Comparative Analysis of XLMiner and Weka for Association Rule Mining and Clustering

A.M. Khattak, A.M. Khan, Tahir Rasheed, Sungyoung Lee, and Young-Koo Lee

Department of Computer Engineering, Kyung Hee University, Korea
{asad.masood,kadil,tahir,sylee}@oslab.ac.kr, yklee@khu.ac.kr

Abstract. Retaining a customer is preferred more than attracting new customers. Business organizations are adopting different strategies to facilitate their customers in verity of ways, so that these customers keep on buying from them. Association Rule Mining (ARM) is one of the strategies that find out correspondence/association among the items sold together by applying basket analysis. The clustering technique is also used for different advantages like; recognizing class of most sold products, classifying customers based on their buying behavior and their power of purchase. Different researchers have provided different algorithms for both ARM and Clustering, and are implemented in different data mining tools. In this paper, we have compared the results of these algorithms against their implementation in Weka and XLMiner. For this comparison we have used the transaction data of Sales Day (a super store). The results are very encouraging and also produced valuable information for sales and business improvements.

Keywords: Association Rule Mining, Clustering, Weka, XLMiner.

1 Introduction

Now-a-days, retaining old customers is preferred more than attracting new customers. Business organizations are adopting different strategies to facilitate their customers in verity of different ways, so that these customers keep on buying from them. Association Rule Mining (ARM) [AIS93] is one of the strategies that have two fold advantages to the business organization after applying the basket analysis. 1) It helps customers to get all the related items from one place and that save their time from visiting different places of the store. 2) It helps organization in more selling of items by placing items closer that are sold together. Different business organizations around the world have used basket analysis technique; among these; *Wal Mart*[1] is the most famous example. Clustering technique is used for classifying data based on some of its characteristics into different classes that eventually help users/organizations to further smoothen their business process. Clustering results in different advantages for business organizations; 1) to recognize the class of most sold products, 2) classifying customers based on their buying behavior and their power of purchase, 3) classifying customers

[1] http://walmartstores.com/

D. Ślęzak et al. (Eds.): DTA 2009, CCIS 64, pp. 82–89, 2009.

arrivals in different time slots based on customers arrival time, and 4) identifying item(s) source for major trade.

Considering high dimensional data with noise and outliers, ARM and Clustering is a challenging task especially when data is very huge and complex. As discussed above, different researchers have provided different algorithms for ARM [LiH04] and Clustering [ESK03 and KaR90] that helps user to properly and efficiently achieve their objectives. The Apriori [LiH04] algorithm is used for ARM; it had a problem of candidate set generation. This problem was removed, so the new improved Apriori algorithm reduce the time of scanning candidate set. It uses the hash tree to store the candidate sets that facilitate in solving the frequent set counting problem and is now more optimized based on time factor. The same way K-Means [KaR90] algorithm is used for Clustering of data based on the parameter(s) specified.

In this paper, we have compared the results of two Data Mining tools i.e. XLMiner [XLMiner] and Weka [WiF05] for ARM and Clustering with Apriori and K-Means algorithms respectively. We have tested these algorithms of both the tools using daily transaction data from Point of Sale system of a super store *Sales Day (SD)*. We tested the Apriori and K-Means algorithms from both Weka and XLMiner on data of year 2007 of SD. By varying the parameters (i.e. support and confidence) for these algorithms; we got very interesting results for ARM discussed later. The same way, we also have tested both the tools with K-Mean algorithm for clustering of the data to identify the different classes of items sold of particular amount and users. The customers are clustered based on their buying power, time and power of buying, most frequent customers, and transaction with amount of transactions that helps in focused advertisement based on customer arrival time and their buying behavior.

Rest of the paper is organized as follows; Section 2 is related to preprocessing of Sales Day data to be used by the algorithms. Section 3 presents experimental results after applying both ARM and Clustering. In Section 4 is based on conclusions and future directions.

