International Journal of Computer Science & Applications
Vol. II, No. II, pp. 85 - 100
© 2005 Technomathematics Research Foundation

Relationship between Product Based Loyalty and Clustering based on Supermarket Visit and Spending Patterns

Chad West, Stephanie MacDonald, Pawan Lingras[1], and Greg Adams
Department of Mathematics and Computing Science
Saint Mary's University, Halifax
Nova Scotia, Canada, B3H 3C3

Abstract

Loyalty of customers to a supermarket can be measured in a variety of ways. If a customer tends to buy from certain categories of products, it is likely that the customer is loyal to the supermarket. Another indication of loyalty is based on the tendency of customers to visit the supermarket over a number of weeks. Regular visitors and spenders are more likely to be loyal to the supermarket. Neither one of these two criteria can provide a complete picture of customers' loyalty. The decision regarding the loyalty of a customer will have to take into account the visiting pattern as well as the categories of products purchased. This paper describes results of experiments that attempted to identify customer loyalty using these two sets of criteria separately. The experiments were based on transactional data obtained from a supermarket data collection program. Comparisons of results from these parallel sets of experiments were useful in fine tuning both the schemes of estimating the degree of loyalty of a customer. The project also provides useful insights for the development of more sophisticated measures for studying customer loyalty. It is hoped that the understanding of loyal customers will be helpful in identifying better marketing strategies.

1. Introduction

Customer loyalty is an important component of marketing analysis in a supermarket. The loyalty of a customer may be apparent through the products bought by the customer. Certain product categories such as bread and eggs may have a higher ability to distinguish between loyal and disloyal customers. Other product categories such as coffee/tea and ketchup may not be deterministic of a customer's loyalty but may simply enhance their degree of loyalty. Establishing a scoring system based on such key product categories is one possible way of determining customer loyalty. However, the dietary habits of some loyal customers may lead to lower loyalty scores if they are based solely on product categories. Studying patterns in transactional records can also provide important clues about the loyal patrons of the supermarket. It is important to conduct parallel analyses of products purchased and transaction patterns for identifying loyal customers. The two separate analyses can also be used for fine-tuning each other.

This paper reports the results of experiments that studied various characteristics of loyal customers based on the products purchased and visiting patterns. The experiments were based on the data obtained from a large national supermarket chain, which was gathered over a thirteen-week period in 2000.

The project was divided into two parallel streams: product based and transaction pattern based analyses. The product based analysis started with a preliminary definition of loyal customers, based on spending levels.

[1] The authors would like to thank NSERC Canada, the Nova Scotia Cooperative Employment Program, and the Senate Research Grant Committee of Saint Mary's University for the financial support. The authors are also grateful to the supermarket chain and it's management for allowing us the use of the data.

This preliminary definition was useful for identification of departments favored by loyal customers. The departmental level analysis in itself was found insufficient for determining the characteristics of loyal customers.

For further understanding, a study of the detailed spending patterns within each department was carried out. A comparison with the AC Nielsen (2001) figures for average consumption allowed a better understanding of loyal customers. It is also possible that high spending level thresholds may exclude smaller families from the analysis. Therefore, adjustments were made to the spending level threshold in an effort to include smaller families. The preliminary data analysis described above provided some information about the relationship between products and loyal customers. This knowledge was used for the development of appropriate loyalty measures based on products favored by loyal customers, and product categories that are under performing. The loyalty measures developed were then used to evaluate the classifications based on the transactional patterns.

Many of the data mining applications use average or total values of certain important attributes such as amount of money spent to create customer profiles. However, temporal variations in values of these variables can also provide important insights into the shopping habits of a customer. Lingras and Young (2001) used time-series of six variables. The customer profiles resulting from the time-series illustrated the advantages of the time-series representation. However, the time-series of many of the chosen variables had similar patterns. Lingras and Adams (2001) revisited the clustering done by Lingras and Young (2001). Various combinations of the six time-series indicated that it is possible to eliminate variables with similar patterns without having significant impact on the resulting customer profiles. The results further underscored the importance of using time-series instead of average values of variables. Experimentation with different weights showed that it is possible to obtain more meaningful clustering by careful fine-tuning of weights of the variables. This study used the weighted clustering scheme suggested by Lingras and Adams for the new data set, which consisted of a larger number of customers.

The product based loyalty scores were calculated for all the clusters created using visits and spending patterns. Some of the flaws in the initial loyalty scoring system became evident after studying the loyalty scores for different clusters. The system was subsequently modified to provide more reasonable loyalty scores. One of the disadvantages of using weekly statistics was also noticed in the cluster patterns. A few customers may shop at the beginning and at the end of a certain week, and not shop in the preceding or following week. Such a shopping behaviour can result in visits and expenditures varying greatly between weeks. The time-series was modified by taking the average for three consecutive weeks. The clustering was performed again based on these modified time series. The resulting shopping patterns tended to have fewer fluctuations and a flatter graphical representation. The loyalty scores were recalculated for the new clusters. The paper provides an analysis of the resulting clusters and their loyalty scores.

2. Review of Literature

This section provides a brief review of data mining from a marketing and techniques points of view. General introduction to data mining techniques is followed by more specific introduction to K-means clustering used in this study.

2.1 Marketing and Data Mining

Marketing analysts consider data mining to be the process of analyzing a company's internal data for customer profiling and targeting. Marketing databases often handle tens of millions of customer records, and in the case of direct marketing even small improvements in the yield for a mailing can mean substantial profits. Database marketing is concerned with predicting customer response to promotions.

