An Application of Inductive Learning for Mining Basket Data

İlhan Uysal **H.Altay Güvenir**

Department of Computer Engineering and Information Sciences,
Bilkent University,06533 Ankara, Turkey
{uilhan,guvenir}@cs.bilkent.edu.tr

Abstract. The development of bar-code technology provided accurate and large market databases for researchers who deal with datasets. Since the data is large both in dimension and size, most exploratory analysis techniques of statistics are not appropriate for such tasks. In this paper, we describe a high-level algorithm, and the application of it on a large basket data, extracted from the database of a big supermarket company. The algorithm have two consecutive steps. Each step is a different popular machine learning method: clustering and classification. In this application, we used KMEANS clustering algorithm and C4.5 classification program respectively. At the end of the application we come up with a set of items that can be employed for promotion. By promotion we aim to increase number of costumers that make their weekly or monthly shopping, which refer to full baskets among transactions.

1 Introduction

The synergy between two important fields, machine learning and statistics that use datasets to analyze and to learn concepts respectively had led the emerging of a new field, called data mining. The new field, which is on the interface of these two disciplines and databases, emerged for two main reasons. The improvements in the database technology yielded us huge amounts of data storage. Most of the algorithms and techniques developed so far in statistics and machine learning do not overcome the memory, I/O, and computational complexity problems emerged, when the dataset of statistics, and the training set of machine learning are such large databases. The other reason is that, new techniques and algorithms are necessary to mine the databases to discover important strategic knowledge to help decision making. Most of today's databases are still waiting to be processed for knowledge, instead of being left as large archives. A description of the field and the process of knowledge discovering in databases are given in [7], and some applications of data mining on scientific and business databases are described in [3, 8].

The development of bar code technology, and its utilization in markets has opened a new application field for data miners. The market databases, so called basket data, filled by the transactions of this technology is large enough to obtain significant results and accurate enough since the transactions are recorded electronically with minimum user intervention. The most popular research topic of data mining on basket data is extracting association rules, which is described first in [1].

We describe the Full Basket Data Mining Application (FBA) in this paper for mining basket data. The application is simple and useful for the managers of supermarkets, in order to increase the sales and profits of their firms. It has a high level algorithm that have two consecutive steps, clustering and classification, respectively. After these steps, we come up with a set of items to be promoted in a market.

Some important tasks of a supermarket management include decisions on items to sell, prices, amount of items to be purchased from producers, items to be promoted, and how to place the items on shelves so that the profit can be maximized. Association rules on basket data is useful for managers for deciding on the last two subjects stated above: They may put the associated items on the same shelves, close to each other, and they can apply promotion on one of the associated items or on all of them together [1]. Among tens of thousands of different items in a supermarket, data mining is useful in order to select a small number of items to be promoted.

FBA helps the managers to decide on items to be promoted in a different way. It does not find the association between items, instead it finds the association between items and large (full) baskets. Some people go these markets to buy their monthly or weekly needs (full basket) and some people to buy a couple of items (empty basket). The purpose of the promotion in this sense is not only to increase the total number of baskets, especially to increase the number of full baskets. The investigation on our market data shows that, even the number of full baskets is much smaller than (less than 15%) empty baskets, the total profit of them is larger than the empty baskets. If we can increase number of baskets 4% with full baskets, the profit of the firm increases 10%, which is a significant increase. The application do not produce spurious results, which is the problem for most data mining applications to be overcome in the evaluation or visualization phase of knowledge discovery and data mining (KDD) process. Consequently the application finds the result of the query stated below:

```
Find minimum number of items that are significantly most frequent
in the transactions of full basket category.
```

The overview of the application is described in the next section. In Section 3 and Section 4 application of clustering and classification on the data to obtain the result of above query is described, respectively. In Section 5, the results are evaluated. An optional clustering phase for very large item sets is discussed in Section 6, and the paper is concluded in Section 7.

2 The Overview of FBA

The algorithm of FBA includes two main steps: Clustering and classification. By clustering, we form the two clusters, full baskets (FB) and empty baskets (EB). We label the transactions with their category by appending an extra attribute, indicating whether they are full or empty baskets. Then, we run a concept learning algorithm, where the features are the items, and training set is the whole set of baskets, and the classes are the clusters found in the previous step. The attributes have Boolean values, either false if that item is not in the basket, or true if it is in the basket. If there are more than one from the same item in a basket, we simply accept it as one item.