2 Data Normalization

Before applying the algorithms on data, first we need to normalize the transaction data for the algorithms to work on. The daily transaction data of Sales Day (SD) store as shown in Figure 1 is in organization required format, which is a high dimensional and complex data that is not useable by the algorithms. For this reason, we have first converted the data from the organization required format to the format required to be used for experimentation. Figure 1-A shows the schema of SD Point of Sale (PoS) system where the SD data is very much redundant as clear from Figure 1-B. For every item sold in a single transaction, there is a complete row for that item and repeating the same order data again and again. We have developed a MS-Visual Basic 6.0 application using MS-SQL 2000 queries to translate the data in to a single row (pivoting) for every ordered transaction as shown in Figure 1-C. We have worked on the transaction data of year 2007 and tested both the tools for ARM and Clustering. The transaction had a variability in number of items contained in them e.g., a person may buy only a milk or a snack pack (i.e. only one item) but a transaction may contain a whole variety of items that range from daily use to occasionally used items

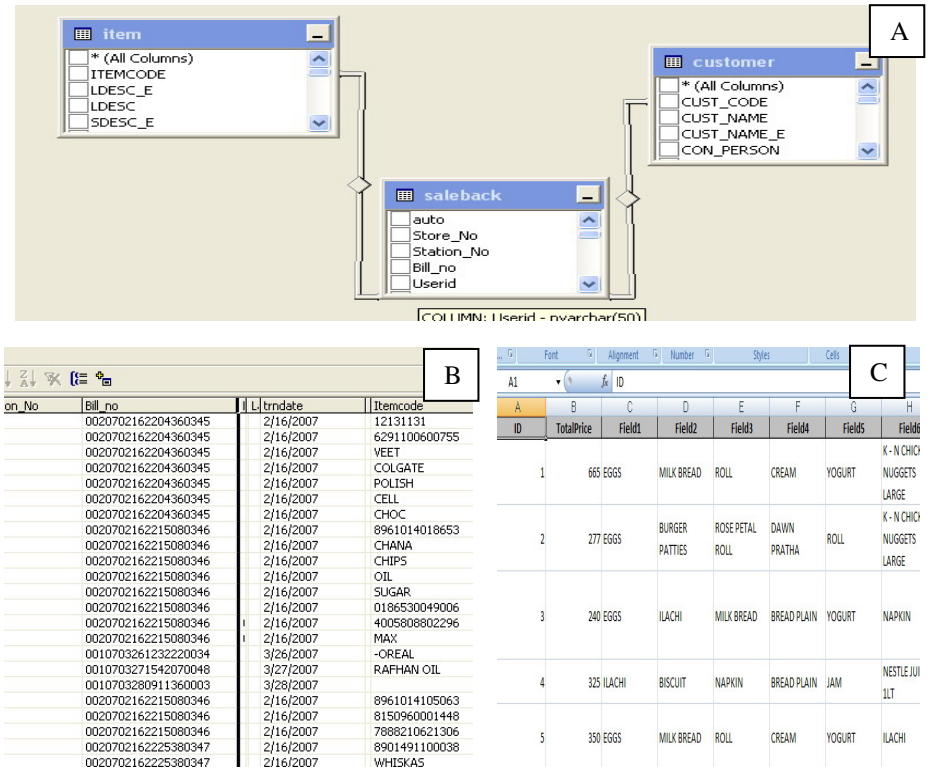


Fig. 1. (A) shows the original structure of transection data storage in Sales Day Point of Sale system, (B) shows that for all the items sold in a transection is having a separate row entry, (C) shows the converted transections from multiple rows for one order to a single row.

that make the item set up to 60-80 products in it. We have fixed the number of items sold in a transaction to 12 items per transaction and any transaction having items more or less than 12 items are discarded.

3 Experimental Results

Here we discuss in detail the comparative study of both the tools (i.e. Weka and XLMiner) for ARM and Clustering results over the transaction data normalized in the previous section. We have divided this work into two sections where first one focus on ARM and the second is focusing on Clustering. In Section 3.1 we will discuss ARM using both Weka and XLMiner and after that in Section 3.2 we will discuss Clustering results. The results from both tools will be compared in these two subsections.