Customer Lifetime Value (LTV), which measures the profit generating potential of a customer, is increasingly being considered a touchstone in customer relationship management. LTV can be used to segment customers, and to determine which customers should be the focus of marketing efforts and dollars. Another measure that is useful in customer relationship management is customer loyalty. Determining customer loyalty is a complicated process that involves many measurements and calculations. To help determine loyalty, customer purchase models can be created based on purchases of non-durable consumer goods (Ehrenberg, 1959). These goods are usually marketed in prepackaged and branded form (Mani, 1999).

The basic unit of time for measuring consumer purchases is usually a week. It is assumed that purchases in one-week will generally be similar to any other week. Most analyses are made over periods of 4 or 13 weeks. One feature of consumer purchasing data is that consumers tend to buy the number of units of a

product equal to the number of weeks covered. Note that the size of individual units will depend on the size of the family. This arises because some customers will tend to buy practically the same number of units nearly every week (Mani, 1999). The periods of 4 or 13 weeks allows the analysis to include those products that are bought only once a month or once a season.

Complete customer profiles can be generated once the proper data is collected. Profiles consist of two parts: factual and behavioral. The factual profile contains information, such as name and address. The behavioral profile models the customer's actions and is usually derived from transactional data (Adomavicius and Tuzhilin, 2001). The LTV and loyalty analyses of customers are examples of items that could appear in their behavioral profile. Profiling customers also allows them to be segmented into subgroups. An example of such subgroups is given by Chatfield and Goldhardt (1970). In two consecutive equal time - periods of n weeks the population can be divided into four subgroups. A "repeat" buyer buys in both periods, a "lost" buyer buys in period I but not in period II, a "new" buyer buys in period II but not in period I, and a non-buyer buys in neither periods. Other more-complicated subgroups can be determined depending on the level of detail of the data collection. The present paper uses some of the results and analysis from earlier studies to describe a loyalty scoring scheme for a supermarket.

2.2 Data Mining Techniques

Data mining, which is also referred to as knowledge discovery in databases, is a process of nontrivial extraction of implicit, previously unknown and potentially useful information (such as knowledge rules, constraints, and regularities) from data in databases (Pelleg and Moore, 1999). Data mining draws on the results from various fields, such as database systems, machine learning, intelligent information systems, statistics, and expert systems (Deogun, *et al*., 1997).

Data mining results are being used frequently by companies to optimize marketing campaigns. Campaigns can be designed to target specific customer groups. A current initiative that draws greatly from data mining results is the IBM-Safeway project (Bellamy, *et al*., 2000). An electronic hand held device has been designed that allows customers to order their groceries remotely. This hand held device collects data about the customer's shopping habits and uses data mining techniques to help compile shopping lists. The device will also offer customer specific discounts. Future applications of data mining will aim to increase customer satisfaction and convenience.

Several typical kinds of knowledge can be discovered by data miners, including association rules, characteristic rules, classification rules, discriminant rules, clustering, evolution, and deviation analysis (Chen, *et al*., 1996). Three of the most widely used techniques are association, classification, and clustering.

Association rule mining finds interesting correlation among a large set of data (Han and Kamber, 2001). These relationships can help managers make smart business decisions. Association rules appear in the form r : F(o) => G(o), where: F is a conjunction of unary formulas, G is an unary formula. Each rule r is associated with a confidence factor c, $0 \leq c \leq 1$, which shows the strength of the rule r (Deogun, *et al*., 1997). A typical example of association rule mining is market basket analysis. For instance, if customers are buying milk, how likely are they to also buy bread (and what kind of bread) on the same trip to the supermarket (Han and Kamber, 2001)?

Data classification is the process that finds the common properties among a set of objects in a database and classifies them into different classes, according to a classification model. The objective of the classification is to first analyze the training data and develop an accurate description or a model for each class using the features available in the data. Such class descriptions are then used to classify future data or to develop a better description for each class (Chen, *et al*., 1997). For example, a classification model may be built to categorize bank loan applications as either safe or risky (Han and Kamber, 2001).

Cluster analysis is one of the basic tools for exploring the underlying structure of a given data set and is being applied in a wide variety of engineering and scientific disciplines. The primary objective of cluster analysis is to partition a given data set of multidimensional vectors (patterns) into homogeneous clusters. Patterns within a cluster are more similar to each other than patterns belonging to different clusters (Su, *et al*., 1997). Data clustering identifies the sparse and the crowded places, and hence discovers the overall distribution patterns of the data set (Chen, *et al*., 1996).

2.3 K-means clustering

There are numerous clustering algorithms ranging from the traditional methods of distance based pattern recognition to clustering techniques in machine learning (Deogun, *et al*., 1997). Distance based approaches are

beneficial due to their straightforward implementation. The drawback to this method is that they are not linearly scalable with stable clustering quality. The clustering must inspect all data points and globally measure their distance from each cluster no matter how close or far away they are. For large data sets the runtime of such an algorithm is intolerably long (Chen, *et al*., 1996). In machine learning, clustering analysis often refers to unsupervised learning, since the class an object belongs to is not pre-specified (Chen *et al*., 1996). This approach can lead to some interesting findings that may be overlooked with traditional clustering methods. Future research is required in making machine learning algorithms readily applicable to large databases due to long processing times and intricacies of complex data (Han and Kamber, 2001).

