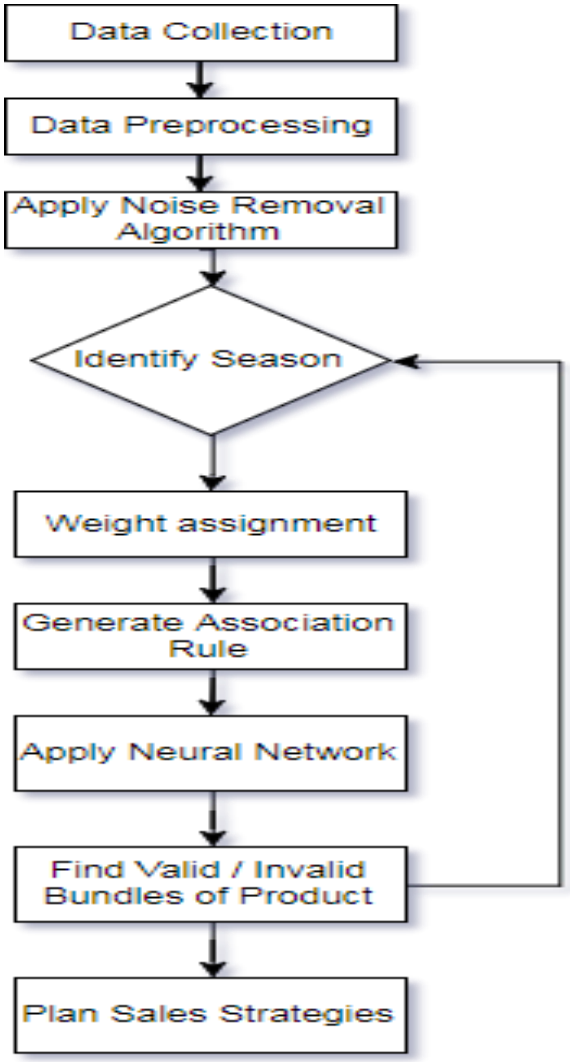

Fig. 2 Proposed System Design for MBA

Algorithm 1: MBANN (Market Basket Analysis with Neural Network Neural Netywork)

Input: D (Database which several items arrangement I1, I2, I 3, I4…In in store). C: possible items combinations;

Output: Valid (CV ∈ C) as well as Invalid (CIV ∈ C) combinations of products / items.

1. D1=ScanDataset(D);
2. CleanD=EHCleaner (D1)
3. Sigma= MBANN (CleanD)
4. if(sigma > 0)
5. sigma=1; //threshold value for valid combinations
6. return CV;
7. else
8. sigma=0; //threshold value for invalid combinations
9. return CIV;

In this algorithm, we used EHCleaner and MBANN Algorithms to remove noise and apply feed forward neural network to find sigma value respectively. If sigma value is positive then that combination consider as a valid combination else set as invalid combination.

Algorithm 2: Extended HCleaner Algorithm (E-HCleaner) for Data Cleaning

```
Input: Transaction set T
Result: Set of noisy objects N, Set of non-noisy objects P

for  i = 1 to ntrans do
        if T[i].contain(HTMLcharcters) then
                T[i].remove(HTMLcharacter)
        end
        if T[i].contain(StopWords) then
                T[i].remove(StopWords)
        end
        if T[i].contain(Expression) then
                T[i].remove(Expression)
        end
        if T[i].contain(Punctuation) then
                T[i].remove(Expression)
        end
        if T[i].contain(URLs) then
                T[i].remove(URL)
        end
end

HCS ← HypercliqueMiner(T); //HCS: Hyperclique Set
T[1...ntrans].assigned ← false;
len_hc ← size(HCS);
for  m = 1 to len_hc do
        for  n = 1 to ntrans do
            if ((!T[n].assigned) && contains(T[n],HCS[m])) then
                T[n].assigned ← true;
            end
        end
end
N ← {};
P ← {};
for  k = 1 to ntrans do
        if T[k].assigned then
                P ← P ∪ T[k];
        end
        else
                N ← N ∪ T[k];
         end
end
return N, P;
```

Algorithm 3: MBANN for Predictive Model

Input: CleanD Dataset, Weight: weight connected with input items;

Output: Sigma Value

```
1.  Search whole feasible combinations(CleanD)
2.  For items I1, I2, I3, I4…In €CleanD
3.      For every feasible combination c ę C do
4.          For every input item Ij in combination c
5.              sigma += input[j] * weight[j]; //summation function
6.              return sigma
7.          end for Ij
8.      end for c
9.  End for In
```

In this algorithm, we calculate sigma value using feed forward network. First calculate all possible combinations. Then for each combination c, calculate sigma value using weight value of each item that provide in weight[i] array and finally return sigma value that may be positive or negative.