The employment of a classification algorithm here is not for the purpose of classifying new coming instances whose classes are unknown. Rather it is applied to assign an ordering to the attributes so that, the attributes which have significant influence on the classes of baskets can be determined. The attributes or items that significantly describe the FB class are what we search. Other items are irrelevant for promotion.

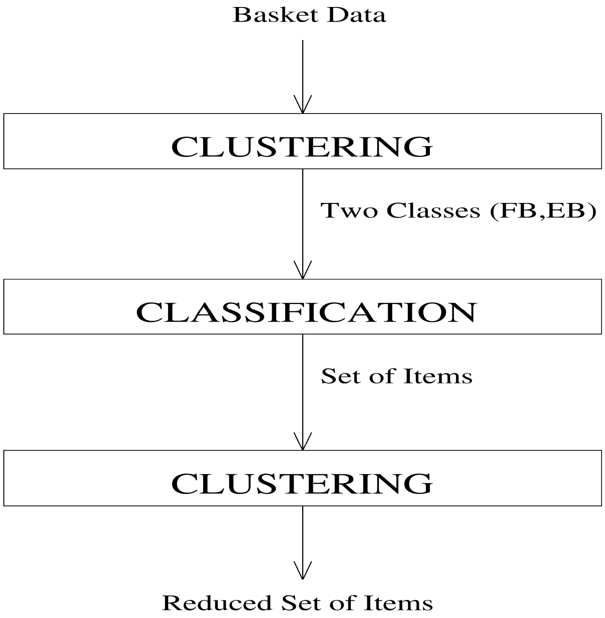

Figure 1. FBA Overview

After completing the second step, a set of items is extracted in a descending order. If the number of items in the set is too large to deal with manually, an optional clustering can be applied to them, to extract significant ones. Another approach is to remove the items under a given threshold ordering value. In our application the resulting number of items is small enough to evaluate manually, so that, we did not apply this optional phase. The overview of the application is shown in Figure 1.

3 Clustering

Clustering is a common technique used in various disciplines under different names. In machine learning it is referred to unsupervised conceptual learning, in statistics it is referred to cluster analysis, and it is named as Q analysis, typology and numerical taxonomy in some other disciplines. Clustering is to partition data into clusters or groups in the sense that all objects (instances) in the same group are similar to each other and not similar to the objects in other groups. Many algorithms and methods about clustering have been proposed in machine learning and statistics. The most common similarity criterion in clustering is the distance. That is why, most of the algorithms are distance-based. And lots of different distance measures can be employed in clustering according to the application such as the sum of squared distances from cluster centers.

In this application, we partition the data into clusters such that the resultant clusters will include either full baskets or empty baskets. By employing clustering we want to categorize all relevant baskets as full baskets, and put all less relevant baskets and outliers into empty basket group. For example a basket having only one item whose cost is very high (e.g. washing machine) or a basket full of different items but whose cost is low is categorized as empty. The baskets that have large number of items and high total costs are categorized as full.