Lingras and Huang (2004) compared a variety of clustering algorithms including hierarchical grouping, Kohonen self-organizing maps, genetic algorithms, and K-means for datasets are various sizes. They found that K-means provides a reasonable balance between accuracy and performance for large datasets. K-means clustering is one of the most popular statistical clustering techniques (Hartigan and Wong, 1979, MacQueen, 1967). The name K-means comes from the means of the *k* clusters that are created from *n* objects using the method.

Let us assume that the objects are represented by *m*-dimensional vectors. The objective is to assign these *n* objects to *k* clusters. Each of the clusters is also represented by an *m*-dimensional vector, which is the centroid or mean vector for that cluster. The process begins by randomly choosing *k* objects as the centroids of the *k* clusters. The objects are assigned to one of the *k* clusters based on minimum value of distance $d(\mathbf{x},\mathbf{y})$ between the object vector x and cluster vector y. With vector representation of x and y, the distance $d(\mathbf{x},\mathbf{y})$ in eq. 1 can be calculated as:

$$d(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{m} (x_i - y_i)^2 \ , \qquad (1)$$

where x_i and y_i are i^{th} components of the vectors x and y, respectively.

After the assignment of all the objects to various clusters, the new centroid vectors of the clusters are calculated as:

$$y_i = \frac{\sum x_i}{\text{Size of cluster } \mathbf{y}} \qquad (2)$$

, where $1 \le i \le m$.

The process stops when the centroids of clusters stabilize, i.e. the centroid vectors from previous iteration are identical to those generated in the current iteration.

3. Preliminary analysis with product based loyalty scores

This section describes the initial results of loyalty scores based on product purchases. The data was obtained from a supermarket chain, which has stores in all of the Canadian provinces.

Since all customers are loyal to varying degrees, one needs to decide on a level of loyalty. It was decided to initially focus on customers who spend between $100 and $150 per week. It was assumed that these customers would be spending the majority of their grocery dollars with the supermarket. The spending behavior of these customers may determine common characteristics of loyal shoppers. Categories important to loyal customers will be helpful in determining category roles.

Customers that average $100-$150 per week spend a larger portion of their grocery dollars in meat and general merchandise. The analysis showed lower expenditures in produce by these customers. Higher spending customers shop frequently in the deli, floral, pharmacy, tobacco, and service case meat departments. They have a lower penetration in produce, dairy and grocery. It was noticed that higher spending customers shopped an average of 11 distinct departments over five weeks. Customers who spend $0-$50 and $50-$100 per week, averaged 7 and 9.5 departments, respectively. This stage revealed interesting tendencies of loyal customers. A more in-depth analysis was required to determine customer characteristics.

The next logical step was to look at the number and type of categories customers shopped in. The first noticeable characteristic of high spending customers was the number of categories they shopped in over five weeks. They averaged 50 distinct categories. Customers who spent $0-$50 and $50-$100 shopped in

approximately 12 and 35 categories, respectively. The study of sales ratios in each category exposed the variations within certain departments. For example, the lower ratio in produce is mainly the result of reduced spending in fresh fruit. Similarly, the higher sales ratio in meat is mainly because of purchases of beef and chicken. The high penetration in deli appears to be due to the increased ratio in fresh luncheon meats. Other categories with high sales ratios are nutritious portable foods, pet food and supplies, laundry detergent, and bathroom tissue.

It is possible that smaller families may have lower spending, but may be equally loyal to the supermarket. The target spend segments were extended to $75-$100 per week shoppers to include smaller families. The next step focused on the supermarket's under-performing categories, as compared to the market. Loyal customers are expected to still shop for products in the supermarket's under-performing categories. Less loyal customers may purchase these products elsewhere. AC Nielsen (2001) figures and previous findings related to customer potential were used to determine key variables related to customer loyalty.

A category sales ratio analysis showed that ratios in many key categories are lower for higher-spending customers. This is reasonable since their purchases are spread over a larger number of categories. There is only so much that can be spent in certain categories. It was decided to look at the combination of categories in which each customer shopped.

Product Grouping	**Loyalty Score**
Fresh Fruit (loose or pre-packaged)	2
Fresh Vegetables (loose or pre-packaged)	2
Meat – Fresh or Packaged Fresh or frozen/boxed	1
Bread – Commercial or In-store	1
Sugar – White sugar or sugar substitute	2
Margarine or Butter	1
Cereal – hot or cold or toaster pastries	1
Salad Dressing (pourable, spoonable) or Spreads or Condiments	1
Cheese – any type (slices, brick, shredded, etc.)	1
Eggs	1
Total Loyalty Score for Required Products	**13**

Table 1. Products which must be purchased and pass the quantity restriction

A loyalty scoring system was created based on the supermarket's performance in each category, as compared to the market. Table 1 shows the name of the products and associated loyalty scores for required categories. Required categories were chosen based on the results of spend segment analyses above. In addition to the required categories, others were chosen which may enhance indication of loyalty. Table 2 lists these extra categories. Some categories were given higher loyalty scores based on their performance against last year's total market figures (AC Neilsen 2001). It is assumed that more customers must be purchasing products from under-performing categories at competitors' stores. Those continuing to purchase from the supermarket's under-performing categories are deemed to be more loyal. Therefore, under-performing categories are given a loyalty score of two. All other categories are given a loyalty score of one.