Design of FFNN

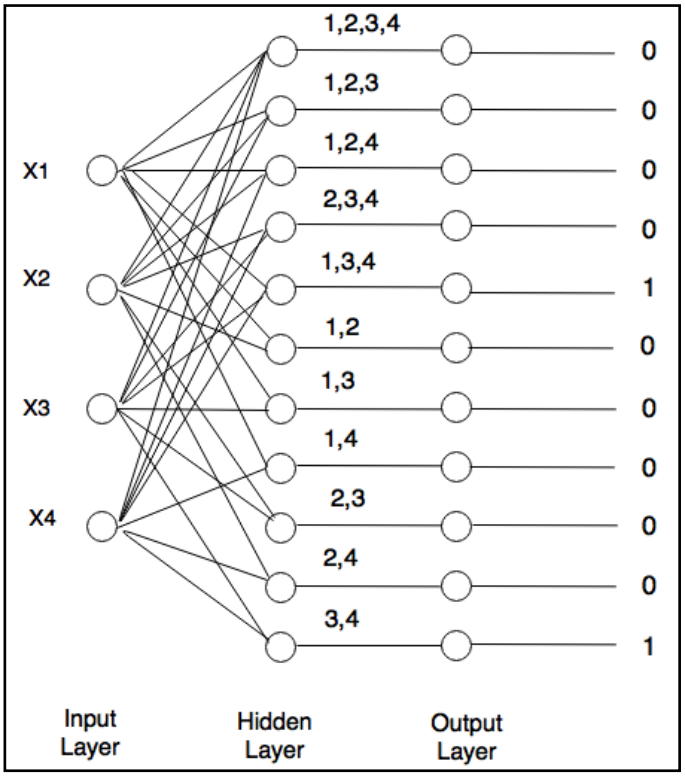

Fig. 1 Single Layer Feed Forward Neural Network

In above figure 1, x1, x2, x3 and x4 are products or items set of coustomer buying and it provides as a input to neural network. Now let these inputs represent Biscuit, Cold drinks, Tea and Fast food respectively. Suppose the values of the input items are x1=1, x=2, x3=3 and x4=4.In rainy season, the inputs items are x4= Fast food and x3=tea are buying mostly by consumer. So, we give the positive weight, say 1, to both. The another input items x2 are buying less regularly in rainy season, so these input items are denoted the negative weight.

Let's assign x2, high negative weight, say -5 and some item like x1 is purchase averaged frequently in rainy season so we assign 0 weights. Here applying the summation function. The initial attachment of neuron having (1,2,3,4), i.e., with connecting links from all the inputs x1, x2, x3, x4, the output will be: (0*1) + (-5*2) + (1*3) + (1*4) = -3.

The negative value, so it is allocated a threshold value is 0. So these are incorrect combination and these cannot be set in the basket. So we can apply the neuron to overall combination.
For second combination (1, 2, and 3) the output of the summation function will be: (0*1) + (-5*2) + (1*3) = -2
For third combination (1, 2, 4) the output of the summation functions: (0*1)+ (-5*2) + (1*4) = -6
For fourth combination (2, 3, and 4) the output will be: (-5*2) + (1*3) + (1*4) = -3
For fifth combination (1, 3, and 4), the output will be: (0*1) + (1*3) + (1*4) = 7
For neuron having 6th integration (1, 2) the outcome of summation consequence will be: (0*1) + (-5*2) = -10
After that for neuron having seventh combination (1, 3) the outcome of summation consequence will be: (0*1) + (1*3) = 3
Then for neuron having 8th mixture (1, 4), the output will be: (0*1) + (1*4) = 4.
For neuron having 9th mixture (2, 3) the output will be: (-5*2) + (1*3) = -7.
Then for neuron having 10th collaboration (2, 4) the output will be: (-5*2) + (1*4) = -6.
For neuron having 11th collaboration (3, 4), the output will be: (1*3) + (1*4) = 7.