3.1 Association Rule Mining (ARM)

ARM finds interesting associations and/or correlation relationships among large set of data items. It infers attribute value conditions that occur frequently together in a given

dataset e.g. Market Basket Analysis. Our goal is also to mine the Association Rules among data items from the transactions data of a super store Sales Day. The association rules provide information in the form of "if-then" statements where these are probabilistic in nature.

Among the different ARM algorithms available like; 1) Apriori, 2) Filtered Associations, 3) Predictive Apriori, and 4) Tertus, we choose to implement Apriori despite its multi scan drawback but the rules generated by Apriori are the most appropriate and finer granulized. To start working with Apriori for ARM, we have specified the environment variables as: 6994 instances of transactions with 12 attributes. The minimum support for ARM is set to 0.6 while minimum confidence is 0.9 with 20 numbers of cycles performed. Based on this input data for Apriori in both Weka and XLMiner, the association rules are mined (see Figure 2).

Results from both the tools depict same rules, while the representation of rules in these tools is different. For instance as shown in Figure 2-A, the confidence for CAKE to be purchased by the customer is 100 % if that customer is going to purchase BISCUIT, MILK, MILK BREAD, and NOODLE. The same rule is also represented in Figure 2-B of Weka where all the items are separately mapped with all the other items, and it also gives the confidence as 1 (which is equivalent to 100 %).

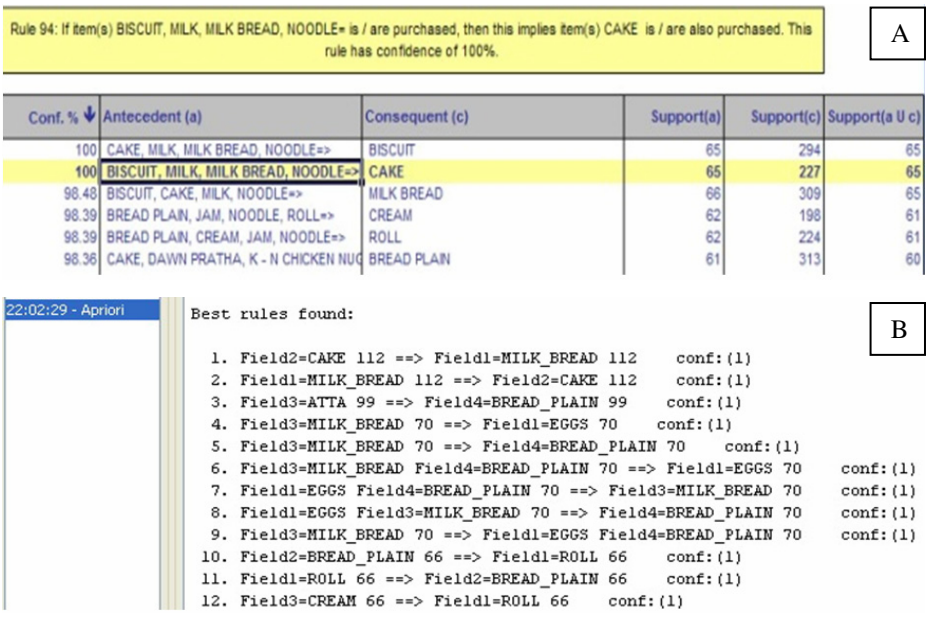


Fig. 2. (A) Shows the association rules identified using XLMiner. The yellow strip above display the complete rule selected in the table. (B) Shows the association rules mined using Weka.