In order to give equal weighting to all categories (except for the under-performance score), a minimum quantity purchased was used as a threshold. Clarke (1993) illustrated the use of thresholds. Variables may be indicative of a characteristic if they meet necessary threshold conditions defined for the situation. Let category X have an average elapsed days of purchase of Y. Transaction data was extracted for Z days. The purchase frequency of category X must be greater than or equal to Z / Y.

The product based loyalty scores are expected to relate to customer spending and visit patterns. Lingras and Young (2001) experimented with a variety of different criteria for classifying customers using sorted time series. Lingras and Adams (2001) refined the approach further by trying to capture the spending potential and

loyalty of customers. It may be interesting to study the relationships between loyalty scores and unsupervised classification.

4. Clustering based on sorted time-series

Classification or clustering plays an important role in supermarket data mining. For example, designing individual promotional campaigns is impractical. It is more feasible to design campaigns for small number of representative classes. The classification can be based on many different criteria. Examples of the criteria include the spending potential of customers and their loyalty to the store. The simplest classification is based on average weekly spending of a customer; however, this classification does not necessarily capture the loyalty of the customer to the store. A more detailed classification should consider many other criteria such as:

Product Grouping	**Loyalty Score**
Potatoes or rice or pasta	1
Milk - liquid or powdered	1
Coffee or tea	1
Soft drinks or water or juice (refrig., frozen, shelf-stable or powdered)	2
Soup - canned or condensed or dry	1
Cooking oils - any type	1
Canned pasta or side dishes	1
Ketchup	1
Jams or jellies or peanut butter	1
Crackers (soda or specialty)	1
Cookies	1
Potato chips or other dry snack	1
Garbage bags - any size	1
Laundry detergent	1
Bleach or fabric softener	2
Paper towels	1
Household cleaners	2
Soap - hand or body or shower	1
Deodorant	1
Shampoo	1
Toothpaste	1
Facial Tissue	1
Canned Meat or frozen vegetables or canned vegetables	1
Dish detergent	1
Bathroom tissue	1
Total Possible Loyalty Score for Extra Products	**28**

Table 2. Products which will add loyalty points if purchased and pass the quantity restriction

- How many different product categories did the customer spend money in? (Examples of categories are meats, fruits and vegetables, etc.)
- How many different sub categories did the customer purchase from? (Subcategories are more specific than categories, e.g. pork, beef, etc.)
- How many products did the customer purchase?
- How much money did the customer spend?
- How often did the customer visit?

Lingras and Young (2001) prepared a data file using the six criteria mentioned earlier. The use of average values for the six variables may hide some of the important information present in the temporal patterns. Therefore, Lingras and Young (2001) used the weekly time series values for the six variables. It is possible that customers with similar profiles may spend different amounts in a given week. However, if the values were sorted, the differences between these customers may vanish. For example, three week spending of customer *A* may be \$10, \$30, and \$20. Customer *B* may spend \$20, \$10, and \$30 in those three weeks. If the two time-series were compared with each other, the two customers may seem to have completely different profiles. If the time-series values were sorted, the two customers will have identical patterns. Therefore, the values of these six variables for 13 weeks were sorted, resulting in a total of 78 variables. A variety of values of *K* (number of clusters) were used in the initial experiments. However, large values of *K* made it difficult to interpret the results. It was decided that five classes of customers might be useful for further analysis. The Kohonen neural network was created using 78 input nodes and five output nodes. The networks were tested for different values of training cycles and learning parameters. The learning parameter of 0.01 and twenty-five training cycles provided the smallest within group error. The results were also compared with another statistical technique called K-means. The Kohonen network was more efficient and provided comparable accuracy.

Based on spending patterns, and variations in visits and discounts, Lingras and Young (2001) described the following five customer groups:

Group 1: Loyal big spenders
Group 2: Infrequent customers
Group 3: Semi-loyal potentially big spenders
Group 4: Loyal moderate spenders
Group 5: Potentially moderate to big spenders with limited loyalty

Lingras and Young's (2001) results indicated that all six time-series may not be necessary for clustering. It is possible that some of the variables do not provide additional information. This observation was possible because of the use of sorted time-series as opposed to single average values of the variables. Lingras and Adams (2001) experimented with different combinations of time-series to create different clustering schemes. From the six clustering schemes, they found a weighted scheme that provided the best results.

The clustering scheme proposed by Lingras and Adams (2001) used more reasonable weighting of the value time-series and visits time-series. The value of groceries was found to be a good indicator of customers' spending potential. The value time-series provides some indication about the customer's loyalty. However, the visits time-series can provide additional information about the tendency of the customer to choose the supermarket over competitors. Lingras and Adams used a weighting scheme to make sure that the value of groceries did not dominate the clustering. On average visits were 50 times smaller than value. Assuming that value is twice as important as visits, the visits data was multiplied by 25.

The reasonable balance in customer loyalty and spending potential was possible because of the weighting scheme. Different emphasis can be obtained by changing the weights of the two time-series. The weighting scheme can be expanded to include other time-series as well. For example, if value-consciousness was an important issue, one could assign an appropriate weight for the discounts time-series. However, there seemed to be limited information gained by including the other three variables, namely, numbers of categories, subcategories, and items.

The present study used the clustering scheme suggested by Lingras and Adams (2001) to cluster customers from seven supermarket stores concentrated in a rural setting. The supermarkets are part of a national chain. The data was taken over a thirteen-week period starting in July 2000. It included information on spending, visits, categories shopped, and other transactional data.