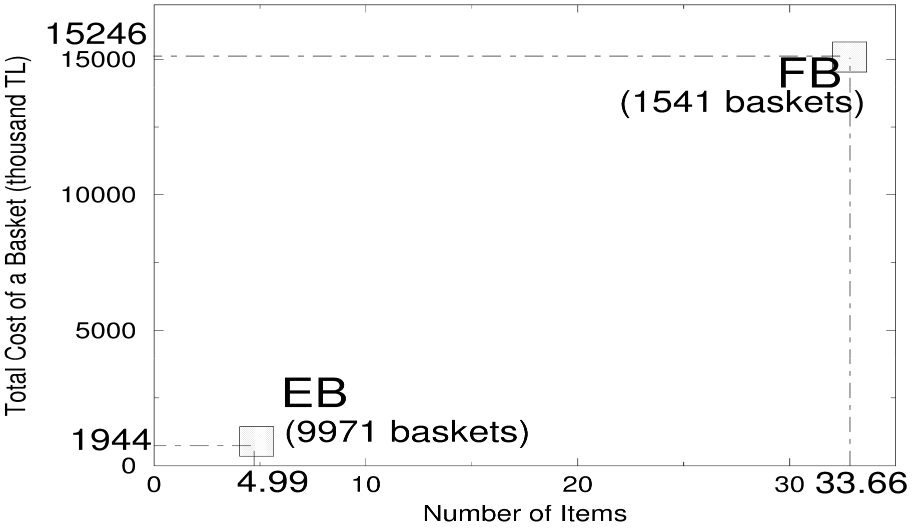

Figure 2. Cluster Centers of Empty and Full Baskets

In order to reach such a categorization, we use two attributes in clustering: Number of items in the basket and total cost of the basket. We used a partitioning clustering algorithm, KMEANS [6, 9], for this purpose. Given the number of clusters a priori, KMEANS algorithm partition the data by an iterative procedure, simply by exchanging instances between clusters. The iteration continues until reaching a minimum distance measure between instances at all clusters. This distance in our application is Eucledian distance. Even though the complexity is exponential, practically it finds results at a short time, with a small memory requirement, which is proportional to number of instances, N. On the other hand, for hierarchical clustering algorithms [6], memory requirement is proportional to N^2, which is not applicable to our data on most computers.

Before clustering the data set both attributes are standardized, by dividing attribute values to their standard deviation. This enables equal contribution of attributes in the computation of distance. A statistical test (hypothesis test) on the resulting clusters shows that the mean values of the formed clusters are significantly different, which proves the important difference between clusters. The number of baskets in clusters and their cluster means are shown in Figure 2.

4 Classification

Classification is one of the most common techniques used in data mining [7, 2, 10]. It is used in machine learning for supervised learning tasks. The aim is to correctly classify new cases according to information extracted from the previously recorded and correctly classified cases. Besides such prediction tasks, interpretation of data by using extracted knowledge is also important for most applications. To make interpretation of the data by using the extracted model is the way most exploratory data analysis methods of statistics work. But statistical methods are not appropriate for this application because of large dimension and large size of the basket data. We will use classification in order to interpret the effect of features (items) on the target, where the target is a categorical feature with two different values or classes (EB, FB) determined in the previous clustering step. The classification methods that form models and enable the interpretation of individual features are appropriate for this step. According to the knowledge about features, we will extract a set of features or items among all features. This process is analogous to feature selection. But, we do not prefer feature selection methods here in order to obtain much smaller set of items and eliminate most spurious or redundant items that appear in full baskets together.

We employed decision tree learning method, which forms models represented by trees. This method is preferred since it allows interpretation such that the important features are placed at the upper levels in the tree. Another important property of this method is that it eliminates redundant or associated features to appear together in the constructed tree, so that, the number of features in the tree is very small when compared to size of the dimensions and especially after applying a pruning. The classification algorithm used in the application is described in the next subsection.

4.1 Decision Tree Learning Algorithm

Using decision trees as a target concept or target function in classification is introduced in [12]. Several decision tree learning algorithms have been implemented, such as ID3, ASSISTANT, and C4.5 [13]. A Decision tree learning algorithm classifies the instances by searching the decision tree starting from the root node, where each node in the tree represents a discrete valued attribute, and each branch is a value of it. The search continues until constructing all leaf nodes, where classes are specified. During the construction of tree, to search all possible trees exhaustively to find one that describe the training data best is not feasible since there are too many possible trees. Instead, a greedy method is employed to find the decision tree. Algorithm starts with an empty tree, and constructs it gradually by making a hill-climbing search using the information gain measure as evaluation function, and it does not backtrack. The attributes that have high information gain are placed close to the root. The inductive bias of the algorithm is that smaller trees are preferred to larger ones. The information gain measure employed in the algorithm is given by the following formula:

$$Gain(S, A) = Entropy(S) - \sum_{v \in values(A)} \frac{|S_v|}{S} Entropy(S_v) \tag{1}$$

where $Entropy(S)$ is defined as

$$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2 p_i \tag{2}$$

where S is the set of examples, p_i is the proportion of S belonging to class i, $Values(A)$ is the set of all possible values for attribute A, and S_v is the subset of S where attribute A has value v [11].

We used the decision tree learning program, C4.5 [13] in our application. We gain three main advantages by using this algorithm in this application. First, since statistical significance of items in detecting the class of instances are important for us, some instances may be misclassified. Misclassification of some instances are not much important for us, as long as significantly large number of them are assigned to true classes. Even though market datasets are accurate and do not have noise, we may refer to some cases as noise. As an example, if 99% of an item is sold in the full baskets and the remaining in the empty baskets we may classify instances including that item as full. The algorithm is robust in this sense, since a statistical measure, information gain, is used in the construction of the tree. As a consequence of the first advantage and bias of the algorithm, some redundant items are eliminated. This is the second benefit which enable us to produce smaller set of items. This is the main reason why we prefer classification algorithm instead of a feature selection method. The third advantage comes from pruning which is applied to most decision trees. By applying pruning to the tree, we both obtain much smaller tree and at the same time avoid possible overfitting problems on the data.