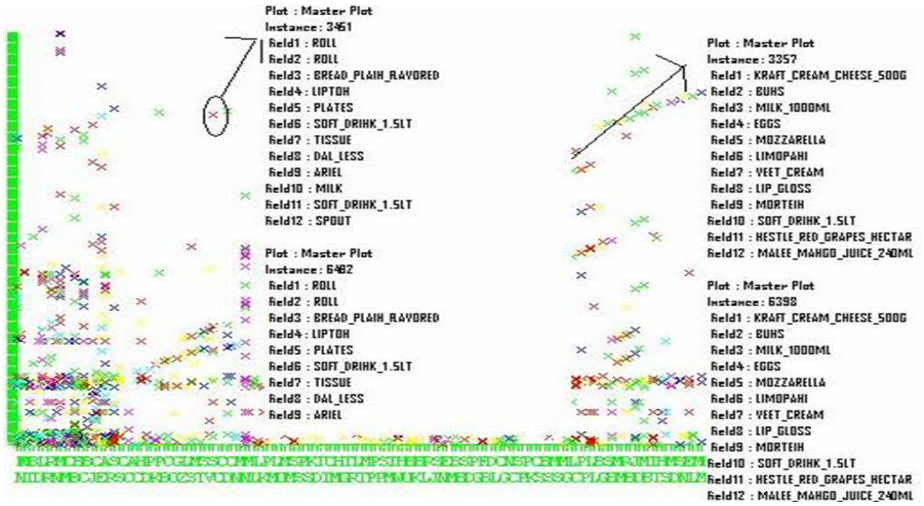


Fig. 3. Visualization of association rules using Weka

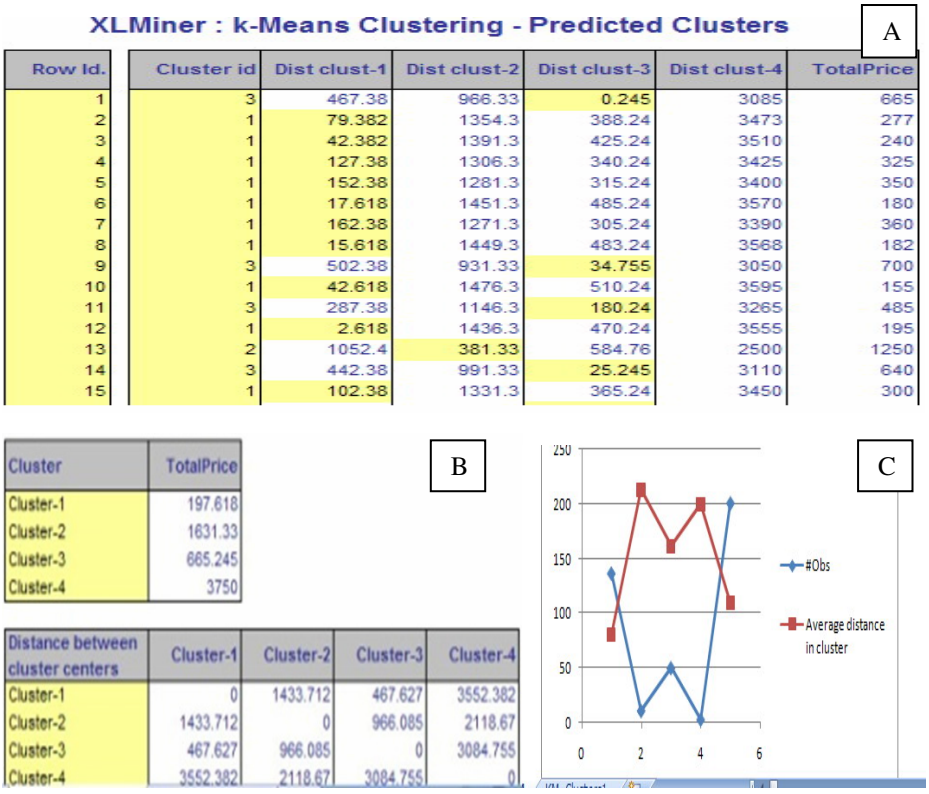
From the above Figure 3, it can be seen that how the associations are distributed over a plot. The circle points indicate that how the two transactions are related on the confidence level and products (items) occurring in their basket. This plot depicts the transactions with their respective basket items.