The clustering was done using the data mentioned above. Weekly totals and visits were used as input to both k-means and Kohonen neural network clustering algorithms. Since the data was taken over a thirteen-week

period there were a total of twenty-six variables for each record. The weighting scheme proposed by Lingras and Adams (2001) was applied. Totals were roughly twice as important as visits during the clustering. Contrary to the findings of Lingras and Young (2001), the k-means method provided more appropriate results. The k-means method showed only a slight loss in efficiency. This difference can be attributed to the larger data set that was used for the current study.

The clustering resulted in groups similar to those obtained by Lingras and Adams (2001). Figure 1 shows the value and visits time-series for the five groups. Based on the patterns shown in Figure 1, the groups can be described as follows:

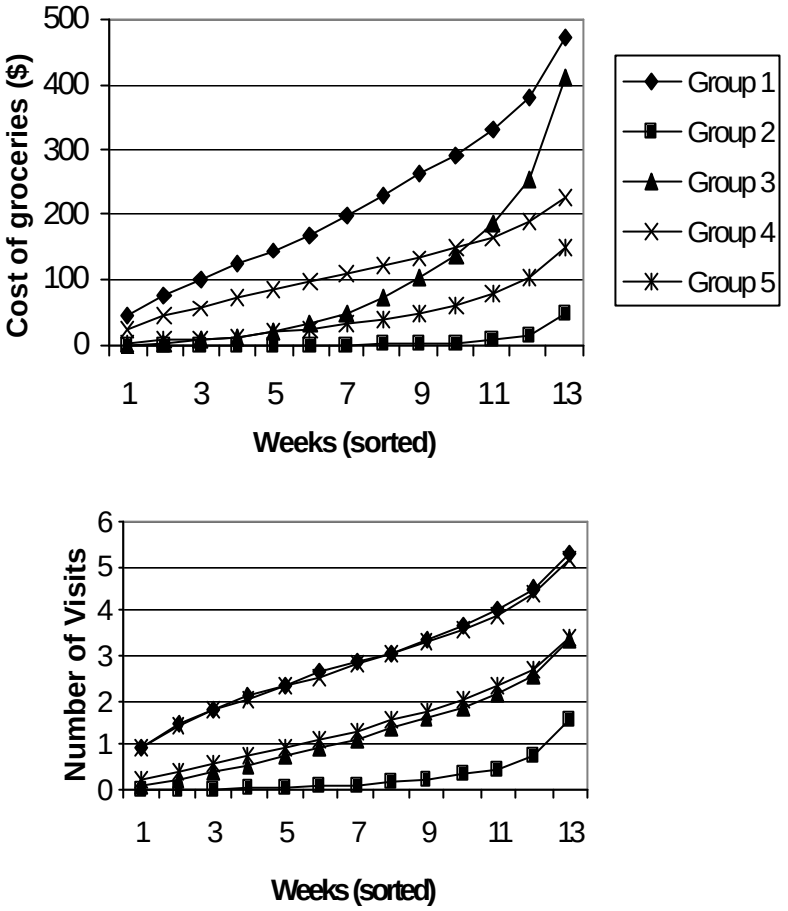

Figure 1. Visits and cost of groceries time-series on 2000 supermarket data

Group 1: Loyal big spenders
This group consists of the largest spenders. The weekly spending ranges from $25 to more than $400. They are frequent visitors and seem to be very loyal to the store.

Group 2: Infrequent customers
Customers from this group are the least loyal to the store among all the groups. They seem to have only visited the store once or twice during the thirteen weeks. Their spending was very limited as well. It is also possible that some of these customers do not use the Supermarket card on a regular basis.

Group 3: Semi-loyal potentially big spenders
In terms of maximum amount spent, this group is comparable to the first group. Based on this observation alone, one may categorize these customers as the second most loyal customers. However, the thirteen-week patterns indicate that for 3-4 weeks these customers tended to stay away from the store. There were additional 4-5 weeks

with limited spending and visits. The supermarket may not be attracting a significant portion of purchases from these customers. More incentives to increase the patronage from these customers may be worthwhile.

Group 4: Loyal moderate spenders

Even though the maximum spending for these customers was smaller than group 3, their spending patterns were the most stable among all the groups. The total number of visits was
almost identical to group 1. These customers may be the most loyal among all the groups. They are not big spenders like the customers from group 1 and 3. They are more likely to be value conscious customers.

Group 5: Potentially moderate to big spenders with limited loyalty

These customers are similar to those from group 2. However, spending and visits over thirteen weeks indicate that these customers are more frequent and spend a little more than those from group 2. It is also possible that they don't always use the supermarket card.

5. Loyalty Scores Based on Product Categories

The loyalty scoring system described in section 2 was applied to the clusters developed in section 3. Initially, the quantity restrictions were not used in the analysis. Table 3 shows the 50th percentile, 95th percentile, and maximum for the five clusters.

Cluster	50th percentile	95th percentile	Maximum
1	36	39	40
2	0	0	37
3	0	39	40
4	33	39	40
5	0	35	40

Table 3. Distribution of loyalty scores

The 50th percentile, 95th percentile, and maximum values were used to provide a clearer picture of loyalty scores from each cluster. Comparison of Table 3 and Figure 1 shows a correspondence between the loyalty scores and the time-series graphs. Group 1 customers are high spenders and frequent visitors. More than half of the customers in group 1 had loyalty scores above 36. Loyalty of group 4 (loyal moderate spenders) was also comparable. More than 50% of group 4 had loyalty scores above 33. Groups 2, 3 and 4 were expected to have limited loyalty. More than half of the customers in these groups had zero loyalty scores. The 95th percentile scores for these three groups confirmed the findings obtained from the cluster analysis. The top 5% of customers in group 3 (semi-loyal potentially high spenders) had loyalty scores above 39. The top 5% of customers in group 5, who were deemed semi-loyal and moderate to high spenders, had loyalty scores above 35. As expected, Group 2 had
the worst loyalty scores. More than 95% of the customers from group 2 had zero loyalty scores. It was considered worthwhile to make a further study of zero and non-zero loyalty scores.

