TEMPORAL MINING OF WEB AND SUPERMARKET DATA USING FUZZY AND ROUGH SET CLUSTERING

by

Rui Yan

MASTER OF SCIENCE IN APPLIED SCIENCE

SAINT MARY'S UNIVERSITY

HALIFAX, NOVA SCOTIA, CANADA

Date of Submission: SEPTEMBER, 2004

Your file Votre référence
ISBN: 0-612-96136-2
Our file Notre référence
ISBN: 0-612-96136-2

Certification

Name: Rui Yan

Degree: Master of Science in Applied Science

Title of Thesis: Temporal Mining of Web and Supermarket Data using Fuzzy and Rough Set Clustering

Examining Committee:

Dr. William E. Jones, Acting Dean of Graduate Studies

Dr. David H. S. Richardson, Program Co-ordinator

Dr. Haiyi Zhang, External Examiner
Acadia University

Dr. Pawan Lingras, Senior Supervisor

Dr. Robert Dawson, Supervisory Committee

Dr. Harold Ogden, Supervisory Committee

Date Certified: June 24, 2004

SAINT MARY'S UNIVERSITY

Date: **September 2004**

Author: **Rui Yan**

Title: **Temporal Mining of Web and Supermarket Data Using Fuzzy and Rough Set Clustering**

Department: **Mathematics and Computing Science**

Degree: **M.Sc.** Convocation: **October** Year: **2004**

Permission is herewith granted to Saint Mary's University to circulate and to have copied for non-commercial purposes, at its discretion, the above title upon the request of individuals or institutions.

Signature of Author

To my family

Contents

List of Tables

List of Figures

Acknowledgements

First, I would like to thank my supervisor, Dr. Pawan Lingras, for his valuable support and thoughtful advice during my studies at Saint Mary's University. He opened the door for me and guided me into the Computing Science world. Thanks to his encouragement and trust in me, I feel he is not only a supervisor but also a mentor and friend. It has been a privilege to work with him and his research team. I cannot express enough thanks to him.

Secondly, I would like to thank the following professors: Dr. Porter Scobey, Dr. Paul Muir, Dr. Stavros Konstantinidis, Dr. Robert Dawson, Dr. Walt Finden, Dr. Art Finbow and Dr. Bert Hartnell for their kind help and valuable suggestions during my studies. All of them are outstanding professors who show great dedication to their teaching careers and think only of the good of the students. I am fortunate to have had the opportunity to study with them.

I want to thank my colleagues: Dr. Ming Zhong, Mr. Chad West, Ms. Jing Xu, Ms. Hui Xu, Ms. Hong Zhao, Ms. Nancy Sun and Mr. Adish Jain for their discussions and friendship. Without them, my life in Halifax would not have been so enjoyable.

In addition, I would like to express my appreciation to the Natural Sciences and Engineering Research Council of Canada (NSERC), the *Faculty of Graduate Studies*, the *Faculty of Science* and the *Department of Math and Computing Science* for the financial support I received during my studies.

I also want to thank the support staff in my department, including Ms. Rose Daurie, our secretary, and Mr. Owen M. Smith, our technician/system administrator, my managers during my time as an intern at IBM, Dr.Bill O'Connel and Dr. Serge Rielau, and all of the other staff who have assisted my Master's studies and research.

Finally, I thank my family, who are always there to support me. I love them and they mean the whole world to me.

Temporal Mining of Web and Supermarket Data using Fuzzy and Rough Set Clustering

Rui Yan

Submitted in September, 2004

Abstract

Clustering is an important aspect of data mining. Many data mining applications tend to be more amenable to non-conventional clustering techniques. In this research three clustering methods are employed to analyze the web usage and super market data sets: conventional, rough set and fuzzy methods. Interval clusters based on fuzzy memberships are also created. The web usage data were collected from three educational web sites. The supermarket data spanned twenty-six weeks of transactions from twelve stores spanning three regions. Cluster sizes obtained using the three methods are compared, and cluster characteristics are analyzed. Web users and supermarket customers tend to change their characteristics over a period of time. These changes may be temporary or permanent. This thesis also studies the changes in cluster characteristics over time. Both experiments demonstrate that the rough and fuzzy methods are more subtle and accurate in capturing the slight differences among clusters.

Chapter 1

Introduction

1.1 Data Mining and its Applications

We live in an information age with an ever-increasing amount of data. All sorts of data are collected and stored so that valuable information can be extracted from them. Today, we have far more information than we can handle: from business transactions and scientific data, to satellite pictures, text reports and military intelligence. Decision-making is based not only on information retrieval, but more importantly on information analysis. Human analysts with no special tools can no longer make sense of enormous volumes of data, that require processing in order to make informed business decisions. Confronted with huge collections of data, we find that new requirements have arisen to help us make better managerial choices. These requirements include the automatic summarization of data, the extraction of the "essence" of stored information, and the discovery of patterns in raw

data. Data mining automates the process of finding relationships and patterns in raw data and delivers results that can either be utilized in an automated decision support system or assessed by a human analyst.

Data mining or knowledge discovery in databases (KDD) has been defined as "the non-trivial extraction of implicit, previously unknown, and potentially useful information from data" [1]. Data mining draws on results from various fields, such as database systems, machine learning, intelligent information systems, statistics, and expert systems. Data mining results are frequently used by companies to optimize marketing campaigns. For example, campaigns can be designed to target specific customer groups.

A recent initiative that makes extensive use of data mining results is the IBM-Safeway project [2]. 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 also offers customers specific discounts. Future applications of data mining will aim to increase customer satisfaction and convenience.

Researchers have studied relational databases and developed many methods and algorithms to perform different data mining tasks. Data mining is being put to use and studied for various types of databases, including relational databases, object-relational databases, object-oriented databases, data warehouses, transactional databases, unstructured and semi-structured repositories such as the World Wide Web (WWW), advanced databases such as spatial databases, multimedia databases, time-series databases, textual databases and even

flat files. Data mining is not specific to any one type of medium or data. There are different types of data format in real databases, such as locations, pictures, time series, etc. Data mining should be applicable to any type of information repository. However, algorithms and approaches may differ when applied to different types of data. Indeed, the challenges presented by different types of data vary significantly. In this research, spatial and time series data are studied, and three different clustering methods are employed to analyze the data sets.

1.2 Temporal Data Mining

Temporal data mining involves interpreting and discovering relationships and patterns from data collected over time [4]. Temporal data usually includes time series data. Most data mining techniques treat data in temporal databases at best as data series in chronological order, and ignore the time values with which the data are stamped [3]. However, valuable information may be missing if the time attributes are ignored. An example given by Chen and Petrounias [3] is the association between butter and bread, *i.e.* people who buy bread also buy butter. If all the supermarket transactions that are available are examined, the association might be found to be true. If, however, the highest concentration of people purchasing butter and bread occurred up to five years ago, then the discovery of the association is not significant for the present and future of supermarket organization. It has also been recognized [4, 5] that time-dependent information is important in data mining, and that temporal patterns or rules should be investigated and discovered from temporal

databases, since they can provide accurate information about an evolving business domain, as opposed to the static approach taken by conventional data mining.

1.3 Data Mining Functionalities

Three important and widely used data mining functionalities and techniques are association, classification, and clustering.

Association analysis is the discovery of the association rules which reveal interesting correlations or relationships in the data set. Businesses are concerned about what to put on sale, how to design coupons, how to place merchandise on shelves in order to maximize profits, etc. These relationships can help managers to make intelligent business decisions. An association rule is an implication in the form $r : X \Rightarrow Y$, where X and Y are sets of items referred to as *itemsets* and $X \cap Y = \phi$. Each rule r is associated with a confidence factor (α) and a support (s). The confidence factor (α) is the ratio of the number of transactions containing $X \cup Y$ to the number of transactions containing X. The support (s) is the percentage of transactions in the database containing $X \cup Y$ [7]. The apriori algorithm is the most well-known algorithm for mining association rules, and is used for most commercial products. The basic idea is to generate candidate itemsets of a particular size and then scan the database to count these to see if they are large [7]. A spatial association rule is a rule indicating a certain association relationship among a set of spatial and possibly some non-spatial predicates [11]. Koperski and Han [11] explore the efficient mining of spatial association rules at multiple approximation and abstraction levels. They propose first to

perform a less costly, approximate spatial computation to obtain approximate spatial relationships at a high level of abstraction, and then to refine the spatial computation only for those data or predicates the refined computation of which, according to the approximate computation, may contribute to the discovery of strong association rules. However, their algorithm is based on the assumption that users have a reasonably good knowledge of what they want to find, and that good knowledge exists (such as concept or operation hierarchies) for non-spatial or spatial generalization.

Classification is a data mining technique which involves the analysis of data to find rules that describe the partition of the database into a given set of classes [8]. The objective of the classification is first to analyze the training data and to develop an accurate description or model for each class by using the features available in the data. Examples of classification applications include image and pattern recognition, medical diagnosis, loan approval, detecting faults in industry applications, and classifying financial market trends [7]. Many classification methods have been proposed for relational databases, including decision tree based algorithms, neural network based algorithms and statistics based algorithms such as regression and Bayesian classification. In the process of spatial classification, the goal is to find rules that partition a set of classified objects into a number of classes using not only non-spatial properties of the classified objects, but also spatial relationships of the classified objects to other objects in the database. Fayyad uses decision tree methods to classify images of stellar objects, to detect stars and galaxies [9]. However, this method is not suitable for the analysis of vector data formats, often used in geographic information systems.

Ester *et al.* [10] proposes an algorithm based on the ID3 algorithm (ID3 uses the method top-down induction of decision trees), however his method does not analyze aggregate values of non-spatial attributes for neighboring objects, and does not perform relevance analyses. Thus, it may produce an overspecialized, poor quality tree. Koperski *et al.* [8] has analyzed the above algorithms and has proposed an efficient two-step method which concentrates on building decision trees. This approach to spatial classification is based on both (1) non-spatial properties of the classified objects and (2) attributes, predicates and functions describing spatial relations between classified objects and other features located in the spatial proximity of the classified objects. Experiments show that the accuracy of the classification increases dramatically, and the time required to build the decision tree is also reduced significantly.

Cluster analysis is one of the basic tools used 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.

Existing clustering algorithms can be classified into two main categories [14]: *hierarchical* methods and *partitioning* methods. Hierarchical algorithms create a hierarchical decomposition of a database D. The hierarchical decomposition is represented by a dendrogram, a tree that iteratively splits D into smaller subsets until each subset consists of only one object. In such a hierarchy, each level of the tree represents a clustering of D. Hierarchical methods are either agglomerative (proceeding from the leaves to the root

by merging) or divisive (proceeding the root to the leaves by dividing). The single-link method is a commonly used agglomerative hierarchical clustering method. Unfortunately, the runtime for this algorithm is very extensive for large databases. Partitioning algorithms construct a partition of a database D of n objects into a set of k clusters. The partitioning algorithms typically start with an initial partition of D and then use an iterative control strategy to optimize an objective function. Well-known clustering methods such as K-means and K-medoids are partitioning algorithms. However, these statistical algorithms are not efficient for high-dimensional data. Kaufman and Rousseeuw [15] developed a partitioning around medoids (PAM) approach, which determines a medoid for each cluster. A medoid can be defined as that object of a cluster, whose average dissimilarity to all the objects in the cluster is minimal. A medoid exists in the data set. Their method works satisfactorily for small data sets, but is not efficient in dealing with medium to large data sets. They later proposed an improved method referred to as clustering large applications (CLARA), to handle large data sets. This method relies on sampling. It improves the converging performance, however if the real medoid is not chosen as a sample, this method cannot obtain an optimal result. Ng and Han [14] explored partitioning algorithms referred to as clustering large applications based on randomized search (CLARANS), which improved the clustering process. The difference between CLARANS and PAM is that the former checks only a sample of the neighbors of a node. Also, unlike CLARA, CLARANS takes a sample of neighbors in each step of a search. This has the advantage of not confining the search to a localized area. Experiments have confirmed that the CLARANS approach is significantly

more efficient than PAM or CLARA.

One conventional statistical clustering technique, the K-means algorithm, and two important non-conventional clustering techniques are employed in the present research for temporal data analysis. Experiments show the accuracy and efficiency of the non-conventional methods as compared with the conventional clustering method.

1.4 Organization of the Thesis

This study focuses on clustering techniques in temporal data mining. A conventional clustering method is compared with the rough and fuzzy methods in analysis of the online behaviors of web users and the shopping activities of supermarket customers. The methods and techniques used in this study are introduced in Chapter 2. Detailed experiments analyzing website users are discussed in Chapter 3. In Chapter 4 the shopping behavior of supermarket customers is analyzed, and the hidden migrations of customers over a six-month period are discovered. Fuzzy and conventional clustering methods are compared in this monthly analysis. This research shows that the three methods successfully identify the clusters in the two different types of data sets. The conventional K-means technique groups each object in precisely one group. The fuzzy method calculates memberships for each object in each cluster. The rough method differentiates objects with respect to the lower and upper bounds, making it possible to provide a rough or unclear boundary for each cluster. Both experiments demonstrate that the rough and fuzzy methods are more subtle in capturing the slight differences among clusters.

Chapter 2

Methods and Techniques

2.1 Methods Review

Clustering is a process of partitioning or grouping a given set of unlabeled patterns into a number of clusters such that similar patterns are assigned to one cluster. The conventional methods lead to crisp clustering (or hard clustering), where there are well-defined boundaries between clusters. Crisp clustering assigns each data point (or feature vector) to one and only one of the clusters. The degree of membership for each data point is either one or zero. However, in the real world, the distinction between clusters tends to be fuzzy or rough. There is a likelihood that an object may be a candidate for more than one cluster, and the characterization of clusters may not be precisely defined.

Joshi and Krishnapuram [16] have argued that clustering operation in web mining involves modeling an unknown number of overlapping sets. Similar arguments can also be

extended to spatial and temporal data mining. Fuzzy clustering is one attractive solution for specifying fuzzy memberships of objects in a cluster. Rough sets provide an alternative method of representing overlapping sets.

Lingras [27] has described how a rough set theoretical clustering scheme could be represented using a rough set genome. The resulting genetic algorithms (GAs) are used to evolve groupings of highway sections represented as interval or rough sets. Lingras [28] has applied unsupervised rough set clustering based on GAs in order to group users of a first year university course website. He hypothesizes that there are three types of visitors to the website: studious, crammers, and workers. Studious download notes from the site regularly. Crammers download most of the notes before an exam. Workers come to the site to finish assigned work such as laboratory and class assignments. Generally, the boundaries of these clusters is not precise. Preliminary experimentation by Lingras [28] has illustrated the feasibility of rough set clustering for developing profiles of website users. However, the clustering process based on GAs seems too computationally expensive for scaling to a larger data set.

The Kohonen neural network or self-organizing map [30] is another widely used clustering technique. The Kohonen network is advantageous for some applications due to its adaptive capabilities. Lingras *et al.* [31] have introduced interval set clustering using a modification of the Kohonen self-organizing maps, based on rough set theory. The proposed algorithm is used to find cluster intervals for web users. The three websites used for the experiments were websites for two first year courses and one second year course.

The students used the websites for downloading class notes and lab assignments; for downloading, submitting and viewing class assignments; for checking their current marks; and for accessing a discussion board. The websites were accessed from a variety of locations. Only some of the website accesses were identifiable by student ID. Therefore, instead of analyzing individual students, it was decided to analyze each visit. This also made it possible to guarantee the required protection of privacy. Lingras *et al.* [31] have also provided a comparison of the user behavior of first and second year students. The experiments show that the modified Kohonen network provides reasonable cluster interval sets by adjusting to changing user behavior.

Lingras and West [29] have provided a theoretical and experimental analysis of a modified K-means clustering approach based on the properties of rough sets. This method is used to classify the visitors to an academic website into the upper and lower bounds of the three classifications mentioned above: studious, crammers, and workers. The modified K-means approach is suitable for large data sets.

Fuzzy C-means (FCMs) is a landmark algorithm in the area of fuzzy C-partition clustering, It was first proposed by Bezdek in 1981 [17]. It is based on the minimization of an objective function with respect to the membership U and the cluster center V. Rhee and Hwang [18] have proposed a type-2 fuzzy C-means algorithm to solve the membership typicality. They point out that since the memberships generated are relative numbers, they may not be suitable for applications in which the memberships are supposed to represent typicality. Moreover, the conventional fuzzy C-means process suffers from noisy data, *i.e.*,

when a noise point is located far from all the clusters, an undesirable clustering result may occur. This algorithm is based on the fact that higher membership values should contribute more than memberships with smaller values, when updating the cluster centers [18].

Another fuzzy C-means algorithm provides an improvement by finding a better initial cluster center [19, 26]. Both methods choose samples from whole data sets. The difference is in how the samples are selected. The method proposed by Cheng *et al.* [26] is called multistage random sampling FCMs (mrFCMs). The mrFCMs has two phases, where phase I is a multistage iterative process of a modified FCMs, and phase II is a standard FCMs with the cluster centers initialized by the cluster center values obtained from phase I [26]. There are four factors that must be determined prior to execution. The first factor is the size of the subsamples, $X_{\Delta\%}$. The second is the number of stages, n. The final size of the data set for mrFCMs phase I is $X_{n\times\Delta\%}$. The other two factors are the stopping condition for the first stage of mrFCMs phase I, $\epsilon_{firststage}$, and the stopping condition for the last stage of mrFCMs phase I, $\epsilon_{laststage}$. Another method, proposed by Hung and Yang [19], is partition simplification FCMs (psFCMs). This algorithm uses a simplified set of the original complete data set to find the actual cluster center. This algorithm also consists of two phases: phase I is a sequence of processes that refines the initial cluster centers. The data set is partitioned into several unit blocks by using the k-d tree method. There must be at least one pattern in each unit block. Thus, the actual number of unit blocks depends on the size and pattern distribution of the data set. For each unit block, the centroid of patterns in the unit block is calculated, and is used to represent all the patterns in the unit block. This

allows data set X_N to be dramatically reduced to a simplified data set X_{PS}, containing the centroids of the original patterns. In phase II, the FCMs algorithm is applied to find the cluster centers of the simplified data set $\bar{X_{ps}} = (\bar{x_1}, \bar{x_2}, ..., \bar{x_{ps}})$, $\bar{x_i} \in R^f$. The number of centroids in the simplified data set is N_{ps}. This is equivalent to the number of unit blocks N_{ub}. Since $N_{PS} < N$, the number of calculations of the norm distance may be reduced, which reduces the overall computation time. The mrFCMs is based on the assumption that a small subset of a data set of feature vectors can be used to approximate the cluster centers of the complete data set. If the actual cluster center is not in the sample chosen, the converging speed will be slower. However, psFCMs reduces the data by splitting the whole data set into several unit blocks. The whole data set is in the unit blocks. The samples, from the unit blocks, can thus represent the whole data set.

In the present research, the conventional K-means clustering method, the modified K-means method proposed in [29], and fuzzy C-means clustering [25, 26] are applied to the three educational websites analyzed earlier by Lingras *et al.* [31] and to three supermarket transactions. The resulting fuzzy and rough clusters provide a reasonable representation of user and customer behaviors for the three websites and supermarkets. The experimental results also demonstrate the good performance of the rough and fuzzy methods. Chapters 3 and 4 discuss the details of the experiments. The three methods employed are introduced in the following sections.

2.2 Review of K-Means

K-means is a least-squares partitioning method, allowing users to divide a collection of objects into K groups. It generates a specific number of disjoint, flat (non-hierarchical) clusters. The K-means method is numerical, unsupervised, non-deterministic and iterative. It is a conventional method, where one object is assigned to one and only one cluster. There are K clusters and there is no overlap between clusters. Every member of a cluster is closer to its cluster than to any other clusters because closeness does not always involve the 'center' of the clusters.

It is assumed 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 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 the minimum value of the distance $d(\mathbf{x}, \mathbf{v})$ between the object vector $\mathbf{v}$ and the cluster vector $\mathbf{x}$. The distance $d(\mathbf{x}, \mathbf{v})$ can be the standard Euclidean distance. After the assignment of all the objects to various clusters, the new centroid vectors of the clusters are calculated as:

$$x_j = \frac{\sum_{object\ v\ was\ assigned\ to\ cluster\ x} v_j}{\mid \mathbf{x} \mid}, \tag{2.1}$$

where $1 \leq j \leq m$. Here $\mid \mathbf{x} \mid$ is the cardinality of cluster $\mathbf{x}$. The process stops when the centroids of the clusters stabilize, *i.e.,* the centroid vectors from the previous iteration are identical to those generated in the current iteration.

The K-means algorithm operates as follows:

- Step 1: Given the cluster number k, Randomly choose the k objects as the initial cluster centroids.
- Step 2: Calculate the distance $d(\mathbf{x}, \mathbf{v})$ between each object $\mathbf{v}$ and the centroid $\mathbf{x}$.
- Step 3: Compare the distances, and assign the objects to the clusters which have the shortest distances.
- Step 4: Update the cluster centroid vectors by equation 2.1.
- Step 5: If the centroid vectors are stable, process stops. Otherwise, go to step 2.