4.2 An Example Basket Data

A sample data having 11 items with 14 instances and its decision tree, which is constructed by C4.5 decision tree induction program, are shown in Figure 3. The notation I in the Figure 3 and Figure 4 represent items (features) in the baskets. If an item is present in a basket or transaction, it is denoted by value 1 meaning *positive*. If it is absent a value 0 is used to denote it. We simply disregard 0's for clarity in Figure 3. The instances having more than two items are classified as positive. Since the values are binary, the resulting decision tree is a binary tree, as shown in Figure 4.

Note that, I_8 is a redundant item since it is covered by I_6, and it is not seen in the decision tree even though it has the second highest information gain after I_6. I_3 and I_4 are also redundant items covered by I_6. Also, one instance is misclassified on the positive leaf of I_{10}.

I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	I11	C
				1		1			1	1	+
			1	1	1		1				+
	1				1	1	1	1			+
		1	1	1	1		1			1	+
					1						+
1					1		1				+
1						1			1	1	+
1				1							+
	1									1	-
				1		1					-
1									1		-
						1		1			-
			1							1	-
									1		-

Figure 3. Sample Basket Data. A 1 represent the existence of the corresponding item in a basket (represented with a row). A (+) represents a full basket and a (−) represents an empty basket.

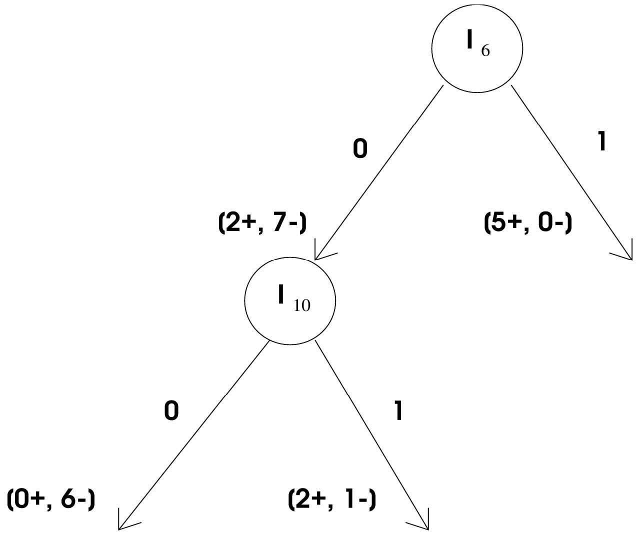

Figure 4. Decision Tree. The terms in parentheses shows number of full and empty baskets, respectively, in a region represented with a leaf node or subtree.

The items having positive leaves on their branches valued with 1, are our initial results in the FBA. In our sample data I_6 and I_{10} are our initial results. The initial results will be filtered further according to a threshold information gain measure, or by an optional clustering at the end.

5 Results on Real Data

We have executed the C4.5 program on the training set having approximately 10000 randomly selected transactions and a test set having approximately 1100 remaining transactions across 4907 attributes (items). After extracting a decision tree with 1163 nodes, we have applied pruning and the program produced a pruned tree having 501 nodes. The upper part of the tree is constructed by the items seen only on the full basket, where their branches are valued with 1, and end on a positive leaf. The most significant 10 items among the resulting set are listed in Table 1.

The measured accuracy of the decision tree (before and after pruning) is shown in Table 2. In the training data 97.3% of the instances are categorized accurately. With the pruned tree this ratio decreases to 95.7%.

The accuracy on the test data is 88%, and it increases to 89% after pruning. The improvement after pruning shows that some overfitting on the decision tree is avoided by pruning.

Item Name	Amount
Jam of Quince	400g
Softener-1	3000g
Softener-2	1000g
Shampoo	200ml
Soup	1 Package
Flour	4000g
Vegetable Oil	2lt
Softener-3	1000g
Cheese	1 Package
Soap	200g

Table 1. Some of the Significant Items in the Full Baskets

Training Data						Test Data					
Before Pruning			After Pruning			Before Pruning			After Pruning		
Tree Size	Number of Errors	Error Ratio	Tree Size	Number of Errors	Error Ratio	Tree Size	Number of Errors	Error Ratio	Tree Size	Number of Errors	Error Ratio
1163	285	2.7%	501	445	4.3%	1163	125	12%	501	115	11%

Table 2. Evaluation of training and test data.