The associations/correspondence among sold items is one of the most useful source/results for the business organization. These associations are used by business organizations to resort their products in a way to place the most frequent sold items together. This also facilitates customers in quick checkout. One other strategy is to place the most sold items in different places. In this case the customer will have to visit different places in the store and will have a look at other different items available, that will increase their probability of been purchased by the customers.

3.2 Clustering

Clusters are often computed using a fast, heuristic method that generally produces good (but not necessarily optimal) solutions. The K-Means algorithm is one such method. We applied K-Means Clustering algorithm on the transaction data using XLMiner and Weka. In XLMiner, to do clustering, we enter the data range that needs to be processed and move the variables of interest to the selected variables box. Here, it is visible in Figure 4-B that the numbers of clusters are 4 and the attribute selected for the clustering is *TotalPrice* of transactions. It also represents the mean of clusters. Figure 4-A shows different transactions that are classified in different clusters, while Figure 4-C represents the objects (clusters) and the inter-cluster distance among them.

Weka uses the centroids positions for calculation of clusters. We have tested the data with K-Means using Weka and got 3 clusters as number of clusters depends on choice of initial centroids, choice of distance measure, and stopping criterion that we defined.



XLMiner : k-Means Clustering - Predicted Clusters

A

Row Id.	Cluster id	Dist clust-1	Dist clust-2	Dist clust-3	Dist clust-4	TotalPrice
1	3	467.38	966.33	0.245	3085	665
2	1	79.382	1354.3	388.24	3473	277
3	1	42.382	1391.3	425.24	3510	240
4	1	127.38	1306.3	340.24	3425	325
5	1	152.38	1281.3	315.24	3400	350
6	1	17.618	1451.3	485.24	3570	180
7	1	162.38	1271.3	305.24	3390	360
8	1	15.618	1449.3	483.24	3568	182
9	3	502.38	931.33	34.755	3050	700
10	1	42.618	1476.3	510.24	3595	155
11	3	287.38	1146.3	180.24	3265	485
12	1	2.618	1436.3	470.24	3555	195
13	2	1052.4	381.33	584.76	2500	1250
14	3	442.38	991.33	25.245	3110	640
15	1	102.38	1331.3	365.24	3450	300

B

Cluster	TotalPrice
Cluster-1	197.618
Cluster-2	1631.33
Cluster-3	665.245
Cluster-4	3750

Distance between cluster centers	Cluster-1	Cluster-2	Cluster-3	Cluster-4
Cluster-1	0	1433.712	467.627	3552.382
Cluster-2	1433.712	0	966.085	2118.67
Cluster-3	467.627	966.085	0	3084.755
Cluster-4	3552.382	2118.67	3084.755	0

C

Fig. 4. (A) Shows classification of transactions in clusters, (B) shows number of clusters and their distance, and (C) visually represent inter-cluster distance

These are overlapping clusters that are obtained by the range of the mean and standard deviation specified. By analysis of the visual plot obtained and shown in Figure 5. It is clear from Figure 5 that the clusters are roughly distributed. The color distribution for clusters is; Blue: Cluster 0, Green: Cluster 1, and Red = Cluster 2. The circle points show that at that point which cluster value it is.

3.3 Analysis of Results

The ARM and Clustering work conducted here in this paper is basically for the purpose of comparative analysis of Weka and XLMiner with Apriori and K-Means algorithms. We have tested both the tools and for ARM they gave exact answers but during different experiments we performed for Clustering generated different clusters. Beside this comparison objective, we also got some very interesting results based on the transactions data. We got the most frequent sold items in a basket that helped the organization in re-organizing the sale strategy for these items to improve their sale. As shown above, in Clustering we have also classified the transactions data on different parameters like; age group of customers, age group plus purchase power, most sold items, time of high customer traffic, and time of high purchase power