For example, assuming there are 10 two-dimensional data sets as shown in Figure 2.1 and it is decided to group these data points into 3 groups. First, data points 1, 2 and 3 are randomly chosen as the centroid vectors. The Euclidean distance between each data point and the three centroids is calculated. In this example, for centroid 1, data points 4 and 5 are the closest points; for centroid 2, data points 6, 7, and 8 are the closest; and for centroid 3, data points 9 and 10 are the closest. Therefore, in the first round, these data

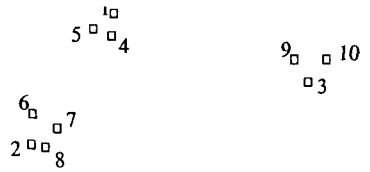


Figure 2.1: K-means: sample 1

points are clustered into the three groups shown in Figure 2.2. In the second step, the new

centroids are calculated as the mean of the data points for each corresponding group. Thus, for group 1, consisting of data points 1, 4, and 5, the centroid is the small dot a. Similarly, for the other two groups, b and c are the centroids. The distance between each data point and the three new centroids, a, b, and c is recalculated. In this example, the centroids of the clusters are finalized, because the elements in each group do not change in the second round. The K-means clustering process is finished when the centroids of the vectors are stabilized. It can be seen that the selection of the centroid vectors is critical for the speed

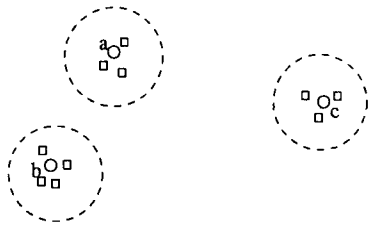


Figure 2.2: K-means: sample 2

of the converging process. For example, if points 1, 4 and 5 are chosen as the centroids, the converging process will take a little longer time.

2.3 Review of Rough Set Theory

The notion of the rough set was proposed by Pawlak [34]. First the concept of rough set theory will be reviewed, and then the details of the second method applied in this research, modified K-means, will be described.

This section provides a brief summary of the concepts of rough set theory essential for introducing the rough set theoretical K-means algorithm.

Let U denote the universe, a finite ordinary set, and let $R \subseteq U \times U$ be an equivalence (indiscernibility) relation on U. The pair $A = (U, R)$ is referred to as an approximation space.

The equivalence relation R partitions the set U into disjoint subsets. Such a partition of the universe is denoted by $U/R = E_1, E_2, ..., E_n$, where E_i is an equivalence class of R. If two elements $u, v \in U$ belong to the same equivalence class $E \subseteq U/R$, then u and v are said to be indistinguishable. The equivalence classes of R are called elementary or atomic sets in the approximation space $A = (U, R)$. The union of one or more elementary sets is called a composed set in A. The empty set $\emptyset$ is also considered to be a special composed set. $Com(A)$ denotes the family of all composed sets.

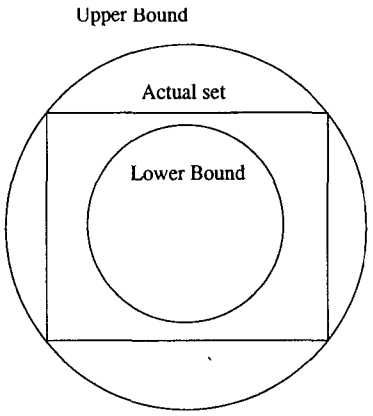


Figure 2.3: Rough set

Since it is not possible to differentiate the elements within the same equivalence class, it may not be possible to obtain a precise representation for an arbitrary set $X \subseteq U$ in terms of elementary sets in A. Instead, any X may be represented by its lower and upper

bounds. The lower bound $\underline{A}(X)$ is the union of all the elementary sets which are subsets of X, and the upper bound $\overline{A}(X)$ is the union of all the elementary sets which have a non-empty intersection with X. The pair $\left(\underline{A}(X), \overline{A}(X)\right)$ is the representation of an ordinary set X in the approximation space $A = (U, R)$, or simply the rough set of X. As shown in Figure 2.3, the elements in the lower bound of X definitely belong to X, while the elements in the upper bound of X may or may not belong to X. Therefore, elements in the lower bound represent the main characteristics of the group. Elements in the boundary area $(\overline{A}(X) - \underline{A}(X))$ exhibit characteristics of more than one group.

Rough set theory, which distinguishes the elements in the lower and upper bounds, can help to analyze customer types. For example, the fact that customers are positioned in the lower bound of group A shows that these customers are loyal to this group. If group A is a high profit group, the manager may consider granting these customers more credit, so that the customers will stay in this group. However, in the case of customers in the boundary, the manager should study these customers to see whether they are changing from the high profit to the lower profit area, or the reverse, or neither. As well, the required policies should be implemented to attract these mobile customers to join the more loyal group. The conventional K-means clustering method does not provide such information. Therefore, a combination of K-means and rough set theory is proposed.

2.4 Modified K-Means Based on Rough Set Theory

Rough sets have been proposed using equivalence relations. However, it is possible to define a pair of upper and lower bounds $(\underline{A}(X), \overline{A}(X))$ or a rough set for every set $X \subseteq U$ as long as the properties specified by Pawlak [34] are satisfied. Yao *et al.* [35] have described various generalizations of rough sets obtained by relaxing the assumptions of an underlying equivalence relation. Skowron and Stepaniuk [37] have discussed a similar generalization of rough set theory. If a more restrictive view of rough set theory is adopted, the rough sets developed in this paper may have to be regarded as interval sets. However, many of the verifiable properties of rough sets within the context of unsupervised learning are obeyed by the interval clusters in this thesis. Therefore, the term rough set is used in the rest of the thesis.

Incorporating rough sets into K-means clustering requires the addition of the concept of lower and upper bounds. Calculations of the centroids of clusters need to be modified to include the effects of lower as well as upper bounds. The modified centroid calculations for rough sets are given by:

$$x_j = \begin{cases} w_{lower} \times \frac{\sum_{v \in \underline{A}(x)} v_j}{|\underline{A}(\mathbf{x})|} + w_{upper} \times \frac{\sum_{v \in (\overline{A}(x) - \underline{A}(x)} v_j}{|\overline{A}(\mathbf{x}) - \underline{A}(\mathbf{x})|} & \text{if } \overline{A}(\mathbf{x}) - \underline{A}(\mathbf{x}) \neq \phi \text{ and } \underline{A}(\mathbf{x}) \neq \phi; \\ \frac{\sum_{v \in (\overline{A}(x) - \underline{A}(x)} v_j}{|\overline{A}(\mathbf{x}) - \underline{A}(\mathbf{x})|} & \text{if } \overline{A}(\mathbf{x}) - \underline{A}(\mathbf{x}) \neq \phi \text{ and } \underline{A}(\mathbf{x}) = \phi; \\ \frac{\sum_{v \in \underline{A}(x)} v_j}{|\underline{A}(\mathbf{x})|} & \text{if } \overline{A}(\mathbf{x}) - \underline{A}(\mathbf{x}) = \phi \text{ and } \underline{A}(\mathbf{x}) \neq \phi. \end{cases} \tag{2.2}$$

where $1 \leq j \leq m$. The parameters w_{lower} and w_{upper} correspond to the relative importance of the lower and upper bounds. $w_{lower} + w_{upper} = 1$. It can be shown that equation 2.2 is

a generalization of equation 2.1. If the upper bound of each cluster were equal to its lower bound, the clusters would be conventional clusters. In this case, the boundary region $\overline{A}(\mathbf{x}) - \underline{A}(\mathbf{x})$ would be empty, and the second term in the equation would be ignored. Equation 2.2 is therefore reduced to the conventional K-means calculation given in equation 2.1. In accordance with rough mereology [37], rough sets are used as patterns for classification. Relevant patterns are discovered by tuning the parameters in such a way that the lower approximation, boundary region, and complement of the upper approximation are relevant, *i.e.*, they are sufficiently included in (or close to) target concepts.

It should be emphasized that the approximation space A is not defined based on any predefined relation on the set of objects. The upper and lower bounds are constructed based on the criteria described above.

The algorithm for the modified K-means based on the rough set theory is shown as follows:

- Step I: Given the cluster number k, Randomly choose the k objects as the initial centroid vectors.
- Step 2: Calculate the distance $d(\mathbf{x}, \mathbf{v})$ between each object $\mathbf{v}$ and the centroid $\mathbf{x}$.
- Step 3: This step determines whether an object belongs to the upper or lower bound of a cluster. For each object vector, $\mathbf{v}$, let $d(\mathbf{v}, \mathbf{x}_i)$ be the distance between it and the centroid of cluster X_i. The differences $d(\mathbf{v}, \mathbf{x}_i) - d(\mathbf{v}, \mathbf{x}_j)$, $1 \leq i, j \leq k$, are used to determine the membership of $\mathbf{v}$. Let $d(\mathbf{v}, \mathbf{x}_i) = \min_{1 \leq j \leq k} d(\mathbf{v}, \mathbf{x}_j)$ and $T = \{j : d(\mathbf{v}, \mathbf{x}_i) - d(\mathbf{v}, \mathbf{x}_j) \leq \mathit{threshold} \text{ and } i \neq j\}$. The value of the threshold is

determined through experimentations.

1. If $T \neq \emptyset$, $\mathbf{v} \in \overline{A}(\mathbf{x}_i)$ and $\mathbf{v} \in \overline{A}(\mathbf{x}_j), \forall j \in T$. Furthermore, $\mathbf{v}$ is not part of any lower bound.
2. Otherwise, if $T = \emptyset$, $\mathbf{v} \in \underline{A}(\mathbf{x}_i)$. $\mathbf{v} \in \overline{A}(\mathbf{x}_i)$.

- Step 4: Update the centroid vectors by equation 2.2.
- Step 5: If the centroid vectors are stable, process stops. Otherwise, go to step 2.

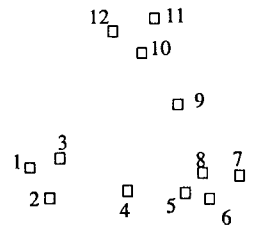


Figure 2.4: Rough K-means: sample 1

Assuming 2-dimensional data points as shown in Figure 2.4, as with the K-means method, it is decided to cluster these data points into 3 groups. Points 1, 6, and 10 are chosen as the centroid vectors for the first round calculation. The Euclidean distance between each data point and the three centroids is calculated. A threshold is set to determine to which group the target point belongs. For example, if the difference between distance (point 1, point 3) and distance (point 3, point 10) is greater than the threshold, then point 3 belongs to the lower bound of the less distant group. As shown in Figure 2.5, point 3 belongs to the lower bound of group A. Similarly, if the difference between distance (point 1, point 4) and distance (point 4, point 6) is less than the threshold, then point 4 is assigned

to the boundary of the two groups. Based on these criteria, the points are assigned to the groups as illustrated in Figure 2.5. In the next round, the centroids are calculated based on equation 2.2 instead of equation 2.1. If the centroids are stable in the next round, the clustering process is finished.

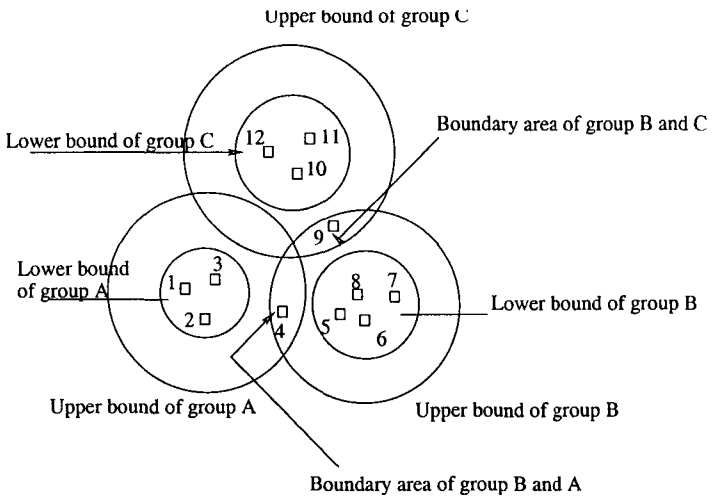


Figure 2.5: Rough K-means: sample 2

The modified K-means method based on rough set theory described above, referred to in this thesis as the rough K-means algorithm (RKMs), depends on the three parameters: w_{lower}, w_{upper}, and $threshold$. Experimentation with various values of the parameters is necessary to develop a reasonable rough set clustering. The design and results of such experiments are described in detail in Chapters 3 and 4. An improved version of this method is presented and a simple simulation is shown in Chapter 3.

2.5 Review of Fuzzy C-Means Clustering

Most real-world classification problems are fuzzy. Fuzzy set theory, to treat fuzziness in data, was proposed by Zadeh in 1965. A fuzzy generalization of clustering uses a fuzzy membership function to describe the degree of membership (ranging from 0 to 1) of an object in a given cluster. There is a stipulation that the sum of the fuzzy memberships of an object in all of the clusters must be equal to 1.

Cannon *et al.* [25] have described an efficient implementation of an unsupervised clustering mechanism that generates the fuzzy membership of objects in various clusters.

The objective of the algorithm is to cluster n objects into c clusters. Given a set of unlabeled patterns: $X = \{x_1, x_2, ..., x_n\}$, $x_i \in R^s$, where n is the number of patterns, and s is the dimension of the pattern vectors (attributes). Each cluster is represented by the cluster center vector V. The fuzzy C-means algorithm minimizes the weighted within group sum of the squared error objective function $J(U, V)$:

$$J(U, V) = \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik}^{m} d_{ik}^{2}. \tag{2.3}$$

where:

- U : the membership function matrix.
- u_{ik}: the elements of U. $u_{ik} \in [0, 1]$, $i = 1, ...n, k = 1, ...c$. $\sum_{k=1}^{c} u_{ik} = 1$, $0 < \sum_{i=1}^{n} u_{ik} < n$.
- V : the cluster center vector, $V = \{v_1, v_2, ..., v_c\}$

- n : the number of patterns.
- c : the number of clusters.
- d_{ik} : the distance between x_i and v_k.
- m : the exponent of u_{ik} that controls fuzziness or the amount of cluster overlap. Gao *et al.* [32] have suggested the use of $m = 2$ in experiments.

The fuzzy C-means algorithm operates as follows.

- Step 1: Given the cluster number c, the initial cluster center V^0 is chosen randomly. The variable m is set to 2; s, the index of the calculations, is set to 0; and the threshold, ϵ, is a small positive constant.
- Step 2: Based on V, the membership of each object U^s is calculated as:

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} (\frac{d_{ik}}{d_{jk}})^{\frac{2}{(m-1)}}}, i = 1, ...n, k = 1, ...c. \tag{2.4}$$

$$d_{ik} = \mid x_k - v_i \mid > 0, \forall i, k. \tag{2.5}$$

  for $d_{ik} = 0$, $u_{ik} = 1$ and $u_{jk} = 0$ for $j \neq i$.

- Step 3: The index s is incremented by one. The new cluster center vector V^s is calculated as:

$$v_i = \frac{\sum_{k=1}^{n} (u_{ik})^m x_k}{\sum_{k=1}^{n} (u_{ik})^m}, \forall i, i = 1, ..n. \tag{2.6}$$

- Step 4: The new membership U^s is computed using equations 2.4 and 2.5, as in step 2.

- Step 5: If $\mid U^s - U^{(s-1)} \mid < \epsilon$, the process is finished; otherwise steps 3, 4, and 5 are repeated.

All three algorithms, conventional K-means, the modified K-means based on the rough set theory, and fuzzy C-means methods, calculate the distance between the centroids and each data point. The difference is that the fuzzy C-means algorithm calculates memberships for each object and a new cluster center vector is calculated based on the memberships. The iteration stops when the memberships for the clusters are finalized, while in the other two methods, the iteration stops when the centroids for the clusters are finalized. All the methods make calls to the distance function $d(\mathbf{x}, \mathbf{y})$. Assuming that the dimensions of the vector are constant at m, the function $d(\mathbf{x}, \mathbf{y})$ requires constant time. The number of calls made to the distance function will determine the order of computational time requirements of each method. The time requirement for K-means method depends on the number of iterations required for the centroid vectors to stabilize. For each iteration, n objects are compared with k clusters, leading to $n \times k$ calls to the distance function. Since k is small and fixed for a given experiment, the time requirement for a single iteration is $O(n)$. For the complete execution of K-means method, the time requirement is $O(n \times iter)$, where $iter$ correspond to the number of iterations [39]. For the rough K-means method, since it has the same calculation process with the K-means, the time complexity is same as that from the K-means method. For the fuzzy C-means method, the system has to calculate the norm distance from each pattern to every candidate cluster center in each iteration. After the distance, the system compute the membership matrix. Therefore, if the dimension of a

dataset is fixed and there are n patterns and k clusters, the time complexity to calculate the membership matrix is $O(n \times k)$ [19]. Assuming k is small and fixed for a given experiment, and $iter$ is the number of iterations, the complexity of the fuzzy C-means clustering will be $O(n \times iter)$.

The next chapter presents experimental results that compare the interval sets created by the three clustering algorithms described above: conventional K-means, rough K-means, and fuzzy C-means.

Chapter 3

Web Data Mining

3.1 Web Mining

Data from the World Wide Web (WWW) can be broadly categorized as content, structure, and usage data. Content data consist of the physical resources on the web, such as documents and programs. Structural data are related to the organization of a website, and to links and relationships between various web resources. Content and structural data represent primary data on the web. Web usage data correspond to the secondary data generated by the interactions of users with the web. Web usage data include data from web server access logs, proxy server logs, browser logs, user profiles, registration files, user sessions or transactions, user queries, mouse clicks, and any other data generated by the interaction between users and the web.

Based on the data sources, web mining can be divided into three classes: content mining, structure mining, and usage mining [21]. Web usage mining applies data mining techniques to discover usage patterns from web data, in order to understand and better serve the needs of web-based applications. Web usage mining involves the creation of user profiles, user access patterns, and navigation paths. The results of web usage mining are employed by e–commerce companies for tracking customer behavior on their sites. Web usage mining consists of three phases: preprocessing, pattern discovery, and pattern analysis.

Clustering analysis is a useful technique in web usage mining, It groups together users or data items with similar characteristics. The clustering process is an important step in establishing user profiles. User profiling on the web consists of studying significant characteristics of web visitors. Due to the ease of movement from one portal to another, web users can be very mobile. If a particular website does not satisfy the needs of a user in a relatively short period of time, the user will quickly move on to another website. Therefore, it is very important to understand the needs and characteristics of web users. Clustering in web mining faces several challenges not present in traditional applications [16]. Clusters tend to have unclear boundaries. The membership of an object in a cluster may not be precisely defined. There is a likelihood that an object may be a candidate for more than one cluster. In addition, due to noise in the recording of data and incomplete logs, there is a high probability that outliers may be present in the data set. In this chapter, experimental analysis using the three methods described in the previous chapter are presented, which

study the activities of web users during an academic term. The extension of the fuzzy C-means algorithm to obtain interval set clustering is described in the following section and comparisons among the three techniques are discussed in sections 3.3 and 3.4. Conclusions are presented at the end of this chapter.

3.2 Extending the Fuzzy C-means Algorithm to Obtain Interval Set Clustering

It is possible to create interval clusters based on the fuzzy memberships obtained using the fuzzy C-means algorithm described in the previous section. Let $1 \geq \alpha \geq \beta \geq 0$. The pattern v_i belongs in the lower bound of cluster k, $\underline{U}(x_k)$, if $u_{ik} \geq \alpha$. Similarly, if $u_{ik} \geq \beta$, the pattern v_i belongs in the upper bound of cluster k, $\overline{U}(x_k)$. In rough set theory, there are three properties describing the membership of objects in the upper and lower bounds:

- P1. An object v can be part of at most one lower bound.
- P2. An object v in the lower bound of a group must also be in its upper bound.
- P3. If an object v is not in the lower bound of any group, it must be in two or more upper bounds.

Since $1 \geq \alpha \geq \beta \geq 0$, if an object belongs to the lower bound of a cluster, it will also belong to its upper bound. If further restrictions are placed on the values of α and β, properties P1 and P3 will also hold. Since the membership values of an object for all of the

clusters sum to 1, no more than one cluster membership can be greater than 0.5. Therefore, if $\alpha \geq 0.5$, an object cannot belong to more than one lower bound; thus property P1 holds, In order to guarantee that property P3 holds, it is necessary to enforce explicitly the condition that if an object belongs to the lower bound of one of the clusters, it cannot belong to the upper bound of any other cluster. Furthermore, β must be set low enough to ensure that at least two memberships are greater than β. If none of the memberships are greater than 0.5, then it can be seen that at least two of the memberships must be greater than or equal to $0.5/(c-1)$, where c is the number of clusters. Therefore, if $\beta \leq 0.5/(c-1)$, property P3 will also hold.

- Rule 1. Let $1 \geq \alpha \geq 0.5 \geq (0.5/(c-1)) \geq \beta \geq 0$. If $u_{ik} \geq \alpha$, then the pattern v_i belongs in the lower and upper bounds of cluster k and does not belong in any other upper bound. Otherwise, the pattern v_i belongs in the upper bounds of clusters k, such that $u_{ik} \geq \beta$. The following theorem can be stated, based on the previous discussion.

Theorem 1. Interval set representations of clustèrs created by Rule 1 will satisfy properties P1 to P3.