These results show that, the costumers that buy cleaning items generally make their weekly or monthly shopping. In other words, most of the baskets that include cleaning items are in the full basket category. From the list of items in the results, we can not understand exactly whether they will increase the profit of a supermarket. The exact evaluation can be done after putting these items on promotion an see whether we achieve a significant increase in the total cost of the baskets containing these items. Some statistical tests (hypothesis test) can be applied to see whether an increase in the profit is significant. Certainly, such decisions on items to be promoted are open to the comments of the managers of supermarkets.

A problem that arises with this experiment is the running time of the classification algorithm. On our data, which have approximately 5000 items and over 11000 transactions, its execution took 13 hours with a machine that have enough memory to hold all these data. On the other hand, we did not meet any memory and execution time problem in the clustering step, by using KMEANS clustering.

6 Optional Clustering

After forming the decision tree and extracting the attributes that have positive leaves on their 1 valued branches, selecting the final item set manually may be time consuming if the number of items is very large. If we have a data set with more than tens of thousands of attributes we may apply clustering on the attributes using their information gain as a similarity criterion.

Other alternative approaches to this situation can be employed. We can employ a threshold value, and extract the attributes above the threshold. Another approach is that, if the number of items to be put on promotion is given, we can select the given number of attributes starting from the most significant one.

7 Conclusion

We have described an abstract or high-level data mining algorithm having two consecutive steps, clustering and classification, respectively, as a new promotion method for supermarkets. We applied it to a real basket data, collected in a period, in order to obtain a set of items that significantly most frequent in the large or full transactions of a big supermarket. The method we described, search for small number of items among thousands, that are too much for a manual work.

We applied KMEANS clustering in the first step and after that, the decision tree learning algorithm in the classification phase of our application. Alternative classification algorithms can be searched in order to extract such items, and their accuracy and efficiency on basket data can be tested. The most important property of a classification algorithm for the applicability in this application is that, it must give a measure about the weights of attributes on the classes. Experiments on the market data showed that, the execution time of C4.5 program increases quadratically when the data size increased. Quadratic running time of C4.5 is also discussed in [4]. In a search for appropriate classification algorithm, efficiency is important. To increase the efficiency, some feature selection and data abstraction methods can be tried, by excluding some items from the data set in the beginning of the process such as described in [4, 5].

Acknowledgments

We thank Beğendik Corp. for providing us their basket data that belongs to one of their supermarkets located in Ankara.

References

1. R.Agrawal, T.Imielinski, A.Swami, Mining Association Rules between Sets of Items in Large Databases, *Proceedings of the ACM SIGMOD*, May 1993.
2. R.Agrawal. T. Imielinski, A.Swami, Database Mining: A Performance Perspective, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 5, No.6, December 1993.
3. R.J. Brachman, T. Khabaza, W. Kloesgen, G. Piatetsky - Shapiro, E. Simoudis, Mining Business Databases, *Communications of the ACM*, November 1996.
4. J.Catlett, Peepholing: Choosing Attributes Efficiently for Megainduction, *Proceedings of 9th Int. Conference of Machine Learning, Morgan Kaufman Publishers Inc.*, pp49-54, 1992.
5. V.Dhar, A.Tuzhilin, Abstract Driven Pattern Discovery in Databases, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 5, No.6, December 1993.
6. R. Duda , P.E. Hart, Pattern Classification and Scene Analysis,*Wiley*, 1973.
7. U. Fayyad, R. Uthurusamy, Data Mining and Knowledge Discovery in Databases, *Communications of the ACM*, November 1996.
8. U. Fayyad, D. Haussler, P. Stolorz, Mining Scientific Data, *Communications of the ACM*, November 1996.
9. L. Kaufman, P.J. Rousseeuw, Finding Groups in Data -An Introduction to Cluster Analysis, *Wiley Series in Probability and Mathematical Statistics*, 1990.
10. C. J. Matheus, P. K. Chan, G. Piatetsky-Shapiro, Systems for Knowledge Discovery in Databases, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 5, No.6, December 1993.
11. T.M.Mitchell,Machine Learning, *McGraw Hill*, 1997.
12. J.R.Quinlan, Induction of Decision Trees, *Machine Learning*, 1, 81-106, 1986.
13. J.R.Quinlan, C4.5: Programs for Machine Learning, *Morgan Kaufmann Publishers Inc.*, 1993.