Cluster	Number of customers	Zero loyalty scores	Average of non-zero loyalty scores
1	1390	26%	37
2	11749	99%	26
3	1936	53%	34
4	3548	38%	35
5	7666	74%	31

Table 4. Analysis of zero and non-zero loyalty scores

Table 4 shows the total number of customers in each cluster, the percentage of the customers with zero loyalty scores, and the average of non-zero loyalty scores. The percentage of zero loyalty scores matches the analysis of cluster patterns. Loyal groups have lower percentages of zero loyalty scores. However, for all the groups, the percentage number of zero loyalty scores seems rather high. Overall, 56% of the customers had zero

loyalty scores. This was one of the disadvantages of the initial loyalty scoring system. The customers received zero loyalty scores if they did not shop in one or two of the required categories. An example of a customer with a zero loyalty score could be a vegetarian household. Since meat is a required category under the current system, vegetarians would be assigned a score of zero. Even if a vegetarian was loyal, and shopped in every other category frequently, the current scheme would lead to a loyalty score of zero.

An additional shortcoming of the existing system was the range of non-zero loyalty scores. This is evident in the averages of non-zero loyalty scores in Table 4. The averages are consistently high for all the groups. Further analysis showed that the lowest non-zero score for all the groups was 19 and the maximum in most cases was 40. There is a large gap between a loyalty score of zero and nineteen. It would be more desirable to have an even distribution of loyalty scores. Based on these observations, the loyalty scoring system was modified as described in the next section.

6. Modified Loyalty Scores

The loyalty scoring system was modified to include the quantity restrictions given by AC Nielsen (2001) figures. The number of required categories was increased from ten to thirteen. The new scoring scheme is outlined in Tables 5 and 6. A customer is now only required to purchase in twelve of the thirteen required categories. This flexible requirement did not unduly penalize customers with dietary restrictions such as vegetarians.

Product Grouping	**Loyalty Score**
Fresh Fruit (loose or pre-packaged)	2
Fresh Vegetables (loose or pre-packaged)	2
Meat – Fresh or Packaged Fresh or frozen/boxed	1
Bread – Commercial or In-store	1
Sugar – White sugar or sugar substitute	2
Margarine or Butter	1
Cereal - hot or cold or toaster pastries	1
Salad Dressing (pourable, spoonable) or Spreads or Condiments	1
Cheese - any type (slices, brick, shredded, etc.)	1
Eggs	1
Milk – Liquid or powdered	2
Soft Drinks or Water or Juice (refrigerated, frozen, shelf-stable or powdered)	2
Potatoes or Rice or Pasta	1
Total Loyalty Score for Required Products	**18**

Table 5. Products which must be purchased and pass the quantity restriction

The resulting loyalty scores provided a more accurate representation of the clusters. Table 7 describes the distribution of modified loyalty scores for all the groups. The separation between 50^{th} percentile, 95^{th} percentile, and maximum scores for all the groups is approximately 7-10 points. The loyalty scores at the 50^{th} percentile, 95^{th} percentile, and maximum levels decrease in order starting from group 1, group 4, group 3, group 5, to group 2 as suggested by the clustering. The differences between these groups are also distinct, ranging from 3 to 10 points.

More customers were able to meet the requirements to be deemed a loyal customer. Under the old scheme 56% of customers had a loyalty score of zero. The new scheme reduced this number to 36%. The analysis of customers with zero loyalty scores is shown in Table 8. The percentages of zero loyalty scores for the two loyal groups 1 and 4 are significantly smaller than the old scoring system. The modified loyalty scoring system also provided a better distribution of the scores. The range of non-zero loyalty scores was increased by twenty. The minimum non-zero score was one and the maximum increased to 41. Lower values were obtained

more frequently for customers in the disloyal clusters. Higher scores continued to represent customers in the loyal clusters. The average of non-zero scores in Table 8 shows a distinct decrease with decreasing loyalty ranging from 22 for group 1 to three for group 2.

Product Grouping	**Loyalty Score**
Coffee or tea	1
Soup – canned or condensed or dry	1
Cooking oils - any type	1
Canned pasta or side dishes	1
Ketchup	1
Jams or jellies or peanut butter	1
Crackers (soda or specialty)	1
Cookies	1
Potato chips or other dry snack	1
Garbage bags - any size	1
Laundry detergent	1
Bleach or fabric softener	2
Paper towels	1
Household cleaners	2
Soap – hand or body or shower	1
Deodorant	1
Shampoo	1
Toothpaste	1
Facial Tissue	1
Canned Meat or frozen vegetables or canned vegetables	1
Dish detergent	1
Bathroom tissue	1
Total Possible Loyalty Score for Extra Products	**24**

Table 6. Products which will add loyalty points if purchased and pass the quantity restriction

Cluster	50th percentile	95th percentile	Maximum
1	22	29	41
2	0	0	20
3	11	21	41
4	17	26	35
5	3	15	29