3.3 Improved Modified K-Means Based on Rough Set Theory

Another solution for calculating the modified centroid for rough sets takes into consideration the size of the upper and lower bounds of the clusters. Instead of using a fixed weighting factor for the upper and lower bounds, the weighting factors are calculated based on the size of the upper and lower bounds. One possible weighting factor, $\underline{f_j}$ and $\overline{f_j}$, is calculated as follows:

$$\underline{f_j} = \frac{w_{lower}}{w_{lower} \times \mid \underline{A}(\mathbf{x}) \mid + w_{upper} \times (\mid \overline{A}(\mathbf{x}) - \underline{A}(\mathbf{x}) \mid)} \tag{3.1}$$

$$\overline{f_j} = \frac{w_{upper}}{w_{lower} \times \mid \underline{A}(\mathbf{x}) \mid + w_{upper} \times (\mid \overline{A}(\mathbf{x}) - \underline{A}(\mathbf{x}) \mid)} \tag{3.2}$$

The centroid for rough sets can be calculated as follows:

$$x_j = \underline{f_j} \times \sum_{v \in \underline{A}(\mathbf{x})} v_j + \overline{f_j} \times \sum_{v \in (\overline{A}(\mathbf{x}) - \underline{A}(\mathbf{x}))} v_j. \tag{3.3}$$

It can be seen that the weighting factors are dynamically adjusted according to the size of the upper and lower bounds. This can provide enhanced applicability in certain cases, for instance, when the size of the lower bound is much larger than the boundary size. Since dynamically weighted factors involve adjusting the weighting based on the size of the upper and lower bounds, better clustering results may be obtained by using the above formulae. A corresponding simulation is presented in Section 3.5.4.

3.4 Study Data and Design of the Experiment

3.4.1 Data Description

The study data were obtained from the web access logs of three university courses. These courses represent a sequence of required courses for the Computing Science program at Saint Mary's University, Halifax, Canada. The first and second courses were first-year courses, while the third course was a second-year course. The first course, "Introduction to Computing Science and Programming", is offered in the first term, for first-year students. The initial number of students in the course was 180. The number decreased during the term to 130 to 140 students. The students in the course typically come from a wide variety of backgrounds, and include computing science major hopefuls, students taking the course as a required science course, and students taking the course as a science or general elective. As is common in a first-year course, the attitude of students toward the course also varies a great deal. The second course, "Intermediate Programming and Problem Solving", is offered in the second term, for first-year students. The initial number of students in the course was around 100. The number decreased during the term to about 90 students. These students typically have backgrounds and motivations similar to those of the students in the first course, however, the student population is less susceptible to attrition. It was hoped that these subtle differences between the two courses would be reflected in the results of the fuzzy and rough clustering methods. The third course, "Data Structures", is offered to second-year students. The students in this course are core computing science students.

There were 23 students in the course. Here it was hoped that the visit profiles would reflect some of the differences among the students. The primary research undertaken by Lingras [28] and Lingras and West [29] showed that visits from students attending the first course could fall into one of the following three categories.

- Studious visitor: These visitors download the current set of notes. Since they download a limited/current set of notes, they probably study class notes on a regular basis.
- Crammer visitor: These visitors download a large set of notes. This indicates that they have not accessed the class notes for a long period of time. It may be assumed that they are planning for pretest cramming.
- Worker visitor: These visitors are generally working on class or laboratory assignments, or are accessing the discussion board.

The above three categories are not the student types. They indicate the behavior of the web visitor. The same student might show different behavior during the study period. The conventional K-means algorithm is a crisp clustering method, which assigns each web user to precisely one of the three clusters mentioned above. The membership for each cluster is thus 0 or 1. The fuzzy C-means algorithm determines the membership of each visitor in the three clusters, with a level of membership ranging from 0 to 1. The rough K-means method assigns each visitor to the lower bound, upper bound and/or boundary of the three clusters.

Data Set	Hits	Hits After Cleaning	Visits	Visits After Cleaning
First	361609	343000	23754	7619
Second	265365	256012	16255	6030
Third	40152	36005	4248	1274

Table 3.1: Descriptions of data sets

3.4.2 Data Preparation

Data quality is one of the fundamental issues in data mining. Poor data quality always leads to low-quality results. Data preparation is therefore an essential step which must be carried out before applying data mining algorithms. The data preparation for the present research consisted of two phases: data cleaning and data transformation.

The data cleaning involved removing hits from various search engines and other robots. Some of the outliers with a large number of hits and document downloads were also eliminated. This reduced the first data set by 5%. The second and third data sets were reduced by 3.5% and 10%, respectively. Details concerning the data are presented in Table 3.1.

The data transformation required the identification of web visits [28]. Certain areas of the website were protected, and could be accessed only by users using their IDs and passwords. The activities in the restricted parts of the website included submitting user profiles, changing passwords, submitting assignments, viewing submissions, accessing the discussion board, and viewing current class marks. The remainder of the website was publicly accessible. Activities in the public portion of the website involved viewing course information, a lab manual, class notes, class assignments, and lab assignments. If users accessed only the public part of the website, their IDs were unknown. Therefore, web

users were identified based on their IP addresses. This also ensured that user privacy was protected. A visit from an IP address began when the first request was made from the IP address. The visit continued for as long as consecutive requests from the IP address had a sufficiently small delay.

The web logs were preprocessed to create an appropriate representation of each user, corresponding to one visit. The abstract representation of a web user is a critical step that requires a good knowledge of the application domain. Previous studies on the students in the course suggested that some of the students printed preliminary notes before a class, and an updated copy after the class. Some students viewed the notes on-line on a regular basis, while others printed all of the notes shortly before important events such as midterms and final examinations. In addition, there were many visits on Tuesdays and Thursdays, when in-laboratory assignments were due. On-campus and off-campus points of access can also provide some indication of the objectives of a user for the visit. Based on these observations, it was decided to use the following attributes for representing each visitor [28]:

- On-campus/off-campus access.
- Daytime/nighttime access. The daytime period was considered to be from 8 a.m. to 8 p.m.
- Access during lab/class days or non-lab/non-class days. All of the labs and classes were held on Tuesdays and Thursdays. The visitors on these days were more likely to be workers.

- Number of hits.
- Number of class-note downloads.

The first three attributes had binary values of 0 or 1. The values of the last two attributes were normalized. The distribution of the number of hits and the number of class note downloads was analyzed in order to determine appropriate weighting factors. Different weighting schemes were studied. The number of hits was set to be in the range of [0,10]. Since the class notes were the focus of the clustering, the last variable was assigned higher importance, with values in the range of [0, 20].

The total number of visits was 23,754 for the first data set, 16,255 for the second data set, and 4,248 for the third data set. The visits where no class notes were downloaded were eliminated, on the assumption that these visits were made by either casual visitors or workers. Elimination of outliers and visits from search engines further reduced the size of the data sets. After cleaning, the number of visits was 7,619 for the first data set, 6,030 for the second data set, and 1,274 for the third data set, as shown in Table 3.1. Three clustering techniques were applied to the cleaned data. For the fuzzy C-means method, the threshold for stopping the clustering process was set to 10^{-11}, with m equal to 2.

3.5 Results and Discussion

3.5.1 Cluster Analysis

Table 3.2 shows cluster center vectors for the conventional K-means method. It was possible to classify members of the three clusters as studious, workers, and crammers from the results obtained using the conventional K-means algorithm. The first three attributes range from 0 to 1. The last two attributes have higher weights. A comparison of the three clusters shows that the crammers had the highest number of hits and class note downloads in every data set. This is reasonable because crammers do not access class notes for a period of time and then try to download many notes at once for pretest cramming. The average number of notes downloaded by crammers varies from one data set to another. The studious downloaded the second highest number of notes. The distinction between workers and studious for the second course was also based on other attributes. For example, in the second data set, the workers were more likely to come to the campus on laboratory days and access the websites from on-campus locations during the daytime. It is also interesting to note that the crammers had higher ratios of document requests to hits. The workers, on the other hand, had the lowest ratios of document requests to hits.

The fuzzy center vectors are shown in Table 3.3. Table 3.4 shows the rough K-means center vectors. These center vectors are comparable to the conventional K-means center vectors. In order to compare fuzzy and conventional clustering, visits with fuzzy membership values greater than 0.6 were grouped together. Similar characteristics can be found in

Course	Cluster Name	Campus Access	Day/Night Time	Lab Day	Hits	Document Requests
First	Studious	0.67	0.76	0.44	2.97	2.78
	Crammers	0.62	0.72	0.32	4.06	8.57
	Workers	0.67	0.74	0.49	0.98	0.85
Second	Studious	0.00	0.68	0.28	0.67	0.55
	Crammers	0.66	0.72	0.36	2.43	2.92
	Workers	1.00	0.82	0.46	0.66	0.51
Third	Studious	0.69	0.75	0.50	3.87	3.15
	Crammers	0.60	0.71	0.44	5.30	10.20
	Workers	0.62	0.74	0.50	1.41	1.10

Table 3.2: The conventional K-means cluster center vectors

Course	Cluster Name	Campus Access	Day/Night Time	Lab Day	Hits	Document Requests
First	Studious	0.68	0.76	0.44	2.30	2.21
	Crammers	0.64	0.72	0.34	3.76	7.24
	Workers	0.69	0.77	0.51	0.91	0.75
Second	Studious	0.60	0.75	0.13	0.63	0.52
	Crammers	0.64	0.73	0.33	2.09	2.54
	Workers	0.83	0.87	0.75	0.62	0.47
Third	Studious	0.69	0.75	0.50	3.36	2.42
	Crammers	0.59	0.72	0.43	5.14	9.36
	Workers	0.62	0.77	0.52	1.28	1.06

Table 3.3: Fuzzy center vectors

Course	Cluster Name	Campus Access	Day/Night Time	Lab Day	Hits	Document Requests
First	Studious	0.67	0.75	0.43	3.16	3.17
	Crammers	0.61	0.72	0.33	4.28	9.45
	Workers	0.67	0.75	0.49	1.00	0.86
Second	Studious	0.14	0.69	0.03	0.64	0.55
	Crammers	0.64	0.72	0.34	2.58	3.29
	Workers	0.97	0.88	0.88	0.66	0.49
Third	Studious	0.70	0.74	0.48	4.09	3.91
	Crammers	0.55	0.72	0.43	5.48	10.99
	Workers	0.62	0.75	0.51	1.53	1.13

Table 3.4: The rough K-means cluster center vectors

Course	Cluster Name	Campus Access	Day/Night Time	Lab Day	Hits	Document Requests
First	Studious	0.70	0.78	0.45	2.37	2.41
	Crammers	0.65	0.72	0.33	3.74	7.92
	Workers	0.67	0.75	0.50	0.82	0.67
Second	Studious	0.52	0.89	0.00	0.49	0.40
	Crammers	0.65	0.75	0.34	2.18	0.96
	Workers	1.00	1.00	1.00	0.52	0.36
Third	Studious	0.69	0.75	0.51	3.69	2.28
	Crammers	0.58	0.70	0.43	5.38	10.39
	Workers	0.60	0.75	0.52	1.19	1.00

Table 3.5: Average vectors for fuzzy C-means with memberships>0.6

these tables. For the second data set, the modified K-means method is more sensitive to the differences between studious and crammers with regard to the first three attributes than are the other two techniques.

Table 3.5 shows average vectors for the fuzzy C-means method with memberships > 0.6. As expected, Tables 3.3 and 3.5 are similar. Table 3.6 shows average vectors for the fuzzy C-means method with memberships > 0.25. Tables 3.7 and 3.8 show the average cluster vectors for the lower and upper bounds for the modified K-means method. The lower bounds seem to provide more distinctive vectors than any other cluster representation. In a comparison of Tables 3.4, 3.7 and 3.8, it is interesting to note that the conventional centroid vectors seem to lie between the upper and lower bounds of the clusters.

Course	Cluster Name	Campus Access	Day/Night Time	Lab Day	Hits	Document Requests
First	Studious	0.67	0.75	0.44	2.53	2.42
	Crammers	0.63	0.71	0.37	3.84	6.35
	Workers	0.67	0.74	0.49	1.13	0.83
Second	Studious	0.58	0.70	0.21	0.77	0.61
	Crammers	0.63	0.70	0.34	2.10	2.31
	Workers	0.77	0.72	0.59	0.86	0.69
Third	Studious	0.69	0.74	0.48	3.43	2.70
	Crammers	0.64	0.72	0.42	5.17	8.52
	Workers	0.61	0.75	0.51	1.41	1.24

Table 3.6: Average vectors for fuzzy C-means with memberships>0.25

Course	Cluster Name	Campus Access	Day/Night Time	Lab Day	Hits	Document Requests
First	Studious	0.67	0.75	0.43	3.23	3.23
	Crammers	0.60	0.72	0.33	4.29	9.60
	Workers	0.67	0.75	0.49	0.98	0.83
Second	Studious	0.00	0.69	0.00	0.61	0.54
	Crammers	0.63	0.72	0.33	2.64	3.44
	Workers	1.00	0.91	1.00	0.63	0.47
Third	Studious	0.70	0.74	0.48	4.13	4.00
	Crammers	0.55	0.73	0.44	5.49	11.09
	Workers	0.62	0.75	0.51	1.50	1.11

Table 3.7: Average lower bound vectors for rough K-means

Course	Cluster Name	Campus Access	Day/Night Time	Lab Day	Hits	Document Requests
First	Studious	0.67	0.75	0.43	2.96	2.97
	Crammers	0.61	0.73	0.32	4.23	9.10
	Workers	0.67	0.75	0.48	1.08	0.93
Second	Studious	0.55	0.71	0.14	0.73	0.59
	Crammers	0.65	0.72	0.35	2.39	2.83
	Workers	0.88	0.80	0.52	0.73	0.56
Third	Studious	0.69	0.74	0.48	3.97	3.65
	Crammers	0.56	0.71	0.43	5.46	10.70
	Workers	0.62	0.75	0.51	1.63	1.18

Table 3.8: Average upper bound vectors for rough K-means

Course	Cluster Name	Conventional K-Means Clusters	Fuzzy C-Means Memberships >0.6	Fuzzy C-Means Memberships >0.25	Rough K-Means Lower Bound	Rough K-Means Upper Bound
First	Studious	1814	1382	2851	1412	1981
	Crammers	406	414	842	288	339
	Workers	5399	4354	5435	5350	5868
Second	Studious	1699	1750	4163	1197	3871
	Crammers	634	397	1045	443	676
	Workers	3697	1322	3803	1677	4347
Third	Studious	318	265	473	223	299
	Crammers	89	84	140	69	77
	Workers	867	717	871	906	974

Table 3.9: Cardinalities of the clusters for the three techniques

3.5.2 Cluster Cardinalities

Table 3.9 shows the cardinalities of the clusters obtained using the fuzzy C-means, rough K-means and conventional K-means techniques. Table 3.10 shows the cardinality percentages of the three clusters for the three techniques. Since K-means is a conventional hard clustering method, the sum of the cardinalities for the three clusters in each data set is equal to the total number of cardinalities for the data set. For the rough K-means clusters, the sum of the cardinalities for the lower bounds is less than or equal to the total number of cardinalities for the data set. In the case of the fuzzy C-means method, the cardinalities for the clusters are determined by the membership threshold. In this research, it was found that the sum of the cardinalities for the three clusters obtained by the fuzzy C-means and rough K-means methods are subsets of the total data set. For example, for the first course, the sum of the cardinalities is 6150 for the fuzzy C-means method, and 7050 for the rough K-means method. The cardinalities for the upper bound determined by the rough K-means

method are somewhat similar to the cardinalities for the fuzzy C-means clusters, where memberships are greater than 0.25. As shown in Figure 3.1(a), the fuzzy C-means method produced the smallest subset and the rough K-means method produced the second smallest subset for the first and third courses. In contrast, as shown in Figure 3.1(b), the rough K-means method produced the smallest subset, and the fuzzy C-means method the second smallest subset for the second course. The size of the fuzzy C-means subset depends on the threshold for the memberships. If the membership threshold 0.6 is decreased, the sum of the cardinalities for fuzzy C-means clusters will increase.

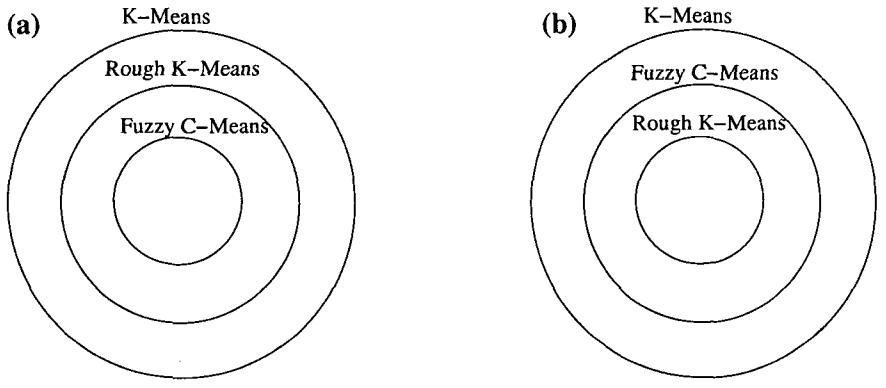


Figure 3.1: Comparison of the cardinalities for the courses

The actual number of members in each cluster varies based on the characteristics of each course. For example, as shown in Figure 3.2, the first-term course had significantly more worker visitors than studious visitors, while the second-term course had more studious visitors than worker visitors. The increase in the percentage of studious visitors in the second term seems to be a natural progression. Interestingly, the second-year course had

Course	Techniques	Total Cardinalities	Percentage of Studious	Percentage of Crammers	Percentage of Workers
First	K-Means	7619	23.81%	5.33%	70.86%
	Fuzzy C-Means	6150	22.47%	6.73%	70.80%
	Rough K-Means	7050	20.03%	4.09%	75.89%
Second	K-Means	6030	28.18%	10.51%	61.31%
	Fuzzy C-Means	3469	50.45%	11.44%	38.11%
	Rough K-Means	3317	36.09%	13.36%	50.56%
Third	K-Means	1274	24.96%	6.99%	68.05%
	Fuzzy C-Means	1066	24.86%	7.88%	67.26%
	Rough K-Means	1198	18.62%	5.76%	75.63%

Table 3.10: Cardinality percentages for the three techniques

a significantly larger number of worker visitors than studious visitors. This seems counter-intuitive. however, it can be explained based on the structure of the websites. Unlike the two first-year courses, the second-year course did not post class notes on web. Thus, the notes downloaded by the students were usually sample programs that were essential during laboratory work. Figures 3.3 and 3.4 show similar characteristics of the three clusters using the K-means and rough K-means methods. A comparison of Figures 3.2, 3.3, and 3.4, indicates that the progression from worker visitors to studious visitors is more obvious with fuzzy clusters than with conventional clusters and rough K-means clusters. The fuzzy C-means method seems more sensitive to the difference between the clusters of studious visitors and worker visitor for the first two courses. The rough K-means method also detects the difference between the two clusters more clearly than the conventional K-means method.

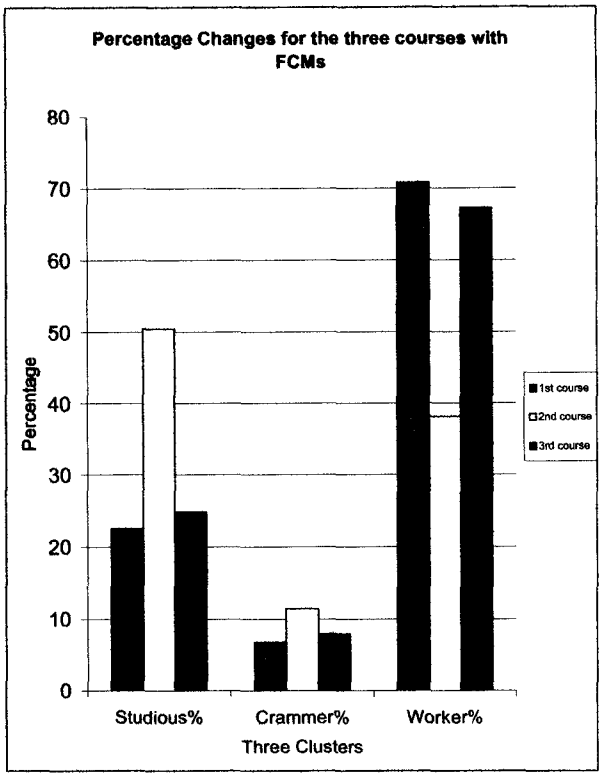


Figure 3.2: Percentage changes for three clusters with FCMs

3.5.3 Overlap Analysis among the Three Techniques

Intersections between conventional clusters and the sets with fuzzy memberships greater than 0.6 provide another indication of the similarity between fuzzy C-means clustering and rough K-means clustering. Table 3.11 shows the ratios of the cardinalities of the intersections: $\frac{|G_c \cap G_k|}{|G_c \cup G_k|}$, where G_c is the set of objects with memberships greater than 0.6 for the

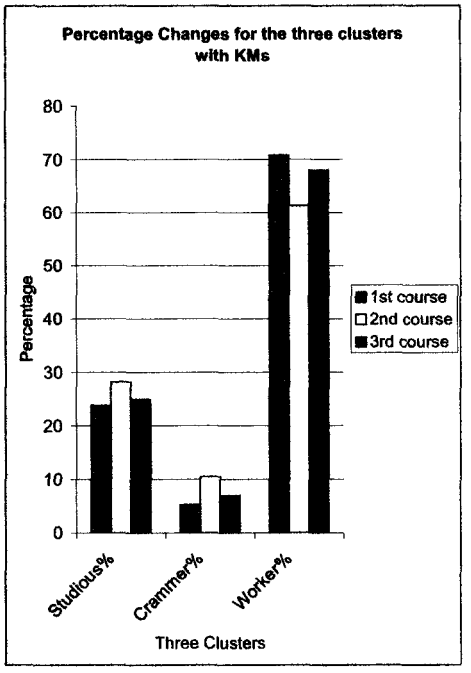


Figure 3.3: Percentage changes for three clusters with KMs

corresponding fuzzy segment, and G_k is the rough K-means cluster. If the two groupings were identical, identity matrices would be obtained. The higher values along the diagonal demonstrate the similarity between the two methods. Somewhat lower values for the first two data sets indicate that the clustering for the first-year courses is fuzzier than that for the second-year course. This observation seems reasonable, since it is easier to characterize the behavior of senior students. The fuzzy representation seems more appropriate for first-year students. Similar observations can be made for Table 3.12.