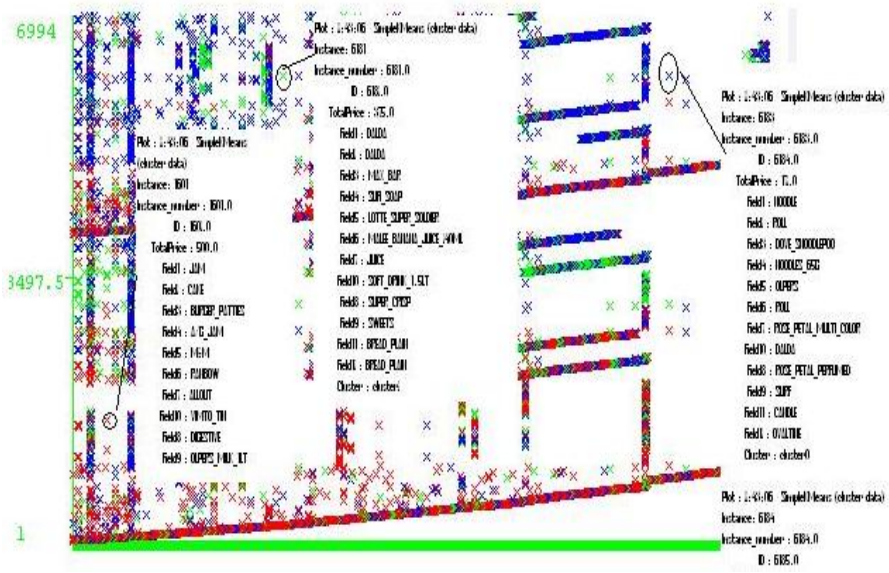


Fig. 5. Visualization of clustered transactions using Weka

customers arrival. These classified (Clustered) transactions are used by the business organization to great advantage. These results are very useful, for instance; with the help of cluster i.e. Time of high purchase power customer arrival, the organization can lunch new high cost products for these customers. Knowing the arrival time of particular type of customers can also be used for focused advertisement of product of interest to the arriving customers. These results also avoid the out-of-stock situation as it gives information about most sold items.

4 Conclusions and Step Ahead

Association Rule Mining and Clustering is a well established area of data mining. These are used for extracting some hidden information from a huge repository of raw data. In this paper we used these two techniques with their Apriori and K-Means algorithms implemented in Weka and XLMiner to analyze the trend of sale at a super store Sales Day. We have compared these algorithms by using Weka and XLMiner over Sales Day data and got very encouraging results that not only satisfy the implementation of these algorithms in both the tools as same but also support the business organization for customer support and future extension in their business. We are planning to extend our work to different tools for more algorithms and use the results to business advantages.

Acknowledgement

This research was supported by the MKE (Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised

by the IITA (Institute of Information Technology Advancement)" (IITA-2009-(C1090-0902-0002)) and was supported by the IT R&D program of MKE/KEIT, [10032105, Development of Realistic Multiverse Game Engine Technology].

This work also was supported by the Brain Korea 21 projects and Korea Science & Engineering Foundation (KOSEF) grant funded by the Korea government(MOST) (No. 2008-1342).

References

[AIS93] Agrawal, R., Imielinski, T., Swami, A.: Mining Associations between seta of items in Massive Databases. In: proc. Of the ACM-SIGMOD 1993 int'l conf. on Management of Data, Washingtom D.c USA, pp. 207–216 (1993)

[ESK03] Ertoz, L., Steinbach, M., Kumar, V.: Finding Clusters of Different Sizes, Shapes, and Densities in Noisy. In: SIAM International Conference on Data Mining, (Feburary 20, 2003)

[KaR90] Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley Series in Probability and Statistics. John Wiley and Sons, New York (1990)

[LiH04] Liu, X.W., He, P.L.: The research of improved association rules mining Apriori algorithm. In: Proceedings of 2004 International Conference on Machine Learning and Cybernetics, August 26-29, vol. 3, pp. 1577–1579 (2004)

[WiF05] Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)

[XLMiner] Data Mining Add-In For Excel, `http://www.resample.com/xlminer/`