Thus, after using summation technology, entire valid as well as invalid collaborations of client purchases have been prosperously established. Combinations which are give output greater than 0, which combination put as

valid combination. Hence in rainy phase, there are just 4 valid collaborations of client purchases: (1,3,4), (1,3), (1,4) and (3,4) and other all are invalid combination. Market basket can be rebuilt on depend upon these collaborations

Results and Discussion

The practical implementation of proposed methodology is done using Java programming language. We implemented our algorithms on extensive transaction dataset of online shopping manually. This section first presents the snapshots for implemented MBA framework for different steps of prediction. Then the performance evaluation results are presented with comparison with previous methods. The first step was applying proposed algorithm for data cleaning, after that the predictive methods are used on pre-processed dataset.

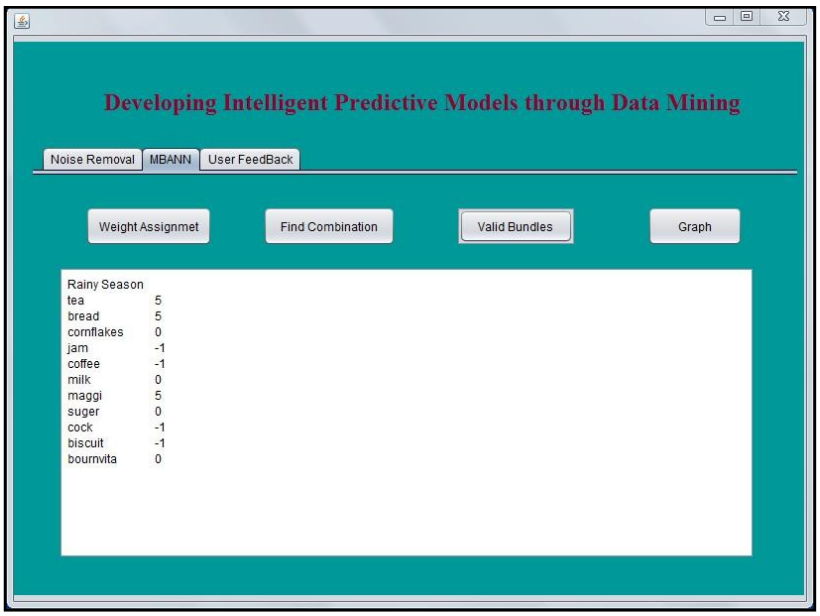

Fig. 3 Result of Weight Assignment

Figure 3 showing the outcomes of weight assignment in which weights are assigned to each product from dataset according the season (ex. Rainy). After weight assignment, our algorithm next computes the all possible combinations for all products as showing in figure 4.