Courses		Studious (RKMs)	Crammers (RKMs)	Workers (RKMs)
First	Studious (FCM)	0.37	0.00	0.03
	Crammers (FCM)	0.06	0.60	0.00
	Workers (FCM)	0.00	0.00	0.81
Second	Studious (FCM)	0.40	0.00	0.00
	Crammers(FCM)	0.00	0.71	0.00
	Workers (FCM)	0.00	0.00	0.79
Third	Studious (FCM)	0.37	0.00	0.07
	Crammers (FCM)	0.00	0.82	0.03
	Workers (FCM)	0.00	0.00	0.79

Table 3.11: Intersections between clusters using FCMs and RKMs

Courses		Studious (KMs)	Crammers (KMs)	Workers (KMs)
First	Studious (FCM)	0.56	0.00	0.04
	Crammers (FCM)	0.02	0.82	0.00
	Workers (FCM)	0.00	0.00	0.81
Second	Studious (FCM)	0.32	0.00	0.20
	Crammers(FCM)	0.00	0.63	0.00
	Workers (FCM)	0.00	0.00	0.36
Third	Studious (FCM)	0.83	0.00	0.00
	Crammers (FCM)	0.00	0.94	0.00
	Workers (FCM)	0.03	0.00	0.67

Table 3.12: Intersections Between clusters using FCMs and KMs

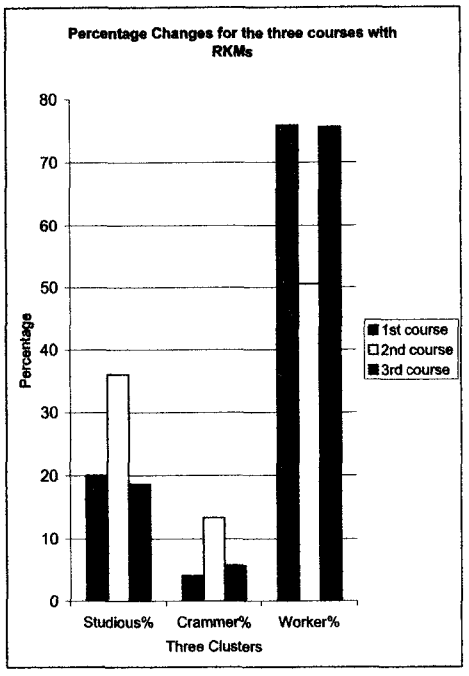


Figure 3.4: Percentage changes for three clusters with RKMs

3.5.4 Cluster Behavior with the Improved Rough K-means Method

Table 3.13 shows the cardinalities of the three clusters for the improved rough K-means method discussed in Section 3.3. It can be seen that the cardinalities of the improved rough K-means method in Table 3.13 are similar to the cardinalities in Table 3.9. Figure 3.5 shows the data distribution for the first data set. Since the last two attributes have more impact on the clustering process, the data set is plotted based on the last two attributes: hits and document requests. Figures 3.6 and 3.7 show the first course cluster distributions for the lower and upper bounds. It is clear that the worker visitors have the fewest hits

Course	Cluster Name	The Improved Rough K-Means Lower Bound	The Improved Rough K-Means Upper Bound
First	Studious	1969	3266
	Crammers	625	800
	Workers	3591	4713
Second	Studious	910	1732
	Crammers	120	159
	Workers	4178	4961
Third	Studious	244	340
	Crammers	159	185
	Workers	717	805

Table 3.13: Cardinalities of the clusters for the improved rough k-means method

and request the lowest number of documents. Crammer visitors have the largest number of hits and document requests. Studious visitors have the second largest number of hits and document requests. The boundary between each of the clusters is quite clear. Visitors for the both the upper and lower bounds are plotted in Fig. 3.8. In this figure, visitors in the upper bound surround those in the lower bound. For example, visitors in the lower bound of the crammers group are plotted as green diamond points in Fig. 3.8, while visitors in the upper bound of the crammers group are shown as red rectangles. It can be seen that the green diamonds are covered by the red rectangles. Studious and worker visitors also exhibit similar characteristics. Figures 3.9 and 3.13 show the data distributions for the other two data sets. Similar results can be found for these two data sets (Figs. 3.9 to 3.16). Here the objects in the lower bound are surrounded by the objects in the upper bound, and the size of the boundary region is not zero. The clustering behavior resulting from this improved rough K-means method is similar to that of the modified K-means based on rough set theory discussed in Section 2.4. Future studies will examine more extreme cases in order

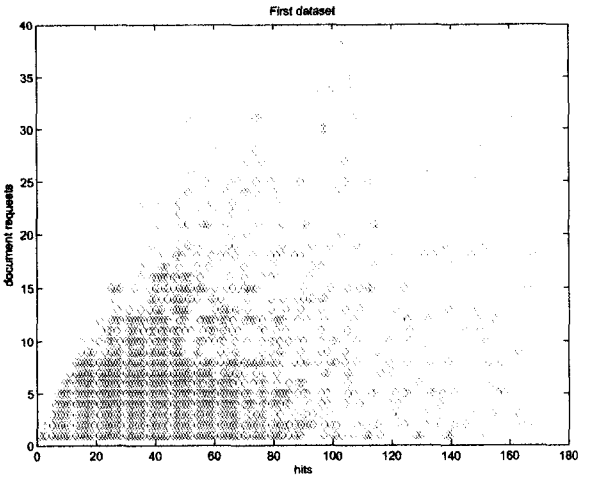


Figure 3.5: Data distributions for the first course

to test the robustness of the improved rough K-means method.

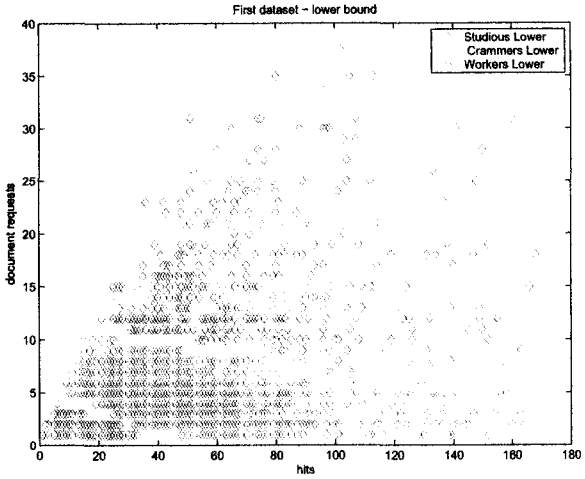


Figure 3.6: First course cluster distributions for the lower bound with the improved RKMs

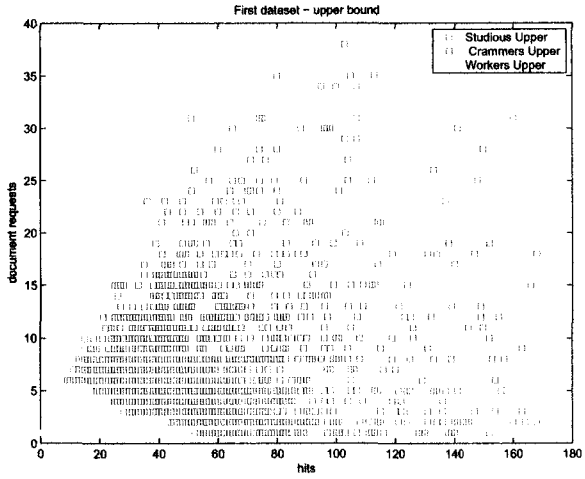


Figure 3.7: First course cluster distributions for the upper bound with the improved RKMs

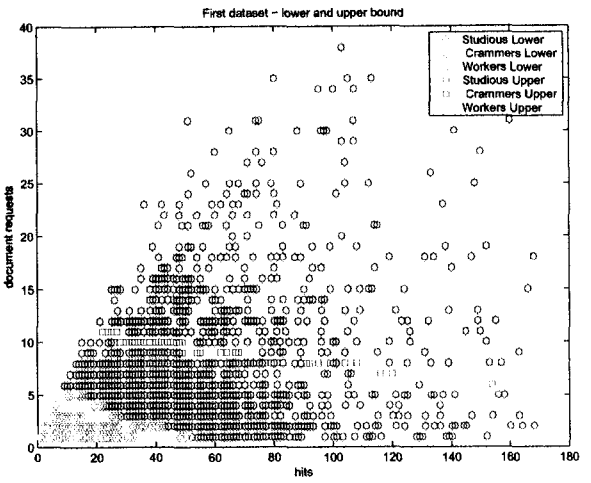


Figure 3.8: First course cluster distributions with the improved RKMs

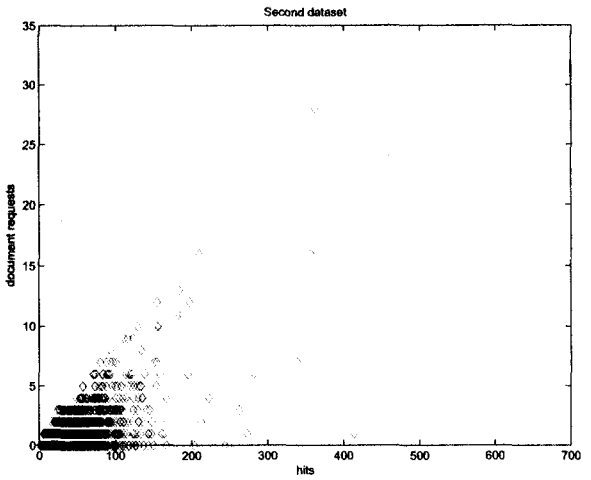


Figure 3.9: Data distributions for the second course

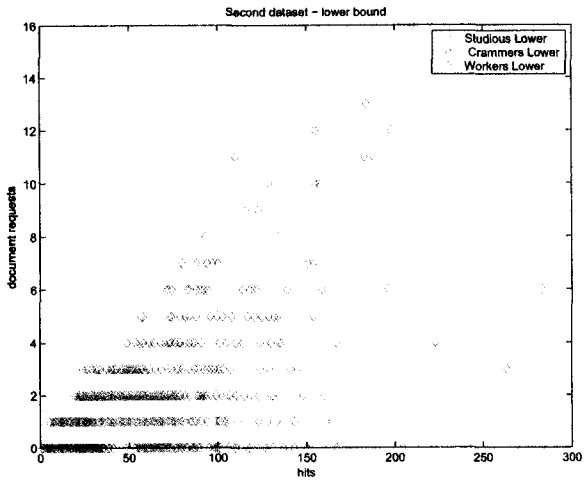


Figure 3.10: Second course cluster distributions for the lower bound with the improved RKMs

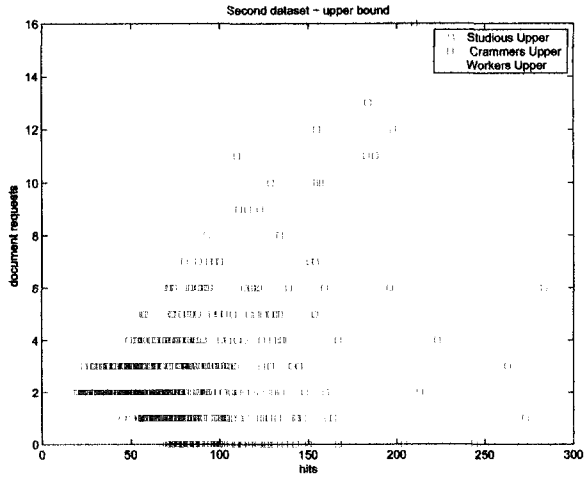


Figure 3.11: Second course cluster distributions for the upper bound with the improved RKMs

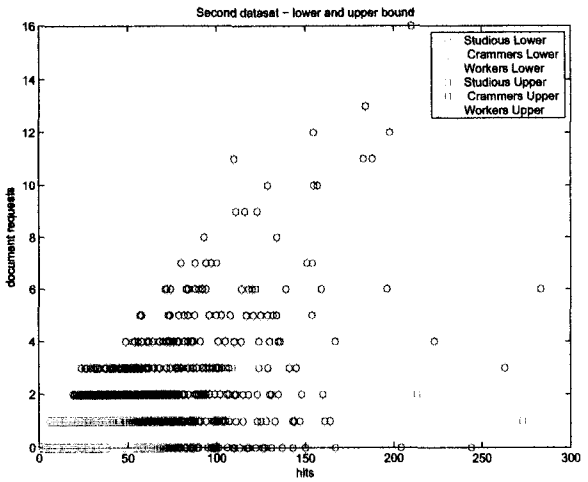


Figure 3.12: Second course cluster distributions with the improved RKMs

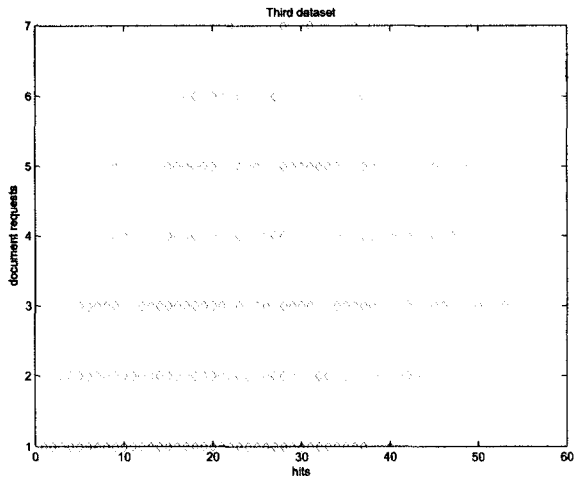


Figure 3.13: Data distributions for the third course

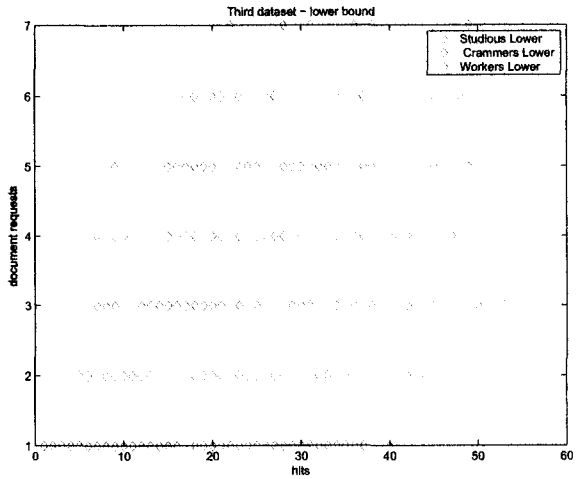


Figure 3.14: Third course cluster distributions for the lower bound with the improved RKMs

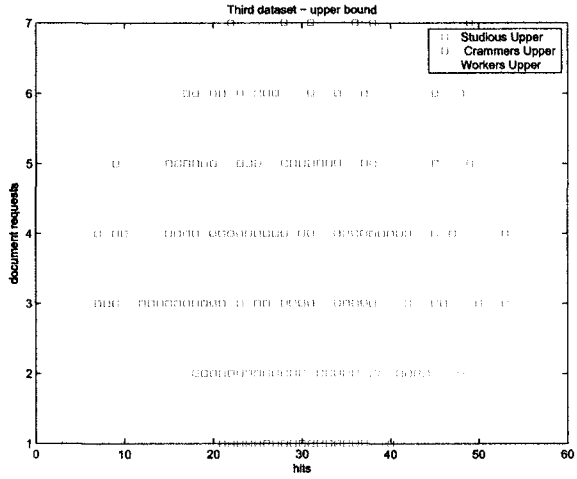


Figure 3.15: Third course cluster distributions for the upper bound with the improved RKMs

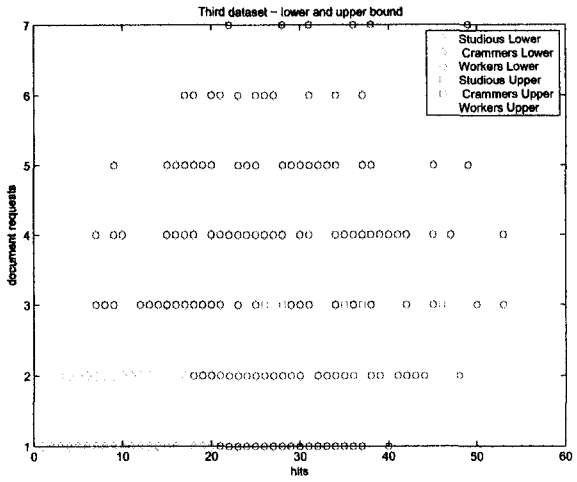


Figure 3.16: Third course cluster distributions with the improved RKMs

Exactly the same experimental setup was used for all three websites. The characteristics of the first two websites were similar. The third website was somewhat different in terms of the site contents, course size, and types of students. The results discussed in this section indicate many similarities among the fuzzy cluster memberships for the three websites. The differences between the results can easily be explained based on further analysis of the websites. It is noteworthy that the fuzzy C-means clustering and the rough K-means clustering were more successful than the conventional K-means algorithm in capturing the subtle differences among the websites. In this experiment, the last two attributes are assigned more weight than the other three attributes. Different weighting schemes may lead to different results. Future work can take different weighting schemes into consideration.

3.6 Summary and Conclusions

This research compares experimental results from a conventional K-means algorithm with results from a fuzzy C-means algorithm and a rough K-means algorithm. Data from visits to three university course websites were used in the experiments. It was expected that the visitors would be classified as studious, crammers, or workers. Since some of the visitors may not precisely belong to one of the groups, the visitors were represented using fuzzy membership functions and rough sets. The experiments produced meaningful clustering of the web visitors using all three clustering techniques. Analysis of the variables used for clustering permitted clear identification of the three clusters as studious, workers, and crammers. Students may exhibit variable attitudes toward the particular websites over a

period of time. The student visits are clustered as the three groups described above. In this experiment, different weighting is assigned to the attributes, causing hits and document requests to have the greatest impact on the clustering process. Different weighting schemes may lead to different clustering results. There were many similarities and a few differences among the characteristics of the conventional clusters, fuzzy clusters and rough sets for the three websites. Subtle differences among the three courses could be identified more easily by using the fuzzy set and rough set representations of the clusters than by using the conventional K-means approach. The groups considered in this study are imprecise. Therefore, the use of fuzzy sets and rough sets seems to provide good results. Another robust rough set method is suggested in this section and a simple simulation is also presented. The corresponding figures illustrate the location of the objects in each cluster. Future work can apply this robust method to some special cases.

Chapter 4

Supermarket Data Mining

4.1 Study Data and Design of the Experiment

Classification and clustering play an important role in supermarket data mining. For example, designing promotional campaigns for individual customers is impractical. It is more feasible to design campaigns for a small number of representative classes. Classifications can be based on many different criteria. Examples of such criteria include the spending potential of customers and their loyalty to the store. The simplest classification is based on the 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:

1. From how many different categories does the customer purchase products? (Examples of categories are meats, fruits and vegetables, etc.)

2. From how many different subcategories does the customer purchase products? (Subcategories are more specific than categories, *i.e.* pork, beef, etc.)

3. How many products does the customer purchase?

4. How much money does the customer spend?

5. How often does the customer visit the store?

With a more complex set of criteria, clustering is more appropriate than classification, at least for the initial analysis. The use of average values for the two variables spending and visits may conceal some of the important information present in the temporal patterns. It is possible that customers with similar profiles may spend different amounts in a given week. However, if the values are sorted, the apparent differences between these customers may vanish. For example, in three weeks customer A may spend $10, $30, and $20 respectively, while customer B spends $20, $10, and $30. If the two time series are compared, these customers might seem to have completely different profiles. However if the time-series values are sorted, the patterns for the two customers will be identical. For this reason, the values of the two variables, spending and visits, were sorted, A 26-week period was used, resulting in a total of 52 values for each customer. A variety of values for K (number of clusters) was 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. Based on spending patterns and variations in visits and discounts, Lingras and Young [38] described the following five customer groups:

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

The results obtained by Lingras and Young [38] indicate that all five 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 due to the use of sorted time series as opposed to single average values of the variables. Lingras and Adams [36] experimented with different combinations of time series to create different clustering schemes. Of the six clustering schemes examined, it was found that a weighted scheme provided the best results. The clustering scheme proposed by Lingras and Adams [36] used reasonable weighting of the spending time series and visit time series. The value of the groceries purchased was found to be a good indicator of customer spending potential. The value time series provides some indication of customer loyalty. However, the visit 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 the groceries did not dominate the clustering. On average, the number of visits was 50 times smaller than the value of the groceries purchased. Since customer spending is more important than the number of visits, it seems reasonable to assign greater importance to

the amount of spending. On the assumption that spending is twice as important as visits, the visit data were multiplied by 25. A reasonable balance between customer loyalty and spending potential was obtained by means of the weighting scheme. A different emphasis can be obtained by changing the weighting of the two time series. The weighting scheme can be expanded to include other time series as well. For example, if value-consciousness is considered an important issue, an appropriate weighting for a discount time series can be assigned. However, there seems to be a limited amount of information gained by including the other three variables: number of categories, number of subcategories, and number of products.