Table 7. Distribution of modified loyalty scores

Cluster	Number of customers	Zero loyalty scores	Average of non-zero loyalty scores
1	1390	5	22
2	11749	98	3
3	1936	23	13
4	3548	10	18
5	7666	45	9

Table 8. Analysis of modified zero and non-zero loyalty scores

7. Clustering based on Modified Time Series

Time series of weekly statistics may not accurately represent spending patterns of a customer. A person may do a significant amount of shopping at the beginning and end of a week, and reduce the shopping in the preceding or following week. This can lead to extreme values in the time-series. The average of the current, the preceding, and the following week can be used to overcome this problem. This data smoothing technique is known as a three-day moving average. Since the first and last weeks do not have either a preceding or following week, the total number of variables is reduced by two. The resulting time series is eleven weeks long compared to thirteen weeks in the original time-series.

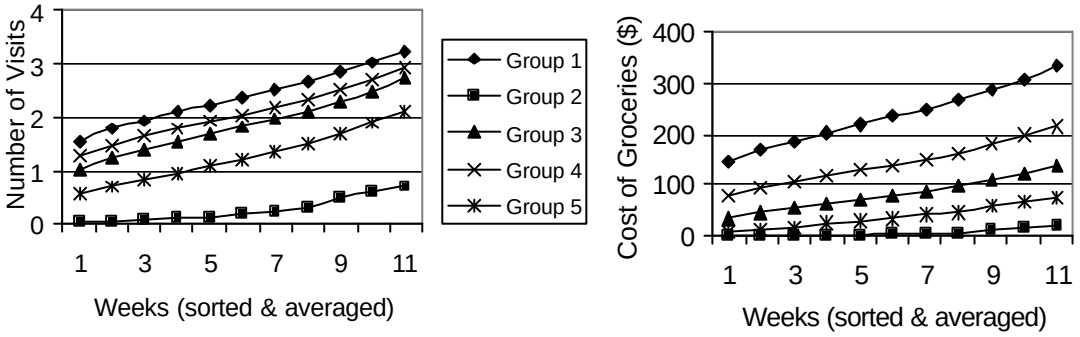

Figure 2. Three-day moving average time-series of visits and value of groceries on 2000 supermarket data

Figure 2 shows the moving average time series for the five clusters. The corresponding analysis of the modified loyalty scores is shown in Tables 9 and 10. Figure 2 seems to suggest that smoothing the data provides a greater distinction between the clusters. Smoothing also causes the time-series to be more stable and linear. The three-day moving average time-series shows that cluster four has consistently higher values of groceries and visits than cluster 3. That means cluster 4 has more loyal customers than cluster three. The value time series for cluster 3 and 4 crossed each other in Figure 1. The actual clusters obtained using the moving average are significantly different from each other. The comparison of the second columns in Tables 8 and 10 show that the sizes of groups 1 and 4 are significantly smaller with the moving average time series. Group 4 (semi-loyal and potentially high spender) is the biggest gainer in terms of size.

Cluster	50th percentile	95th percentile	Maximum
1	23	28	37
2	0	0	8
3	12	21	29
4	19	27	35
5	0	12	20

Table 9. Distribution of modified loyalty scores for moving average clustering

Cluster	Number of customers	Zero loyalty scores	Average of non-zero loyalty scores
1	673	4%	22
2	11014	99%	2
3	4134	19%	13
4	2263	9%	19
5	6366	56%	7

Table 10. Analysis of modified zero and non-zero loyalty scores for moving average clustering

The clustering based on the moving averages had small but important effects on the loyalty scores shown in Tables 9 and 10. The percentages of zero loyalty scores are higher for loyal groups (groups 1 and 4) and lower for less loyal groups, such as group 2. The maximum score for the least loyal group 2, is significantly smaller with the moving average based clustering. The new clustering scheme also had a slight effect on the range of loyalty scores. A more detailed analysis of the customers will be necessary to determine whether the clustering obtained with moving average time series is better than the conventional time series.

8. Summary and Conclusions

This paper describes the relationship between product-based loyalty and clustering based on time series of supermarket data. Clustering was done on visits and total weekly expenditures using Kohonen neural network and k-means methods. The results of the clustering were graphed as time-series to analyze the effectiveness of a loyalty scoring system.

A scoring system was proposed to evaluate the loyalty of supermarket customers. Points were assigned to customers based on their purchases within key product categories. The system was not optimal because 56% of the customers were unable to meet the specified requirements. The scoring system did not always show distinct differences between loyal and disloyal clusters.

A modified scoring system was derived from the original loyalty scoring system. The changes included the addition of quantity restrictions and the modification of the required categories. The modified system allowed more of the customers to meet the new requirements. An increased distribution of scores was also obtained under the new scheme. The minimum non-zero score decreased from 19 to one.

Finally, a three-day moving average was introduced into the clustering and loyalty scoring systems. This system was implemented to compensate for irregularities in customer shopping behavior. Visits and totals values for a week were averages of the preceding, current, and following weeks. The data was then sorted and plotted as time-series graphs. The moving average based clusters were significantly different in size compared to the conventional time series. The moving average patterns of the clusters were more distinguishable from each other. There were small but significant differences between the loyalty scores for the two clustering schemes. A more detailed analysis at individual customer level will be necessary to study the desirability of the moving average patterns.

References

AC Nielsen, 2001. Market Track Report for 52 Weeks ending December 2, 2000.