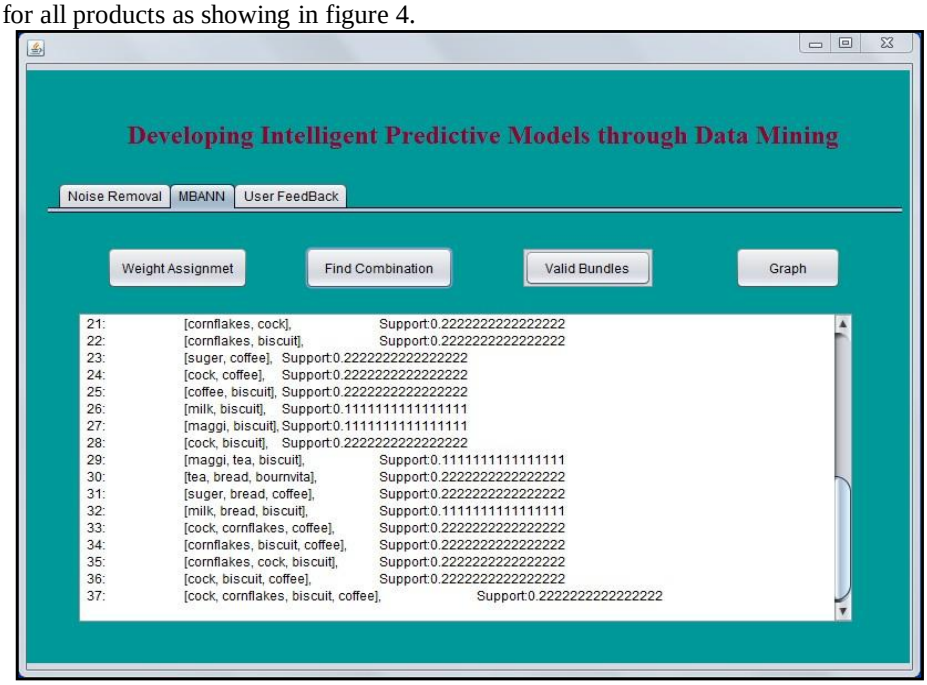

Fig. 4 Extraction of Possible Items Combination with Support

Finally valid bundles are generated using FFNN classifier showing in figure 5. Valid Bundle tab provides all possible valid bundles. These valid bundles are nothing but the predicted outputs for frequent buying items to increases the sales of shopping store. We have tested the different outcomes with different seasons.

In evaluation section, we measured the performance of previous methods those are studied in literature review such as case based prediction model [5], collaborative filtering [6] and existing method [7]. We modified the existing approach to improve the performance.

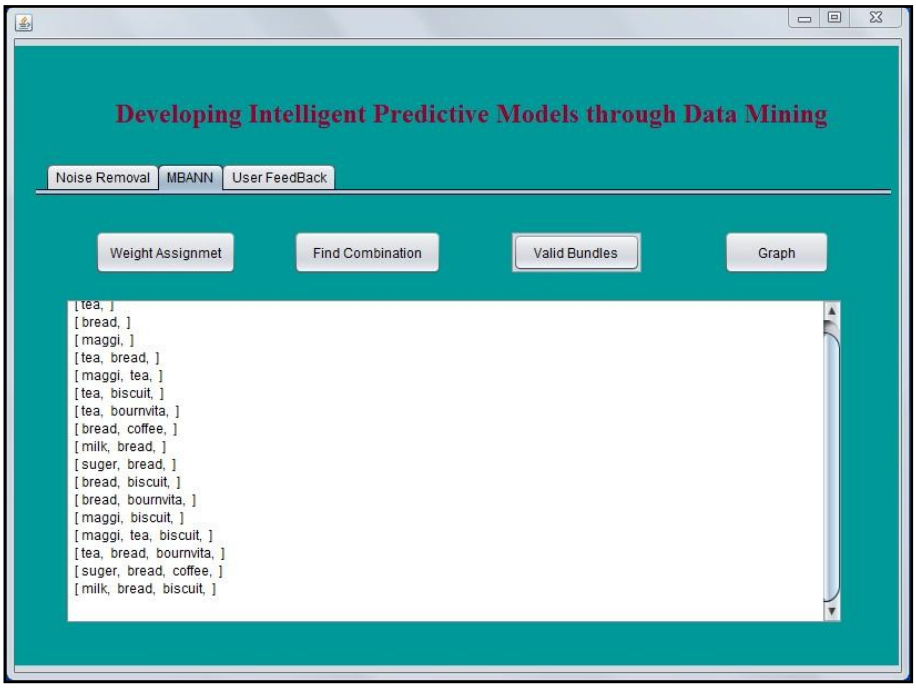

Fig. 5 Valid Bundles Prediction

The evaluation is done using three well know quality metrics such as precision, recall and f-measure rates. Figures 6, 7 and 8 are showing the comparative results for F-measure, recall and precision rates respectively. Y axis of all 3 graphs shows methods name and x-axis shows f-measure, recall, and precision respectively. The values are represented in tables.