The present study uses the customer representation suggested by Lingras and Adams [36]. The experiment is designed to analyze the customers of twelve supermarkets concentrated in a rural setting. The analysis is used to create interval clusters based on the three algorithms described in Chapter 2. As described in the previous chapter on web mining analysis, comparisons are made among the three methods. A monthly analysis of customer shopping behavior is also performed. The target supermarkets are part of a national chain. The data were collected over a 26-week period beginning in May, 2001. The data collected include information on spending, visits, shopping categories, and other transactional data. In order to test the validity of the results for different regions, data sets from three regions are used. The smallest region has only one store. The database for that store contains 22,447 households and 3,691,611 transaction records. The second region has five stores. The number of households is 29,520 and the number of transaction records is 9,296,004.

The largest region has six stores. A total of 58,982 households shop at these stores and 15,719,786 transactions were recorded.

4.2 Results and Discussion

4.2.1 Cluster Analysis

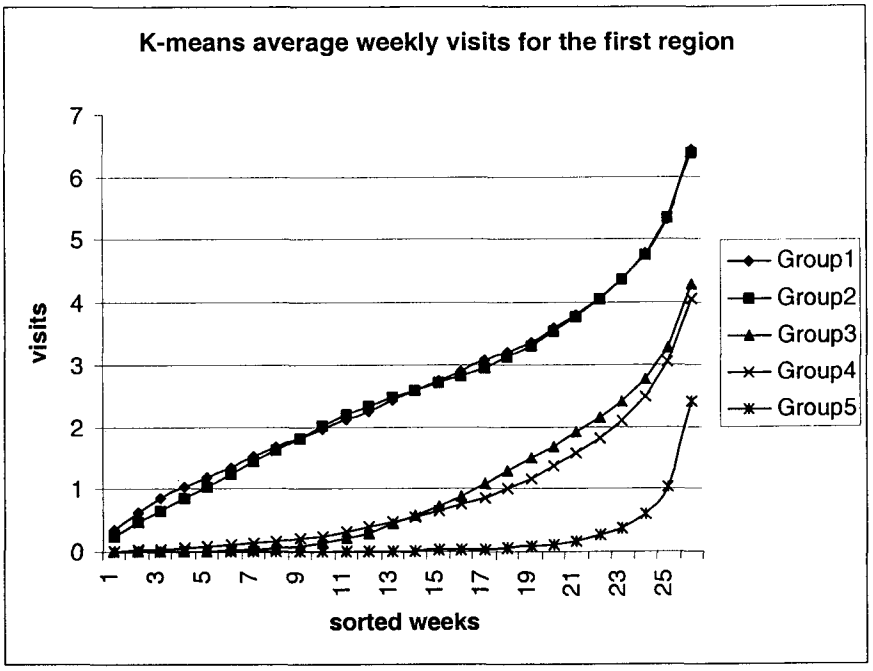


Figure 4.1: K-means average weekly visits for the first region

Figure 4.1 and 4.2 show an analysis of the average weekly visits and spending time series for the five groups for the first region, using the K-means method. Based on the patterns shown in this figure, the groups can be described as follows:

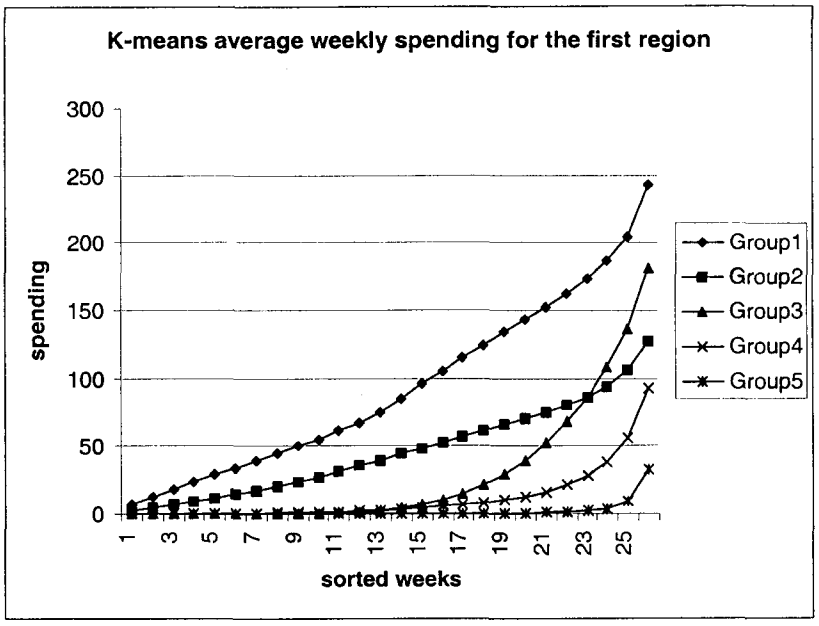


Figure 4.2: K-means average weekly spending for the first region

- Group 1: Loyal big spenders. This group consists of the biggest spenders. In this case, weekly spending ranges from $6 to more than $200. These customers are frequent visitors, sometimes with more than six visits in one week. They seem to be very loyal to the store. Table 4.1 shows the group cardinalities for the first region. There are 1,246 customers assigned to this group. Table 4.2 shows the cardinality percentages. Approximately 5.6% of the customers are in this group. These values differ depending on the region and the method of analysis used. For example, as shown in Table 4.2, for the second and third regions, the percentages for group 1 are 3.47% and 4.87%, respectively.

Cluster Name	Conventional K-Means Clusters	Fuzzy C-Means Mem-berships >0.4	Rough K-Means Lower Bound	Rough K-Means Upper Bound	Rough K-means Boundary
Group 1	1246	1041	1268	1422	154
Group 2	2427	1559	2127	2781	654
Group 3	2309	2537	2238	4041	1803
Group 4	6218	4785	2102	3692	1590
Group 5	10040	10077	11782	12716	1534

Table 4.1: Vector representation of clusters for first region

Cluster	Techniques	Total Cardi-nalities	Percen. of Group1	Percen. of Group2	Percen. of Group3	Percen. of Group4	Percen. of Group5
First	K-Means	22240	5.60%	10.91%	10.38%	27.96%	45.14%
	Fuzzy C-Means	19999	5.21%	7.80%	12.69%	23.93%	50.39%
	Rough K-Means	19517	6.50%	10.90%	11.47%	10.77%	60.37%
Second	K-Means	27637	3.47%	9.80%	16.35%	24.72%	45.66%
	Fuzzy C-Means	25595	6.66%	10.00%	14.16%	22.04%	47.15%
	Rough K-Means	23893	10.64%	7.28%	11.36%	19.67%	51.05%
Third	K-Means	56964	4.87%	11.19%	8.87%	24.95%	50.12%
	Fuzzy C-Means	51505	6.35%	10.39%	15.00%	25.50%	42.76%
	Rough K-Means	48870	7.77%	13.95%	13.46%	11.29%	53.54%

Table 4.2: Cardinality percentages for the three techniques

- Group 2: Loyal moderate spenders. Customers in this group spend less than those in group 1, however the total number of visits is almost identical to the figure for group 1. Even though the maximum spending of these customers is less than in group 3, their spending patterns are the most stable among all the groups. These customers may be the most loyal among all the groups. They are not big spenders like the customers in groups 1 and 3. They are more likely to be value-conscious customers or customers with small families. For the first region, there are 2,427 customers in this group, which is 10.91% of the total data set. For the second and third regions,

9.80% and 11.19% customers, respectively, are in this group.

- Group 3: Semi-loyal potentially big spenders. In terms of the maximum amount spent, this group is comparable to the first group. The 26-week patterns indicate that for around 10 weeks, these customers tend to stay away from the store. There are an additional 10 weeks with limited spending and visits. However, in the remaining six weeks, these customers exhibit increased spending, and their spending is almost the highest. Around 10.38% of the customers in the first region belong to this group. In the second and third regions, 16.35% and 8.87% of the customers, respectively, belong to this group.

- Group 4: Potentially moderate to big spenders with limited loyalty. These customers are similar to those in group 2. However, spending and visits over the 26-week period indicate that these customers are more frequent visitors and spend a little more than those from group 2. It is also possible that they do not always use the supermarket card. The percentages of customers in this group are 27.96%, 24.72%, and 24.95% for the three regions, respectively.

- Group 5: Infrequent customers. Customers in this group are the least loyal. They seem to visit the store only once or twice during 13 weeks. Their spending is limited (less than $40). It is also possible that some of these customers do not use the supermarket card on a regular basis. The majority of customers belong to this group.

Similar characteristics for other regions can be observed from Figures 4.3 to 4.6. The

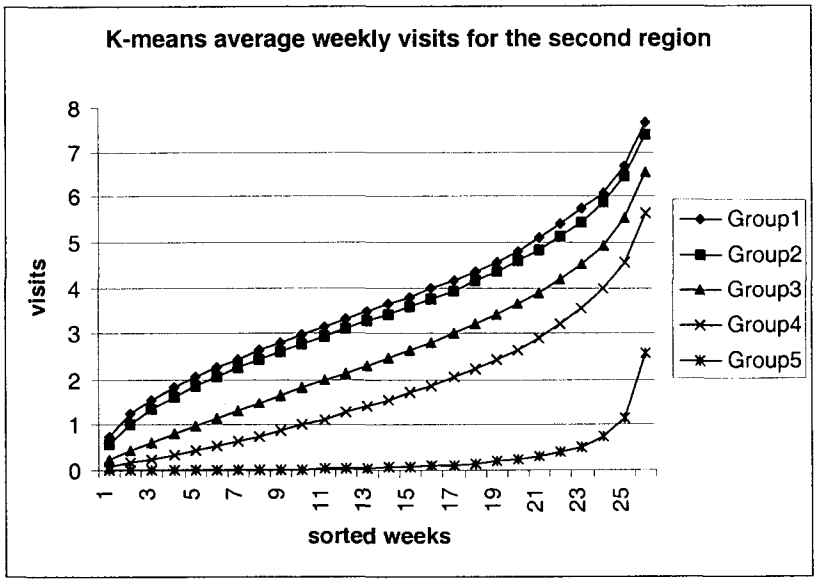


Figure 4.3: K-means average weekly visits for the second region

numbers of visits for groups 1 and 2 are similar to the figures for the other groups. However spending by these two groups is different. For the first and third regions, the spending curves for groups 2 and 3 cross in the 22th week.

Figure 4.7 and 4.8 show the weekly visits and spending obtained using the fuzzy C-means method for the first region. The results are more consistent for the five groups. There is no crossing among the groups. Figures 4.9 to 4.12 show the results for the other two regions using the fuzzy C-means method.

A comparison of the figures for the two clustering methods indicates that the K-means method yields higher values for spending, while the fuzzy C-means method is more sensitive to the number of visits. For example, in Figure 4.7, the number of visits for the five groups is distinguished clearly, while in Figure 4.1 the visits for groups 1 and 2 are similar,

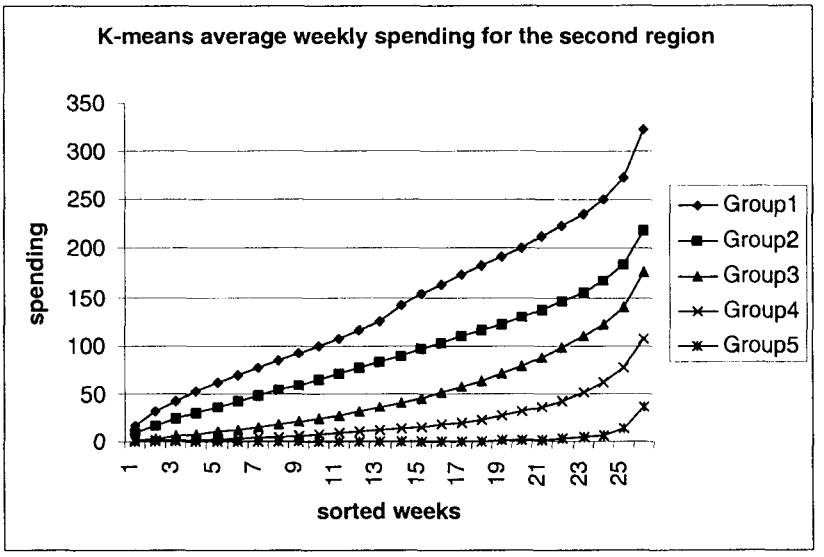


Figure 4.4: K-means average weekly spending for the second region

as are the visits for groups 3 and 4. Moreover, with the fuzzy C-means method, the highest average number of visits for customers in group 1 is almost 7, while the K-means method yields a value of 6. With regard to spending, the highest average value for spending for the most loyal group of customers is found to be around \$225 using the fuzzy C-means method and approximately \$250 using the K-means method.