Adomavicius, G. and Tuzhilin, A., 2001. Using Data Mining Methods to Build Customer Profiles. IEEE Computer. Volume 34, Issue 3, pp. 74 – 82.

Bellamy, R.K.E., Kellogg, W.A., Richards, J.T., and Swart, C.B., 2000. Exploring A New Paradigm for E-Groceries. www.ibm.com/ibm/easy/eou_ext.nsf/Publish/1203/$File/Kellogg.pdf

Berry M.J.A. and Linoff G., 1997. Data Mining Techniques for Marketing, Sales, and Customer Support. John Wiley & Sons. New York.

Bradley, P.S., Fayyad, V., and Reina, C., 1998. Scaling Clustering Algorithms to Large Databases. Proceeding of the 4th International Conference on Knowledge Discovery and Data Mining, (KDD-98). AAAI Press.

Chatfield, C. and Goldhardt, G.J., 1970. The Beta-Binomial Model for Consumer Purchasing Behaviour. Applied Statistics. Volume 19, Number 3. pp. 240-250.

Chen, M., Han, J., and Yu, P., 1996. Data Mining: An overview from database perspective. IEEE Transactions on Knowledge and Data Engineering. Volume 8, Number 6, pp. 866-833.

Clarke, R., 1993. Profiling: A Hidden Challenge to the Regulation of Data Surveillance. Dept. of Computer Science, Australian National University, Canberra, Australia.

Deogun, J.S., Raghavan, V.V., and Sarkar, A., and Sever, H., 1997. Data Mining: Research trends, challenges and applications. in Rough Sets and Data Mining: Analysis of Imprecise Data. Boston, MA., pp. 9-45.

East, R., Harris P., Lomax W. and Willson G., 1997. First-Store Loyalty to US and British Supermarkets. Kingston Business School, Kingston University, Kingston, United Kingdom.

Ehrenberg, A.S.C., 1959. The Pattern of Consumer Purchases. Applied Statistics. Volume 8, Number 1. pp. 26-41.

Groth, R., 1998. Data Mining, A Hands-on Approach for Business Professionals. Prentice Hall. Upper Saddle River, New Jersey.

Han, J. and Kamber, M., 2001. Data Mining Concepts and Techniques. Morgan Kaufmann Publishers. San Francisco.

Hartigan, J.A. and Wong, M.A., 1979. Algorithm AS136: A K-Means Clustering Algorithm. Applied Statistics. Volume 28, pp. 100-108.

Kasabov, N., 1996. Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering. MIT Press, Boston.

Kersten, P.R., 1995. The Fuzzy Median and the Fuzzy MAD. Proceeding of 1995 Annual conference of the North American Fuzzy Information Processing Society, pp. 85 – 88.

Lingras, P.J. and Adams, G., 2001. Selection of Time-Series for Clustering Supermarket Customers. Department of Mathematics and Computer Science, Saint Mary's University, Halifax, Nova Scotia.

Lingras, P.J. and Huang, X., 2004. Statistical, Evolutionary, and Neurocomputing Clustering Techniques: cluster-based versus object-based approaches. To appear in AI Review.

Lingras P.J. and Young, L., 2001. Multi-criteria Time-Series based Clustering of Supermarket Customers using Kohonen Networks. Proceedings of the 2001 International Conference on Artificial Intelligence (IC-AI'2001), Las Vegas, Nevada, USA.

MacQueen, J., 1967. Some Methods fir Classification and Analysis of Multivariate Observations. Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability, L.M. Le Cam and J. Neyman, Ed., vol. 1, pp. 281-297.

Mani, D.R., Drew, J., Betz, A., and Datta, P. 1999 Statistics and Data Mining Techniques for Lifetime Value Modeling. Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, USA. pp. 94-103.

Pelleg, D. and Moore, A., 1999. Accelerating Exact k-means Algorithms with Geometric Reasoning. Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, USA. pp. 277-281.

Piatetsky-Shaprio, G. and Frawley, W.J., 1991. Knowledge Discovery in Databases AAAI/MIT Press.

Potts, W.J.E., 1999. Generalized Additive Neural Networks. Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, USA. pp. 194-200.

Russell, S. and Lodwick, W., 1999. Fuzzy clustering in data mining Telco database marketing campaigns. Proceeding of the 1999 18th International Conference of North American Fuzzy Information Processing Society. New York, USA.

Su, M-C., DeClaris, N., Liu, T-K., 1997. Application of Neural Networks in Cluster Analysis. Proceedings of 1997 IEEE International Conference on Computational Cybernetics and Simulation, Volume 1, pp 1 – 6.

Sungjune P., 2000 Neural Networks and Customer Grouping in E-commerce. Proceedings of Academia/Industry Working Conference on Research Challenges. pp. 331 – 336.

Too, L.H.Y., Souchon, A.L. and Thirkell, P.C., 2000, Relationship Marketing and Customer Loyalty in a Retail Setting: A Dyadic Exploration. Aston University, Birmingham, United Kingdom.

Tung, A.K.H., Lu, H., Han, J., and Feng, L, 1999. Breaking the Barrier of Transactions: Mining Inter-Transaction Association Rules. Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, USA . pp. 297-301.

Venkatesh S., Smith A.K., Rangaswany A., 2000, Customer Satisfaction and Loyalty in Online and Offline Environments. PennState Universtiy, University Park, Pennsylvania. http://www.ebrc.psu.edu/papers/pdf/02-2000.pdf.