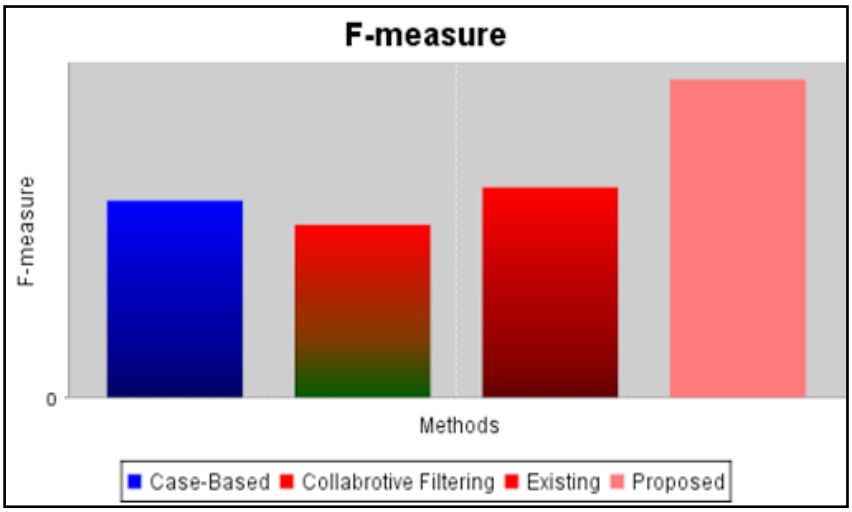

Fig. 6 Performance Analysis of F-measure

Figure 6 is showing that proposed approach showing the improved performance for F-measure as compared to other predictive models for MBA. There is significant improvement in proposed MBA predictive model in performances. Similarly the precision rate and recall rate is better as compared to existing techniques. The values of measured results are in the range of [0, 1].

Table 1 shows the actual outcomes for all three performance metrics F-measure, recall, precision and accuracy rates respectively. These results are satisfying the objective of our research designs and algorithms.

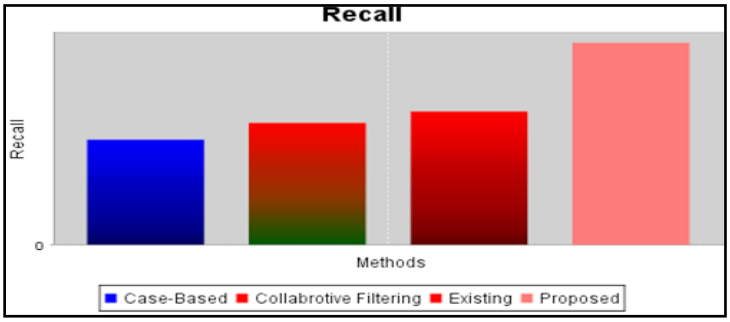

Fig. 7 Recall Rate Performance Analysis

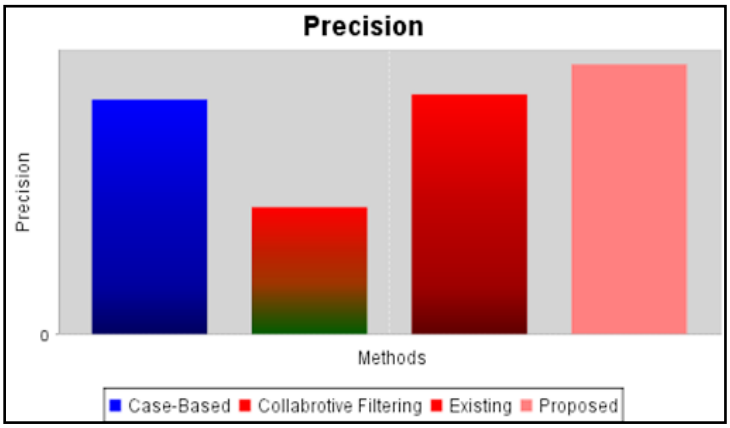

Fig. 8 Precision Rate Performance Analysis

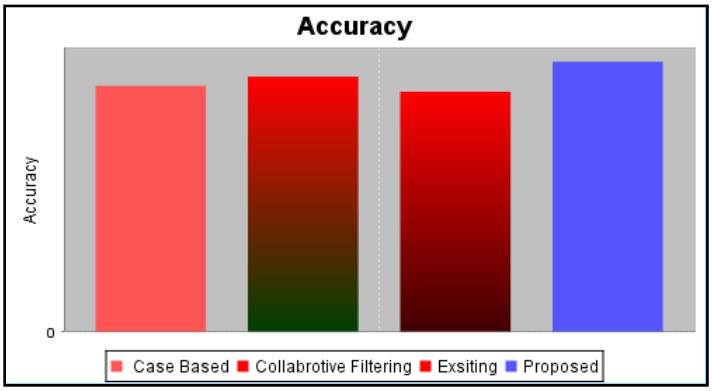

Fig. 9 Performance Analysis of Accuracy

TABLE 1: Comparison of F-MEASURE, RECALL, PRECISION AND ACCURACY

	Case-based [5]	**Collaborative [6]**	**Apriori based [7]**	**Proposed**
F1-Measure	0.25	0.22	0.27	0.40
Recall Rate	0.17	0.20	0.22	0.33
Precision Rate	0.44	0.24	0.45	0.51
Accuracy	0.82	0.85	0.80	0.90