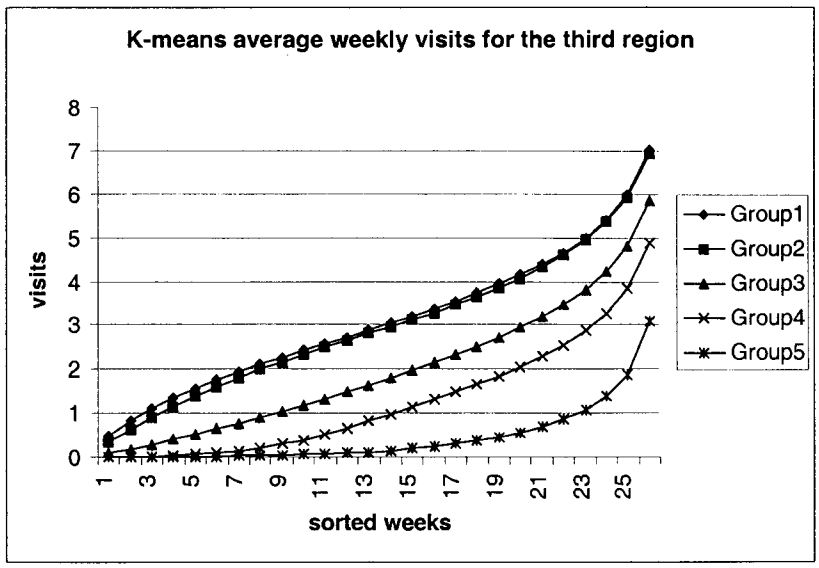


Figure 4.5: K-means average weekly visits for the third region

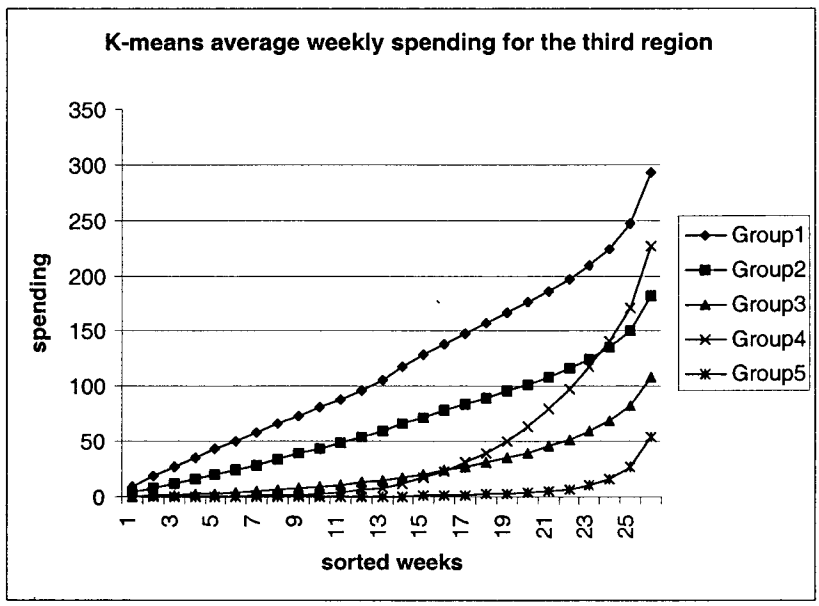


Figure 4.6: K-means average weekly spending for the third region

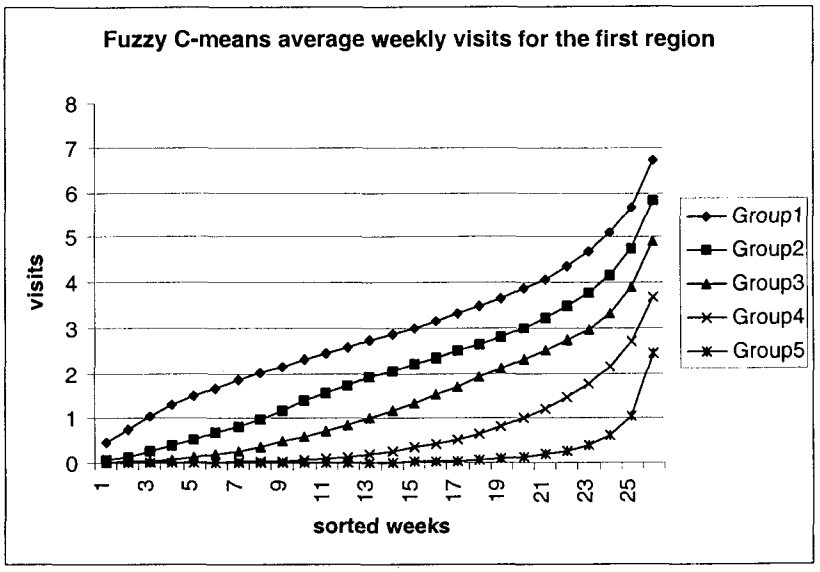


Figure 4.7: Fuzzy C-means average weekly visits for the first region

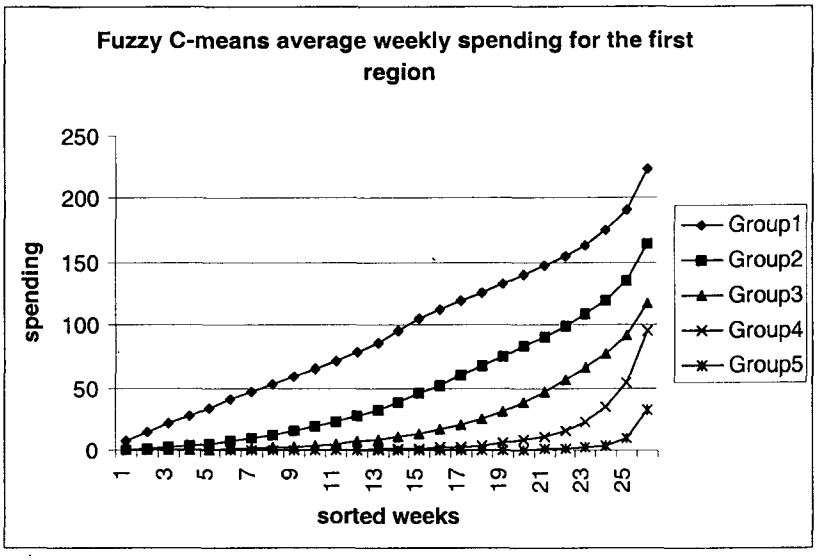


Figure 4.8: Fuzzy C-means average weekly spending for the first region

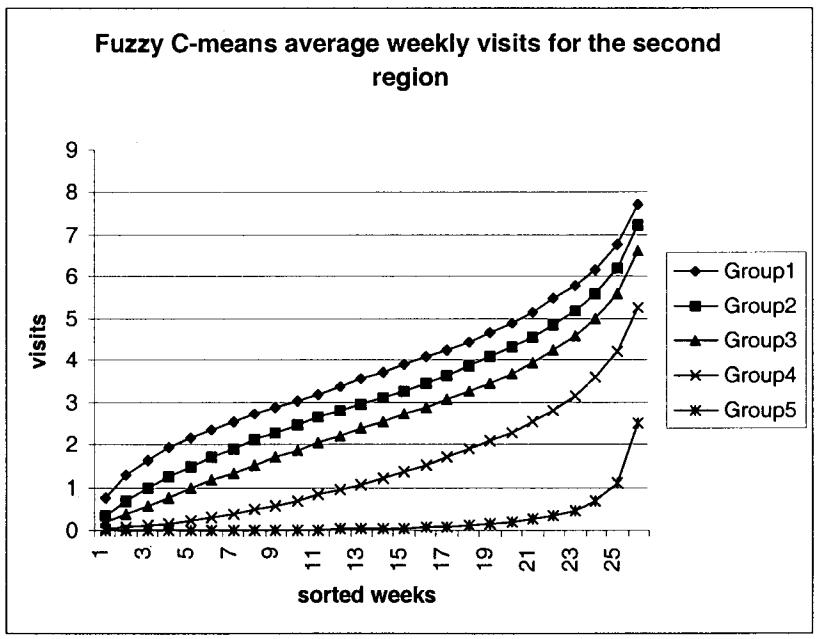


Figure 4.9: Fuzzy C-means average weekly visits for the second region

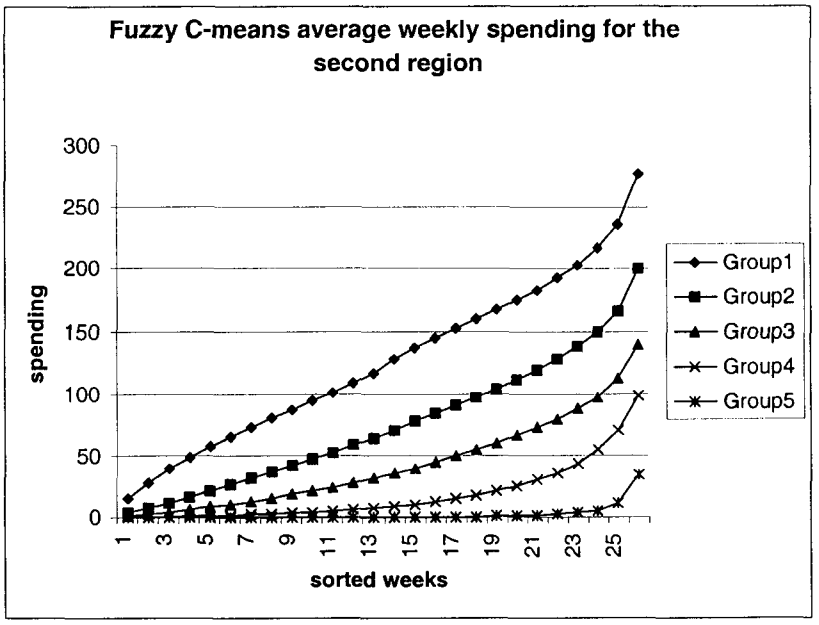


Figure 4.10: Fuzzy C-means average weekly spending for the second region

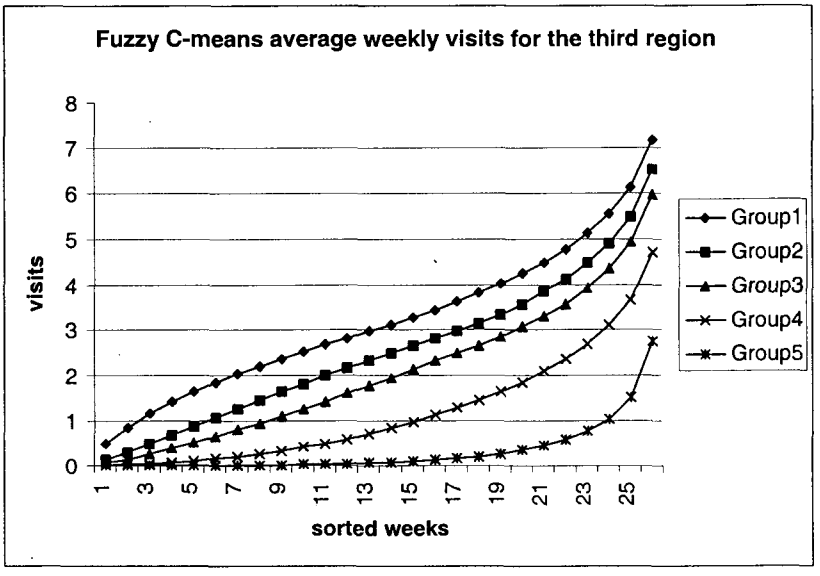


Figure 4.11: Fuzzy C-means average weekly visits for the third region

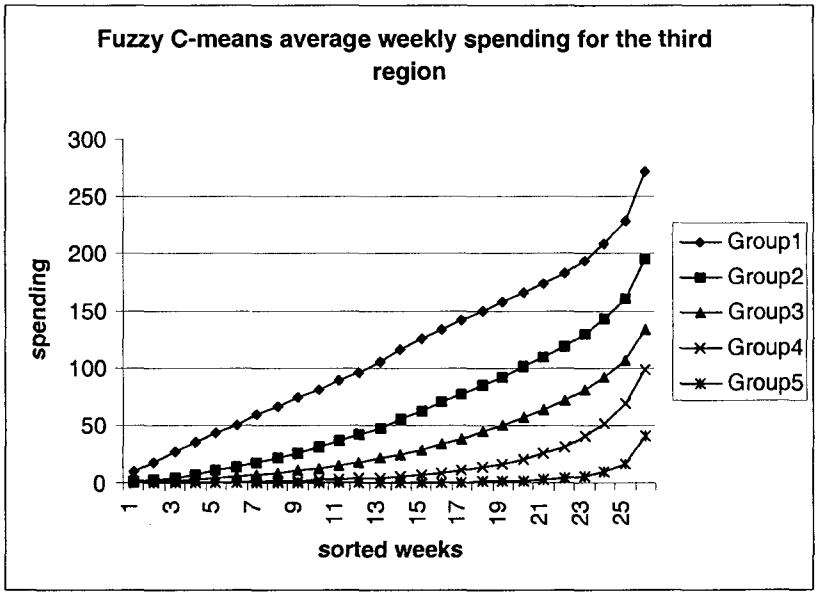


Figure 4.12: Fuzzy C-means average weekly spending for the third region

4.2.2 Cluster Cardinalities

The cardinalities for the three regions are shown in Tables 4.1 to 4.4. In each case, the most loyal group was the smallest, while the least loyal group was the largest. For example, for the first region, the three methods assign over 1,000 customers to group 1 and assign more than 10,000 customers to group 5. In Table 4.1, it can be seen that the K-means method assigns 1,246 customers to the loyal big spenders group (group 1), 2,427 customers to group 2, 2,309 customers to group 3, 6,218 customers to group 4 and 10,040 customers to the least loyal group (group 5). The fuzzy C-means method assigns 1,041 customers to group 1, 1,559 customers to group 2, 2,537 customers to group 3, 4,785 customers to group 4 and 10,077 customers to group 5. For the third method, the rough K-means method, the lower bound of group 1 consists of 1,268 customers, while the upper bound contains 1,422 customers. The lower bound of group 2 consists of 2,127 customers, while the upper bound contains 2,781 customers. The lower bound of group 3 consists of 2,238 customers, while the upper bound contains 4,041 customers. The lower bound of group 4 consists of 2,102 customers, while the upper bound contains 3,692 customers. The lower bound of group 5 consists of 11,782 customers, while the upper bound contains 12,716 customers. As loyalty decreases, the number of customers increases. The cardinality percentages for the three techniques are shown in Table 4.2.

Cluster Name	Conventional K-Means Clusters	Fuzzy C-Means Memberships >0.4	Rough K-Means Lower Bound	Rough K-Means Upper Bound	Rough K-means Boundary
Group 1	959	1704	2541	2854	313
Group 2	2709	2559	1739	2780	1041
Group 3	4519	3624	2715	4131	1416
Group 4	6832	5640	4700	7634	2934
Group 5	12618	12068	12198	14124	1926

Table 4.3: Vector representation of clusters for second region

Cluster Name	Conventional K-Means Clusters	Fuzzy C-Means Memberships >0.4	Rough K-Means Lower Bound	Rough K-Means Upper Bound	Rough K-means Boundary
Group 1	2776	3268	3798	4330	532
Group 2	6375	5353	6815	9031	2216
Group 3	5051	7727	6577	11992	5415
Group 4	14210	13132	5517	10409	4892
Group 5	28552	22025	26163	30252	4089

Table 4.4: Vector representation of clusters for third region

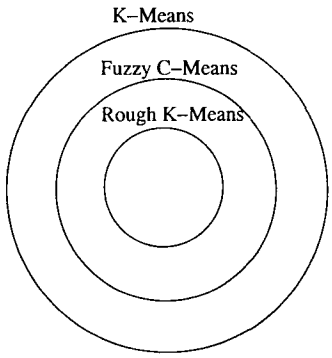


Figure 4.13: Cardinality comparison among the three methods

As mentioned previously, the K-means method is a crisp clustering method. The sum of the group cardinalities for the K-means method is equal to the total data set. However, the fuzzy and rough methods yield subsets of the total data set. In this study, the rough K-means method assigns the lowest number of customers to the five groups, while the fuzzy C-means method yields the second largest data set. Figure 4.13 shows the relationships among the cardinalities for the three methods.

Cluster	Intersec (FCMs and KMs)	Union (FCMs and KMs)
Group1	969	1318
Group2	1173	2813
Group3	1259	3587
Group4	4698	6305
Group5	9959	10158

Table 4.5: Intersection and union for fuzzy C-means (FCMs) and K-means (KMs) for first region

4.2.3 Overlap Analysis among the Three Techniques

As in the previous experiment using web data, the intersections among the three clustering methods are also analyzed. Tables 4.5 to 4.10 provide precise figures for the intersection and union among sets. Table 4.11 shows the intersection ratio for the fuzzy C-means and K-means methods for the first region. It can be easily seen that the highest values are found on the diagonal of the table. The intersection ratio for the two methods for group 1 is 0.74, and the ratios for groups 4 and 5 are higher than 0.5. Groups 2 and 3 seem more similar than the other groups, and the intersection ratios for these groups are lower. This is also the case for the second region, (see Table 4.13). The third region is somewhat different in this respect. Here it is groups 3 and 4 which have lower ratios, (see Table 4.15). Tables 4.12, 4.14 and 4.16 show the intersection ratios for the fuzzy C-means and rough K-means methods for the three regions. In a comparison of the fuzzy C-means and rough K-means methods, groups 3 and 4 have lower ratios for the first and third regions, (see Tables 4.12 and 4.16).

Cluster	Intersec (FCMs and RKMs)	Union (FCMs and RKMs)
Group1	1041	1268
Group2	1463	2223
Group3	1139	3636
Group4	973	5914
Group5	10073	11786

Table 4.6: Intersection and union for fuzzy C-means (FCMs) and modified K-means (RKMs) for first region

Cluster	Intersec (FCMs and KMs)	Union (FCMs and KMs)
Group1	899	1764
Group2	1511	3757
Group3	2572	5571
Group4	5176	7296
Group5	12068	12618

Table 4.7: Intersection and union for fuzzy C-means (FCMs) and K-means (KMs) for second region

Cluster	Intersec (FCMs and RKMs)	Union (FCMs and RKMs)
Group1	1704	2541
Group2	472	3826
Group3	1300	5039
Group4	3663	6677
Group5	12041	12225

Table 4.8: Intersection and union for fuzzy C-means (FCMs) and modified K-means (RKMs) for second region

Cluster	Intersec (FCMs and KMs)	Union (FCMs and KMs)
Group1	2599	3445
Group2	4193	7535
Group3	1834	10944
Group4	7400	19942
Group5	22025	28552

Table 4.9: Intersection and union for fuzzy C-means (FCMs) and K-means (KMs) for third region

Cluster	Intersec (FCMs and RKMs)	Union (FCMs and RKMs)
Group1	3268	3798
Group2	4957	7211
Group3	3253	11051
Group4	2282	16367
Group5	22025	26163

Table 4.10: Intersection and union for fuzzy C-means (FCMs) and modified K-means (RKMs) for third region

Cluster	FCMs Group1	FCMs Group2	FCMs Group3	FCMs Group4	FCMs Group5
KMs Group1	0.74	0.03	0.02	0.02	0.01
KMs Group2	0.03	0.42	0.14	0.03	0.02
KMs Group3	0.02	0.11	0.35	0.06	0.04
KMs Group4	0.01	0.02	0.14	0.75	0.08
KMs Group5	0.01	0.02	0.03	0.07	0.98

Table 4.11: Intersection ratios for fuzzy C-means (FCMs) and K-means (KMs) for first region

Cluster	FCMs Group1	FCMs Group2	FCMs Group3	FCMs Group4	FCMs Group5
RKMs Group1	0.82	0.01	0.00	0.00	0.00
RKMs Group2	0.00	0.66	0.04	0.00	0.00
RKMs Group3	0.00	0.00	0.31	0.01	0.00
RKMs Group4	0.00	0.00	0.14	0.17	0.00
RKMs Group5	0.00	0.00	0.00	0.12	0.85

Table 4.12: Intersection ratios for fuzzy C-means (FCMs) and modified K-means (RKMs) for first region

Cluster	FCMs Group1	FCMs Group2	FCMs Group3	FCMs Group4	FCMs Group5
KMs Group1	0.51	0.22	0.01	0.01	0.00
KMs Group2	0.00	0.40	0.18	0.02	0.01
KMs Group3	0.01	0.01	0.46	0.14	0.02
KMs Group4	0.00	0.02	0.04	0.71	0.07
KMs Group5	0.00	0.01	0.02	0.05	0.96

Table 4.13: Intersection ratios for fuzzy C-means (FCMs) and K-means (KMs) for second region

Cluster	FCMs Group1	FCMs Group2	FCMs Group3	FCMs Group4	FCMs Group5
RKMs Group1	0.67	0.12	0.00	0.00	0.00
RKMs Group2	0.00	0.12	0.11	0.00	0.00
RKMs Group3	0.00	0.25	0.26	0.00	0.00
RKMs Group4	0.00	0.00	0.09	0.55	0.00
RKMs Group5	0.00	0.00	0.00	0.01	0.99

Table 4.14: Intersection ratios for fuzzy C-means (FCMs) and modified K-means (RKMs) for second region

Cluster	FCMs Group1	FCMs Group2	FCMs Group3	FCMs Group4	FCMs Group5
KMs Group1	0.75	0.01	0.01	0.01	0.01
KMs Group2	0.08	0.56	0.04	0.03	0.02
KMs Group3	0.01	0.15	0.17	0.05	0.04
KMs Group4	0.01	0.03	0.38	0.37	0.06
KMs Group5	0.01	0.02	0.03	0.22	0.77

Table 4.15: Intersection ratios for fuzzy C-means (FCMs) and K-means (KMs) for third region

Cluster	FCMs Group1	FCMs Group2	FCMs Group3	FCMs Group4	FCMs Group5
RKMs Group1	0.86	0.00	0.00	0.00	0.00
RKMs Group2	0.00	0.69	0.05	0.00	0.00
RKMs Group3	0.00	0.00	0.30	0.12	0.00
RKMs Group4	0.00	0.00	0.14	0.14	0.00
RKMs Group5	0.00	0.00	0.00	0.11	0.84

Table 4.16: Intersection ratios for fuzzy C-means (FCMs) and modified K-means (RKMs) for third region

Cluster	Membership for Group1	Membership for Group2	Membership for Group3	Membership for Group4	Membership for Group5
Group1	0.52	0.24	0.11	0.07	0.06
Group2	0.13	0.41	0.28	0.12	0.07
Group3	0.06	0.23	0.43	0.19	0.09
Group4	0.02	0.05	0.23	0.51	0.20
Group5	0.00	0.01	0.03	0.13	0.84

Table 4.17: Group memberships for first region, using K-means (KMs)

4.2.4 Membership Analysis

The group memberships for the three regions obtained using the K-means technique are shown in Tables 4.17, 4.18, and 4.19, respectively. It can be seen that for the second region (see Table 4.18), group 1, obtained using the K-means method, has a membership of 0.60 for the first group, 0.19 for the second group, 0.10 for the third group, 0.07 for the fourth group and 0.07 for the fifth group. The fact that the membership values decrease makes sense, because groups 1 and 2 are more similar than groups 1 and 5. Group 5 has a membership of 0.86 for the fifth group and 0.11 for the fourth group, and has even lower values for the other three groups. Moreover, it can be easily seen that the higher values appear on the diagonal of the table. Similar characteristics can also be found in the other two tables. Unlike the third region, the first and second regions are both small towns. In the third region, there is a population of more than 125,000, and approximately 175,000 people live within a radius of 50 kilometers [40]. Including two small towns, the second region had a population of about 80,601 [41] in 1997. There are no other competitive grocery stores in the second region. Since the first two regions have similar geographical characteristics, the memberships of the five groups in the first two regions are similar (Table 4.17, 4.18).

Cluster	Membership for Group1	Membership for Group2	Membership for Group3	Membership for Group4	Membership for Group5
Group1	0.60	0.19	0.10	0.07	0.07
Group2	0.32	0.44	0.14	0.07	0.04
Group3	0.06	0.29	0.45	0.14	0.06
Group4	0.02	0.06	0.24	0.54	0.14
Group5	0.00	0.01	0.02	0.11	0.86

Table 4.18: Group memberships for second region, using K-means (KMs)

Both regions exhibit a low membership for groups 2 and 3 and a high membership for the last group. A possible explanation is that during the study period (May to October), the first two regions experienced considerable tourism. For example, the second region has a large, important harbor, which is considered to be the best harbor on the northern shore of the state. Therefore, during the summer season many tourists visit this region. During their stay, tourists went to the stores and did their grocery shopping. However, since their stay was temporary, the stores lost these customers when they left the region. Further studies could be done, for the two regions, of transactions during the other six months of the year. In the second region, in winter, the water is frozen from December to April. Thus, the membership of group 5 can be expected to be lower than indicated in Table 4.18, since there are fewer tourists during the winter period. The third region is not a popular tourist destination, and there are other competitive grocery stores in the neighborhood. The membership of group 5 is therefore less than in the other two regions due to the lower impact from tourism.

Cluster	Membership for Group1	Membership for Group2	Membership for Group3	Membership for Group4	Membership for Group5
Group1	0.59	0.19	0.10	0.07	0.05
Group2	0.19	0.48	0.20	0.08	0.05
Group3	0.08	0.30	0.35	0.19	0.09
Group4	0.02	0.09	0.38	0.40	0.11
Group5	0.01	0.02	0.05	0.24	0.69

Table 4.19: Group memberships for third region, using K-means (KMs)

4.2.5 Monthly Analysis

In this section, the monthly shopping behavior of customers is analyzed. It was found in an initial study that the use of the five groups discussed in the previous section did not result in clear distinctions among the groups. It was therefore decided to group the customers into the following three clusters:

- Group 1: Loyal big spenders.
- Group 2: Semi-loyal spenders.
- Group 3: Least loyal spenders.

First the results of the K-means method are analyzed. Figures 4.14 to 4.25 show number of visits and the average spending for the first region, for the six months in sequence. It can be seen that the differences among the groups with regard to number of visits and spending patterns are quite clear. For example, in Figure 4.14 and 4.15, the average number of visits for group 1 is 4, and the spending is over \$180. The average number of visits for the second group is less than 3.5, and the spending is lower than \$100. The last group is the least loyal group. Over a period of four weeks, customers in this group visit only around once and spent less than \$20. Similar characteristics can be observed for the remaining months.

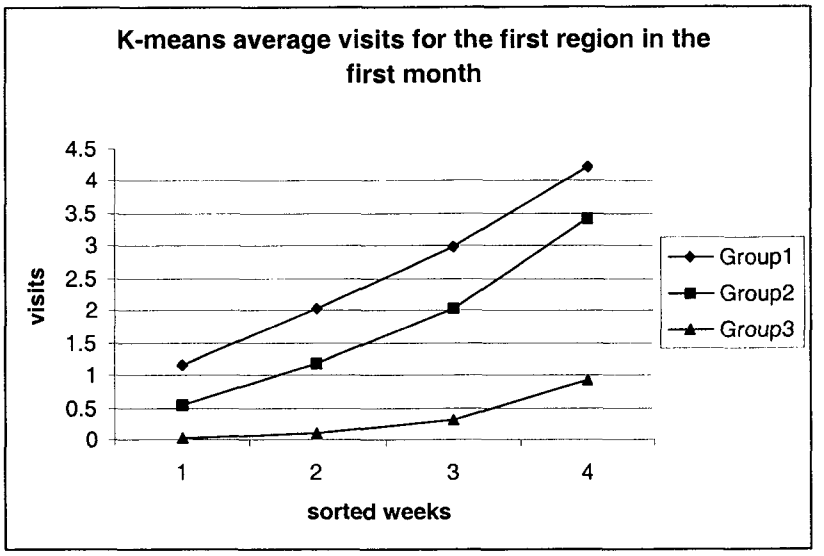


Figure 4.14: K-means average visits for the first region in the first month

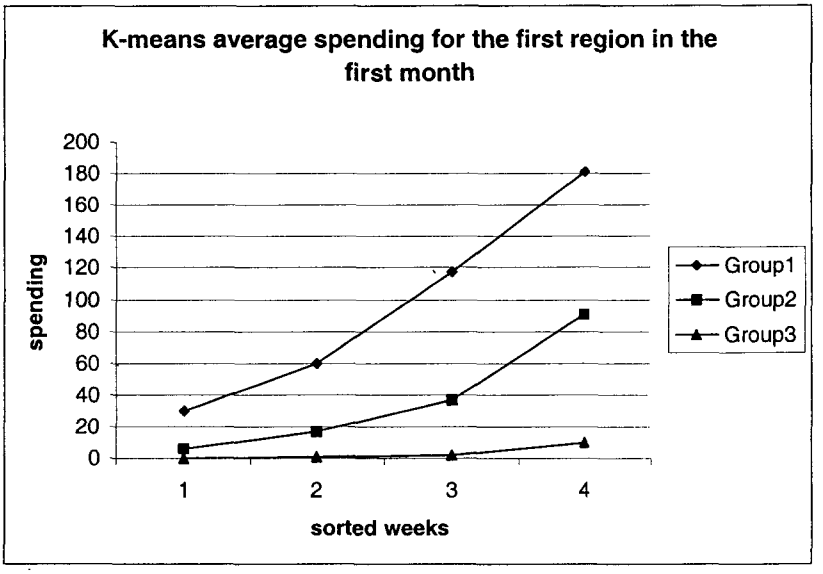


Figure 4.15: K-means average spending for the first region in the first month

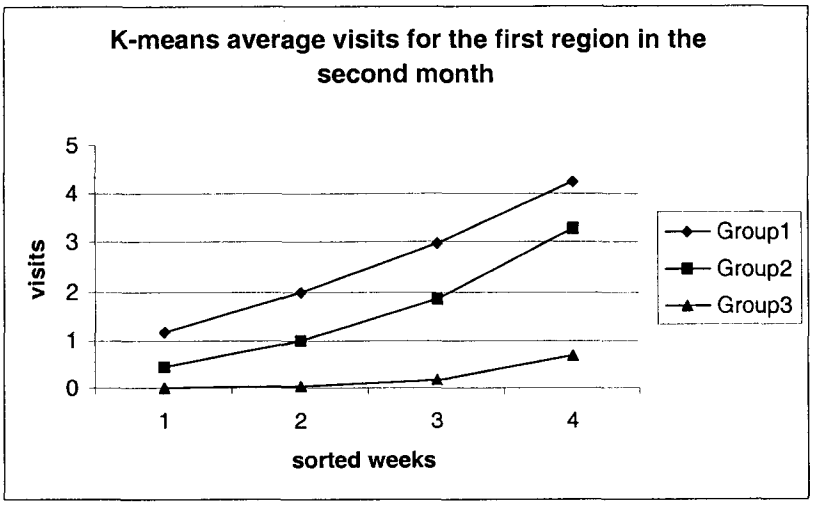


Figure 4.16: K-means average visits for the first region in the second month

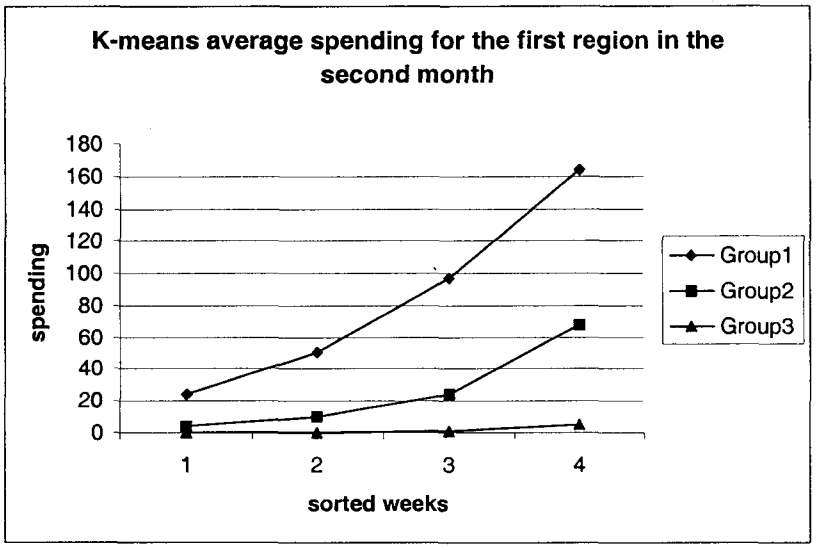


Figure 4.17: K-means average spending for the first region in the second month

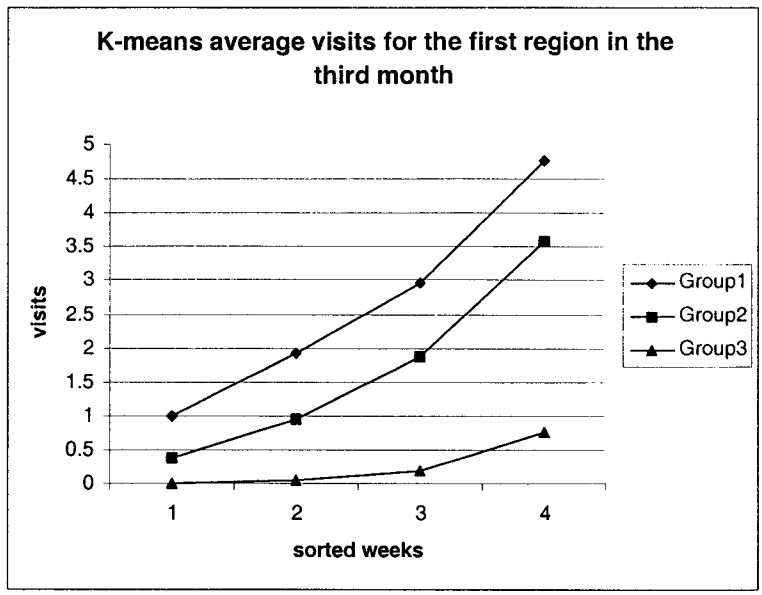


Figure 4.18: K-means average visits for the first region in the third month

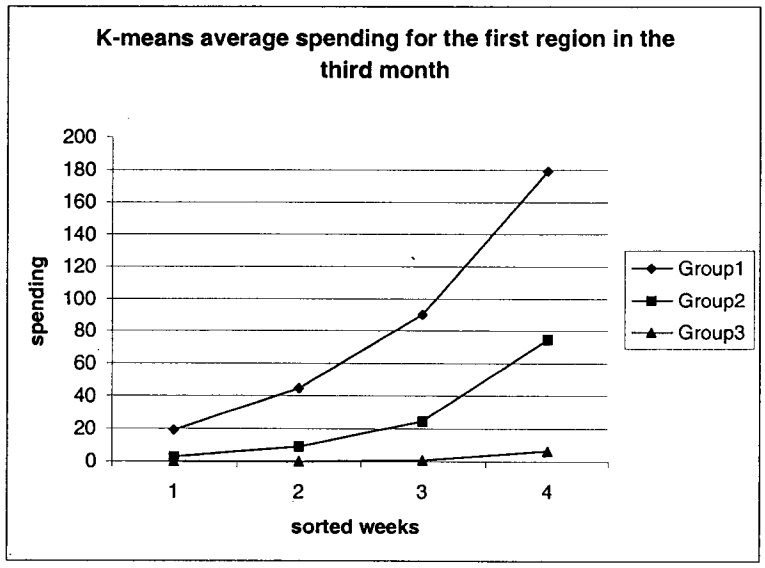


Figure 4.19: K-means average spending for the first region in the third month

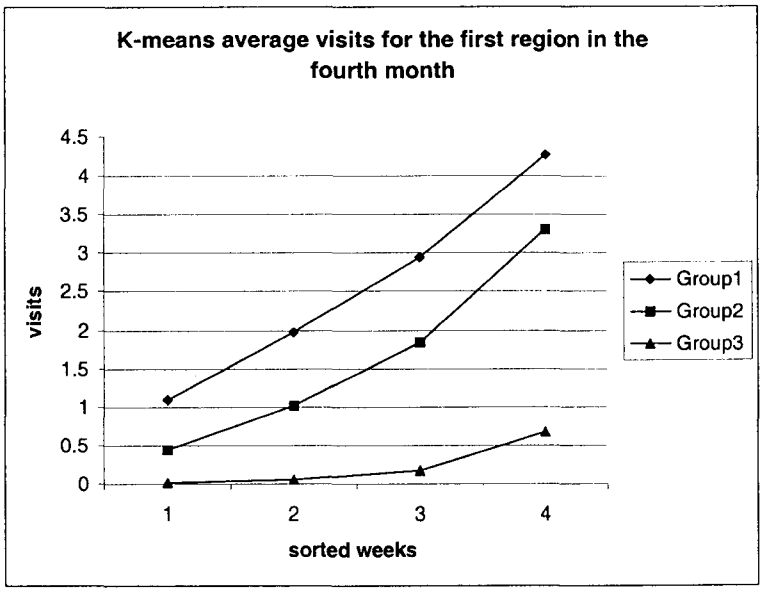


Figure 4.20: K-means average visits for the first region in the fourth month

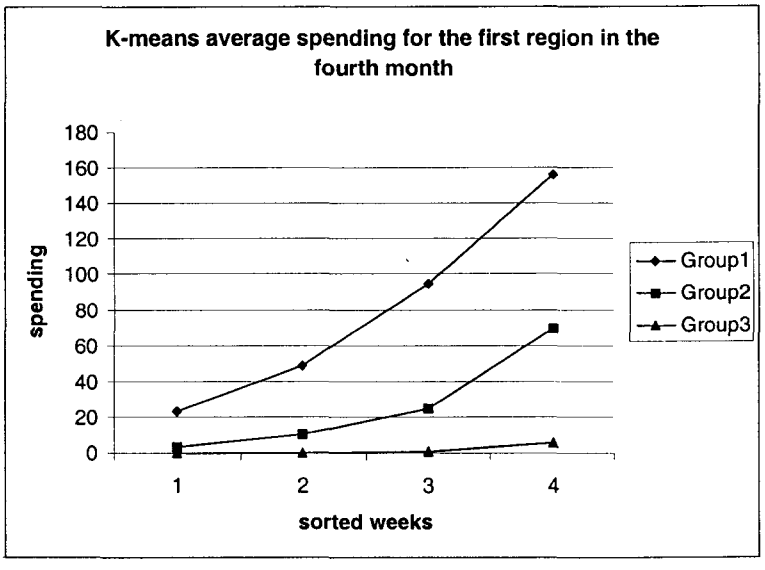


Figure 4.21: K-means average spending for the first region in the fourth month

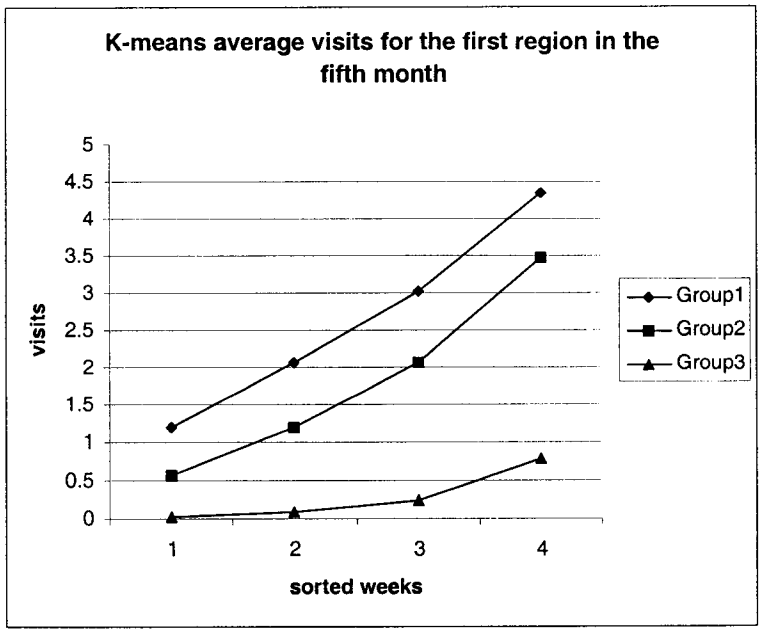


Figure 4.22: K-means average visits for the first region in the fifth month

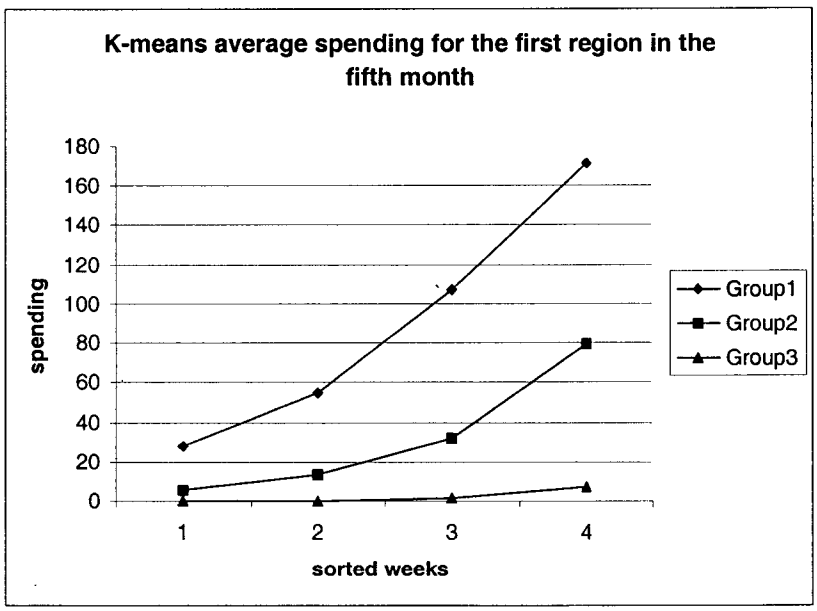


Figure 4.23: K-means average spending for the first region in the fifth month

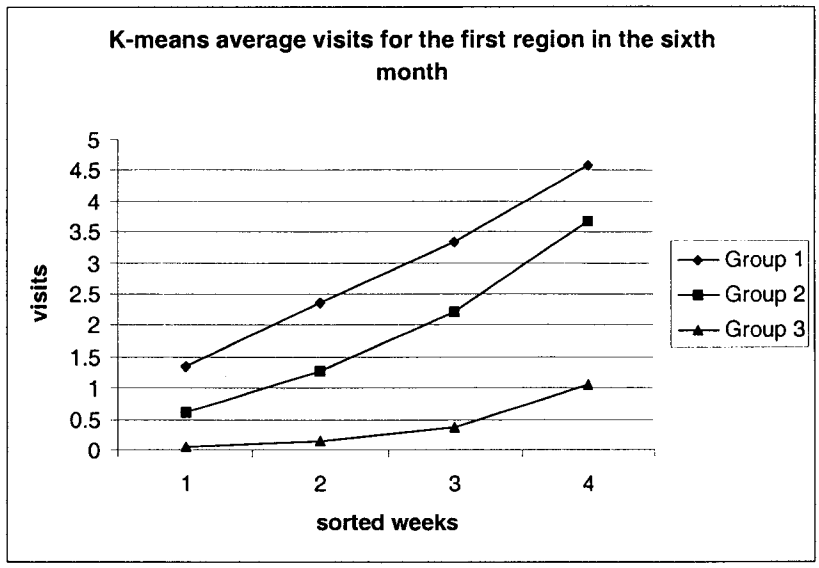


Figure 4.24: K-means average visits for the first region in the sixth month

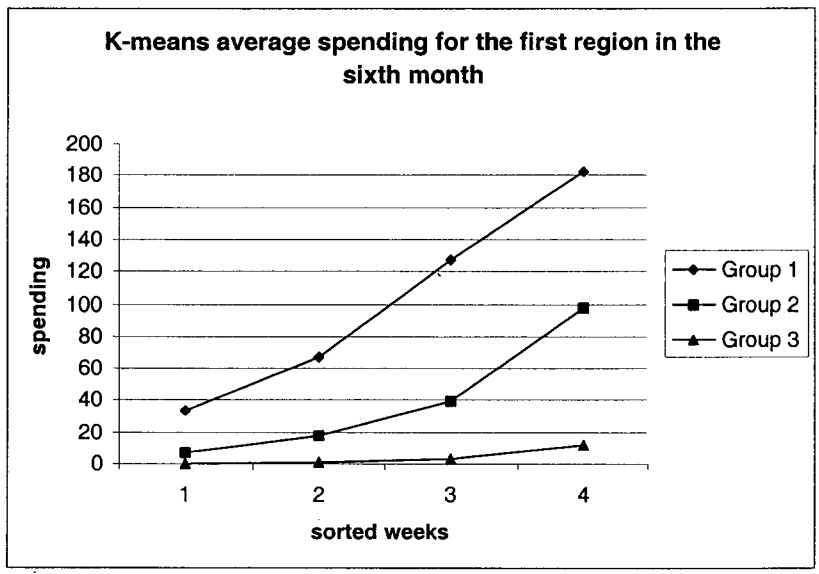


Figure 4.25: K-means average spending for the first region in the sixth month

Cluster	Fuzzy C-Means	K-Means
Group1	2006	1181
Group2	4851	4872
Group3	14372	16184

Table 4.20: Cardinality comparison for first month

Cluster	Fuzzy C-Means	K-Means
Group1	0	2494
Group2	0	6338
Group3	3803	13408

Table 4.21: Cardinality comparison for second month

Tables 4.20 to 4.25 show the monthly cardinalities obtained using the K-means method. Figures 4.46.a, 4.47.a and 4.48.a graphically illustrate the cardinality changes over a period of six months for each group. For example, the most loyal group, group 1, changes markedly during the six months. In the first month (May), there are about 1,181 customers. This number increases for the next two months. In the third month (July), there are a total of 2,660 customers. However, this group starts to lose customers over the following three months (August to October). The second most loyal group, group 2, also begins to lose customers after the second month (June). This make sense, since in the summer season, especially in the case of loyal customers, people do not spend so much time on price comparisons, since they want to spend more time on outdoor activities. Having noted this phenomenon, managers may also study what policies have changed over the six-month period.

Cluster	Fuzzy C-Means	K-Means
Group1	2144	2660
Group2	5141	6059
Group3	13900	13521

Table 4.22: Cardinality comparison for third month

Cluster	Fuzzy C-Means	K-Means
Group1	2101	2554
Group2	5019	5706
Group3	14073	13980

Table 4.23: Cardinality comparison for fourth month

Cluster	Fuzzy C-Means	K-Means
Group1	2050	1956
Group2	5040	5394
Group3	14113	14890

Table 4.24: Cardinality comparison for fifth month

Cluster	Fuzzy C-Means	K-Means
Group1	2110	1361
Group2	5338	5148
Group3	13691	15731

Table 4.25: Cardinality comparison for sixth month

The fuzzy C-means method also yields meaningful results in the monthly analysis. Figures 4.26 to 4.35 show the average visits and spending for the first region for five months, excluding the second month (June). Unexpectedly, in June, the average memberships for the three groups were found to be 0.50, 0.50 and 0.01, respectively. In this case, memberships of more than 95 percent for the three groups correspond to the values 0.50, 0.50 and 0.74. Since the memberships for the first two groups are so similar, the fuzzy C-means method does not assign the customers to groups for this month. More studies should be performed regarding this situation. Figures 4.36 to 4.45 show the vector centers for each month. Tables 4.20 to 4.25 show the cardinalities for the six month from the fuzzy C-means and K-means methods. Figures 4.46.b, 4.47.b and 4.48.b graphically illustrate the cardinality changes during the 6-month period. Due to the fact that the second month cardinalities are missing, the cardinalities are compared for only five months. A comparison with the six figures (Figure 4.46 to 4.48) shows clearly that both methods detect the increase in customers for group 1 in the third month (July). The difference is that for the sixth month, the fuzzy C-means method indicates an increase in customers for group 1, while the K-means method does not detect this migration. Similarly, for the other two groups, the two methods indicate differing customer behaviors in the sixth month. In future, a more detailed analysis should be carried out with regard to this discrepancy.

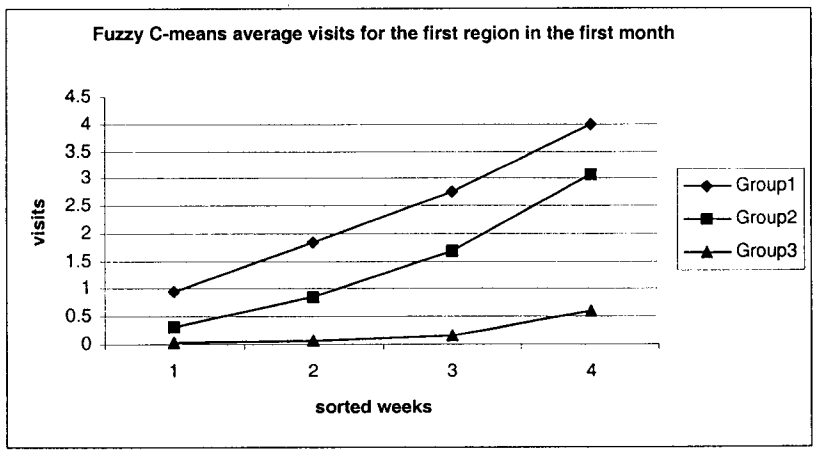


Figure 4.26: Fuzzy C-means average visits for the first region in the first month

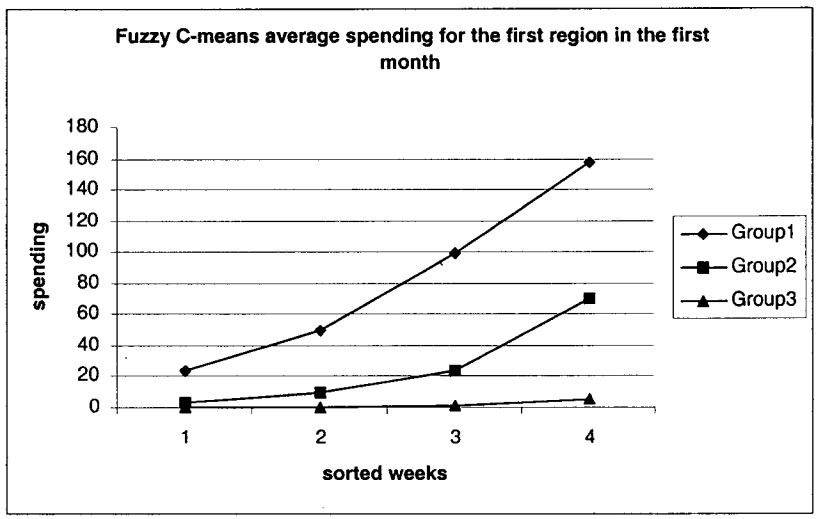


Figure 4.27: Fuzzy C-means average spending for the first region in the first month

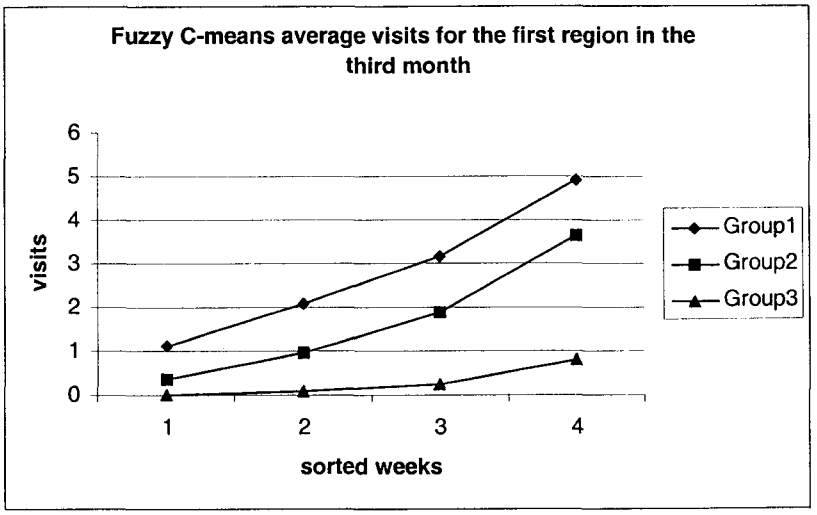


Figure 4.28: Fuzzy C-means average visits for the first region in the third month