Conclusion and Future Work

In this paper, we aimed to design optimized technique for MBA with goal of predicting and analyzing the customers buying behaviors. Earlier and recently there are few attempts that are made in data mining domain for MBA. However optimization is still main challenge for efficient MBA processing. First challenge is data cleaning as none of existing techniques even thought about possibility of raw data or noisy data in history of transactions. Secondary, the demands of the customers are continuously changing with respect to seasons and time. Also output of market basket analysis is totally dependent on time and seasons and so we need to perform it over and over. Hence there is need of dynamic and automatic MBA framework. In this paper, we introduced novel algorithms based on data cleaning; Apriori and FFNN solved said challenges. The performance of proposed approach was evaluated against existing methods. The results are significant and promising to claim efficiency of proposed approach. For future work, one can think on use of optimization methods like Genetic Algorithm as well as end user feedback to improve the quality of proposed work.

References

[1] J. Han and M. Kamber, "*Data mining concepts and techniques*", 2nd ed., Morgan Kaufmann, pp. 228-240,2006.

[2] S. Haykin, "*Neural Networks and Learning Machines*", 3rd ed., Pearson Education, pp. 10-23, 2009.

[3] Shepherd, "*Protein Secondary Structure Prediction with Neural Networks: A Tutorial*", http://www.biochem.ucl.ac.uk/~shepherd/sspred_tutorial/ss-nn.html, 2000.

[4] Warnia Nengsih, "*A Comparative Study on Market Basket Analysis and Apriori Association Technique*", 2015 3rd International Conference on Information and Communication Technology (ICoICT)

[5] Anna Gatzioura, Miquel Sànchez-Marrè, "*A Case-Based Recommendation Approach for Market Basket Data*", IEEE INTELLIGENT SYSTEMS, January/February 2015.

[6] Yi Zuo, Katsutoshi Yada, "*Using Statistical Learning Theory for Purchase Behavior Prediction via Direct Observation of In-store Behavior*", IEEE, Electronic ISBN: 978-1-5090-0713-4, 2015

[7] Anshul Bhargav, Robin Prakash Mathur, Munish Bhargav, "*Market Basket Analysis using Artificial Neural Network*", International Conference for Convergence of Technology – 2014

[8] Djoni Haryadi Setiabudi, Gregorius Satia Budhi, "*Data Mining Market Basket Analysis*' Using Hybrid-Dimension Association Rules, Case Study in Minimarket X"

[9] Andrej Trnka, "*Market Basket Analysis with Data Mining Methods Six Sigma methodology improvement*", 201O International Conference on Networking and Information Technology

[10] XIE Wen-xiu, "*Market basket analysis based on text segmentation and association rule mining*", 2010 First International Conference on Networking and Distributed Computing

[11] H. Sorensen, "*The science of shopping*" Marketing Research, vol. 15, pp. 30–35, 2003.

[12] J. S. Larson, E. T. Bradlow, and P. S. Fader, "*An exploratory look at supermarket shopping paths*" International Journal of Research in Marketing, vol. 22, no. 4, pp. 395–414, 2005.

[13] S. K. Hui, E. T. Bradlow, and P. S. Fader, "*Testing behavioral hypotheses using an integrated model of grocery store shopping path and purchase behavior*" Journal of Consumer Research, vol. 36, no. 3, pp. 478–493, 2009.

[14] K. Yada, "*String analysis technique for shopping path in a supermarket*" Journal of Intelligent Information Systems, vol. 36, no. 3, pp. 385–402, 2011.

[15] C.-C. Chen, T.-C. Huang, J. J. Park, and N. Y. Yen, "*Real-time smartphone sensing and recommendations towards context-awareness shopping*" Multimedia Systems (2015) 21: 61. https://doi.org/10.1007/s00530-013-0348-7