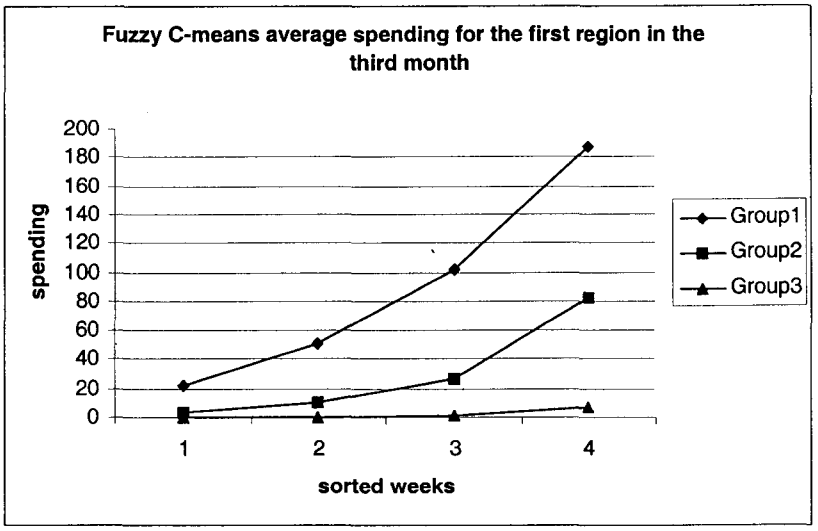


Figure 4.29: Fuzzy C-means average spending for the first region in the third month

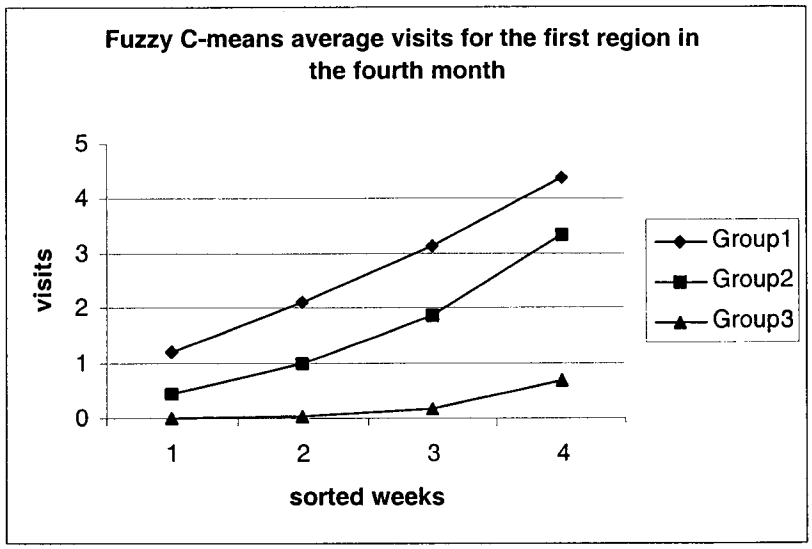


Figure 4.30: Fuzzy C-means average visits for the first region in the fourth month

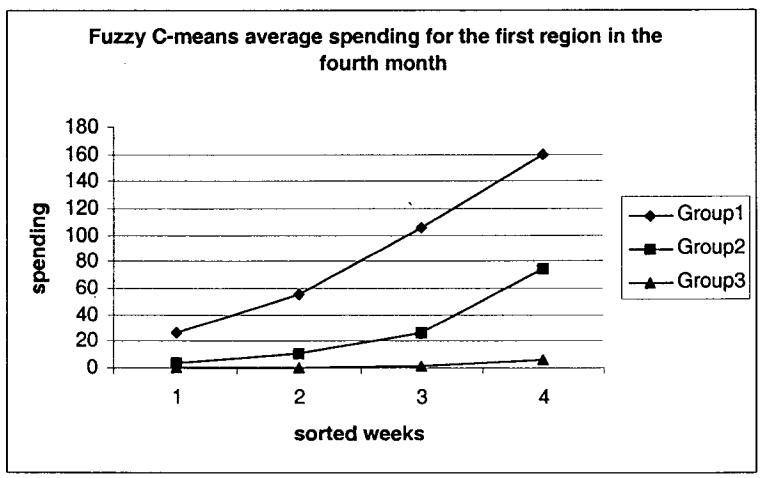


Figure 4.31: Fuzzy C-means average spending for the first region in the fourth month

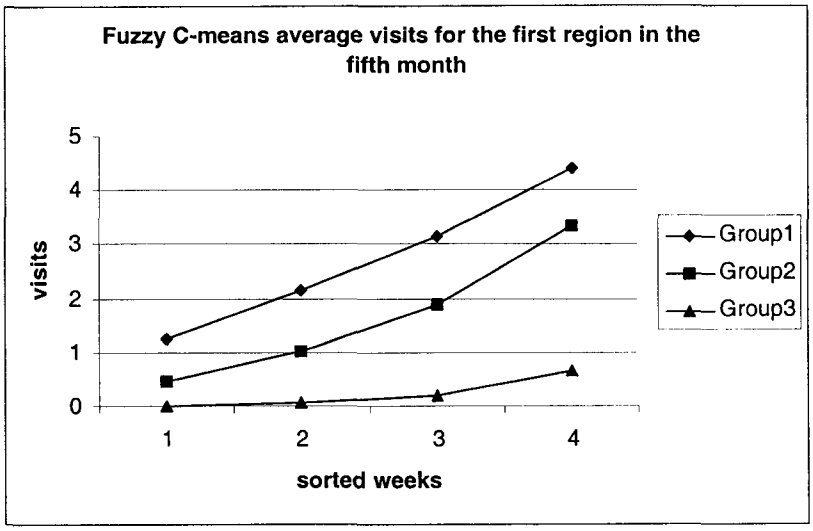


Figure 4.32: Fuzzy C-means average visits for the first region in the fifth month

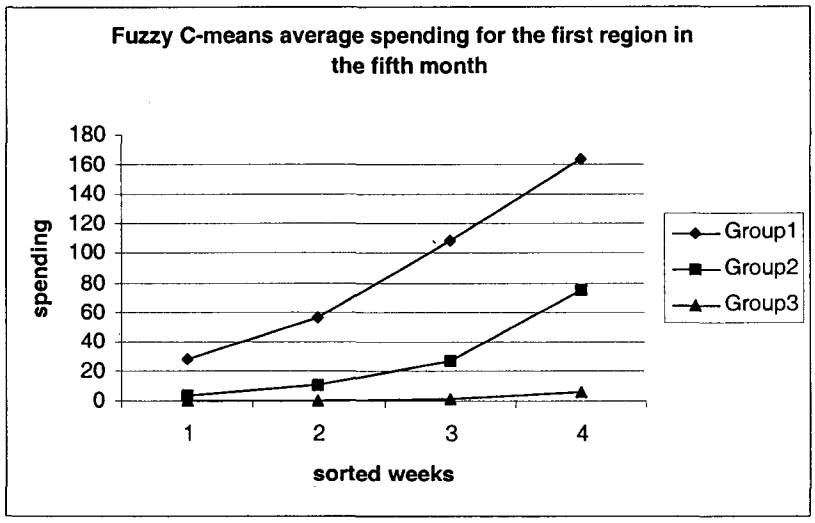


Figure 4.33: Fuzzy C-means average spending for the first region in the fifth month

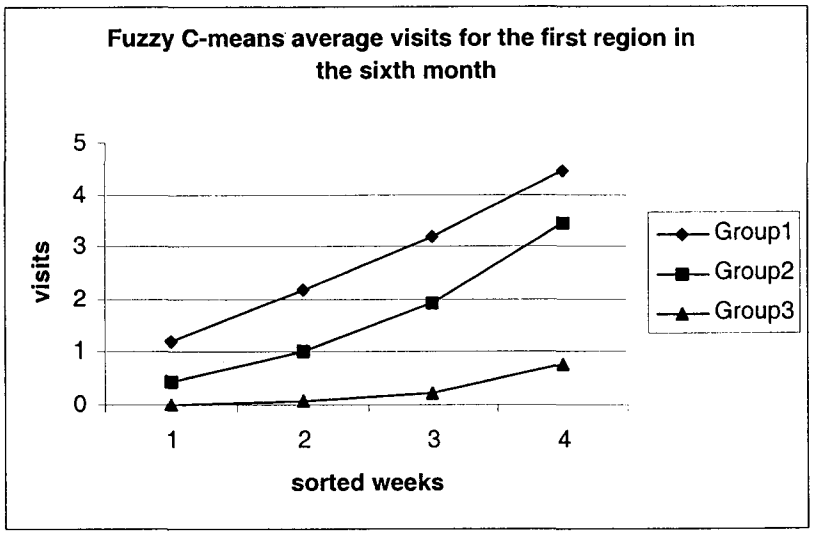


Figure 4.34: Fuzzy C-means average visits for the first region in the sixth month

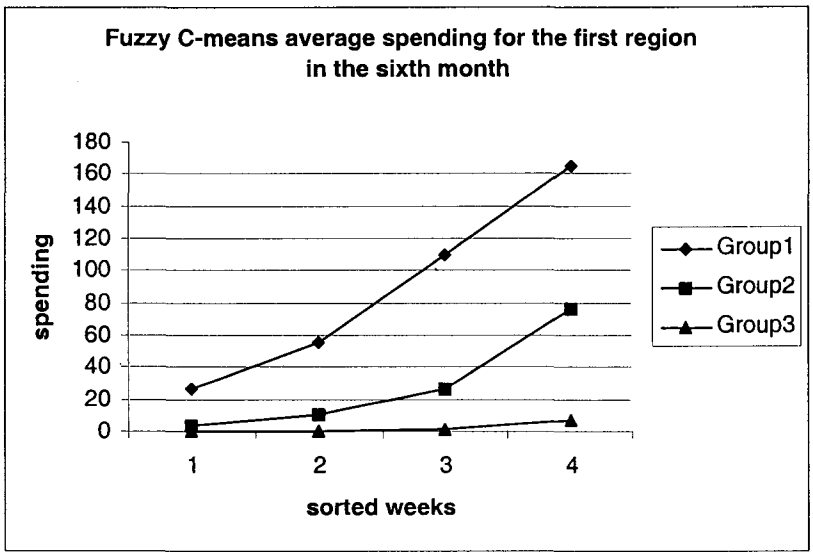


Figure 4.35: Fuzzy C-means average spending for the first region in the sixth month

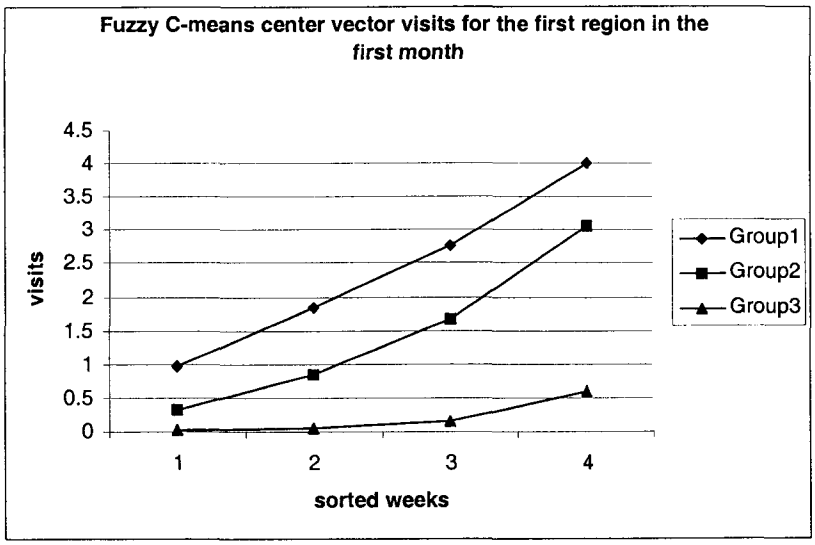


Figure 4.36: Fuzzy C-means center vector visits for the first region in the first month

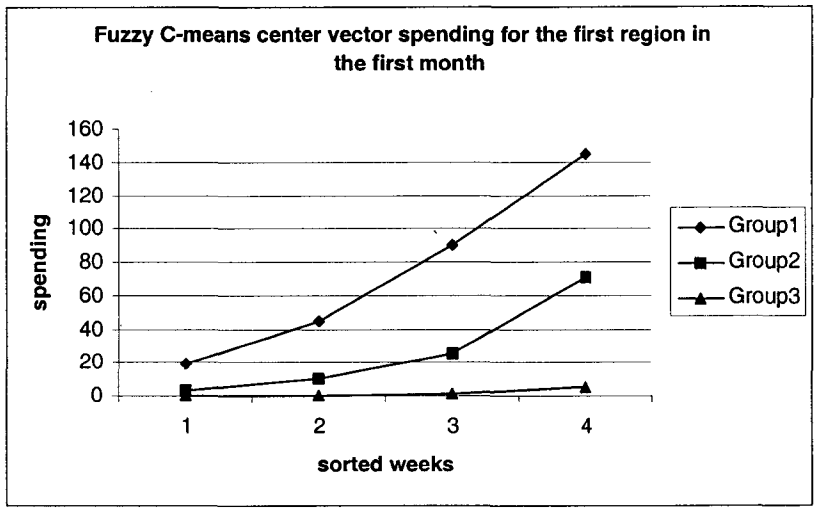


Figure 4.37: Fuzzy C-means center vector spending for the first region in the first month

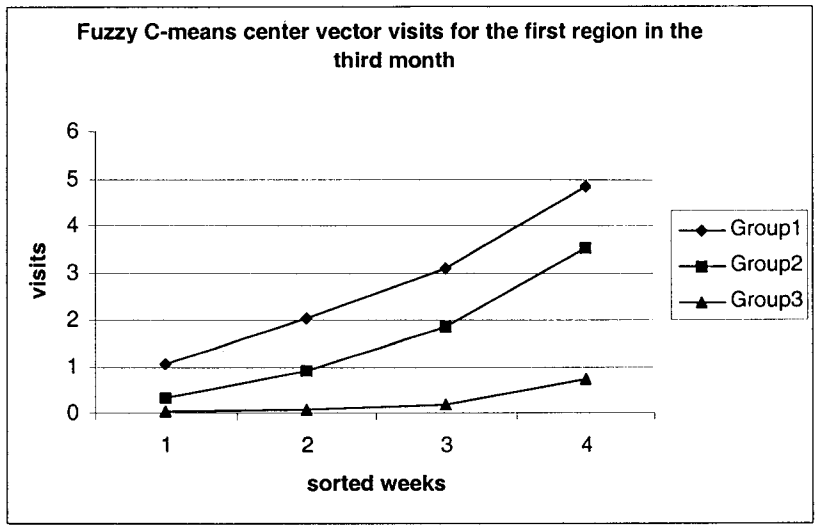


Figure 4.38: Fuzzy C-means center vector visits for the first region in the third month

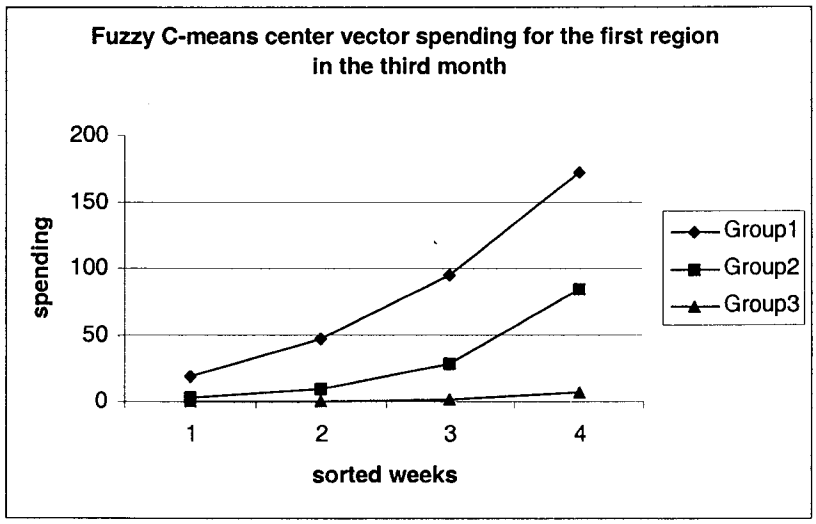


Figure 4.39: Fuzzy C-means center vector spending for the first region in the third month

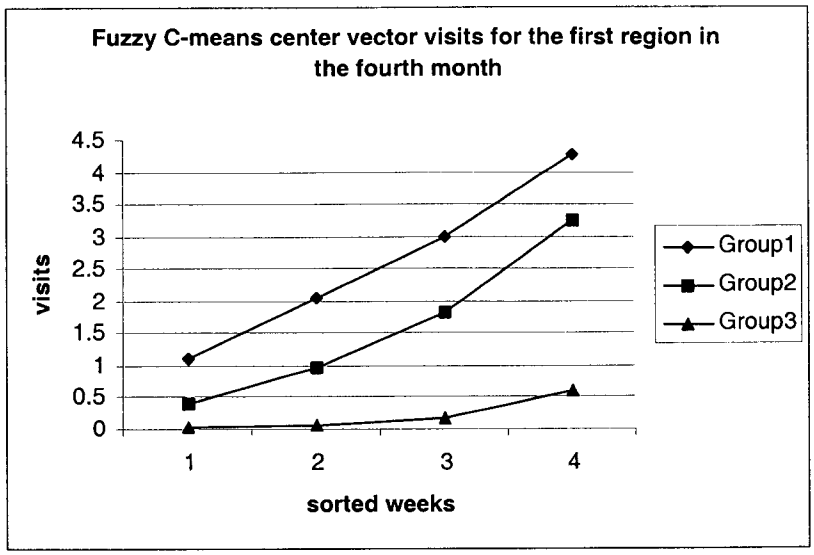


Figure 4.40: Fuzzy C-means center vector visits for the first region in the fourth month

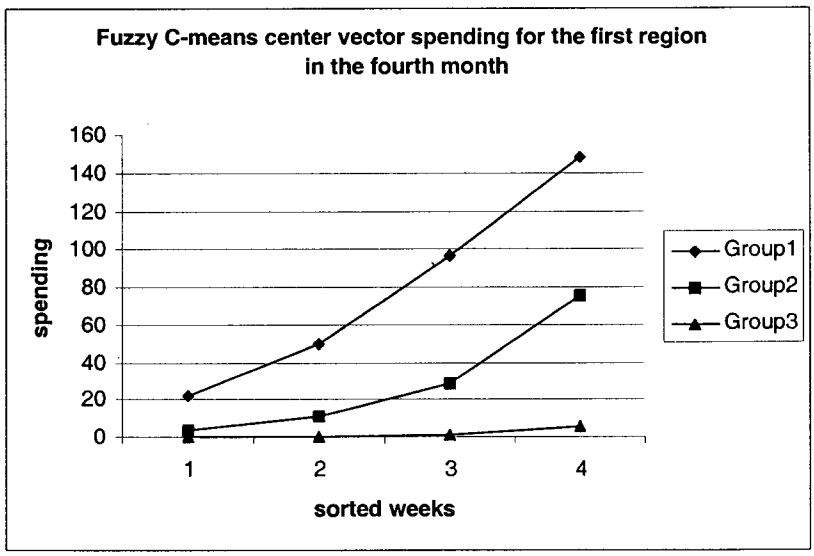


Figure 4.41: Fuzzy C-means center vector spending for the first region in the fourth month

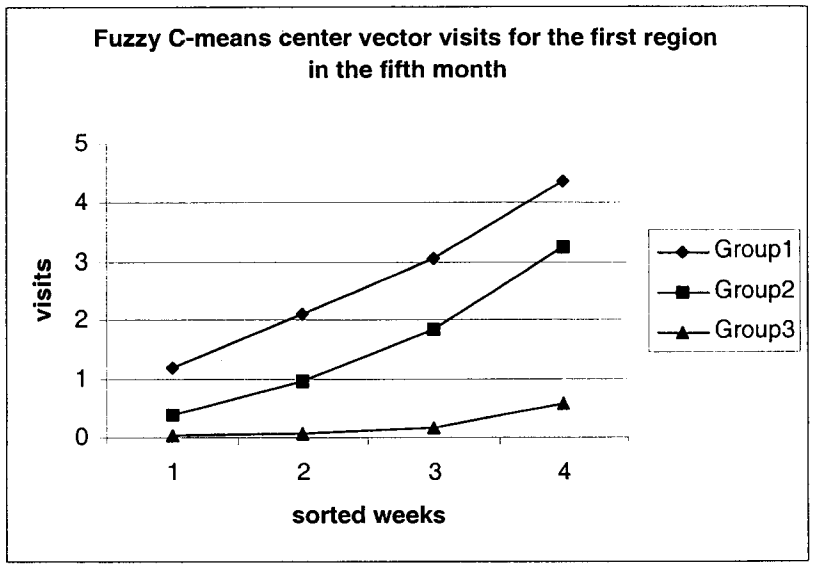


Figure 4.42: Fuzzy C-means center vector visits for the first region in the fifth month

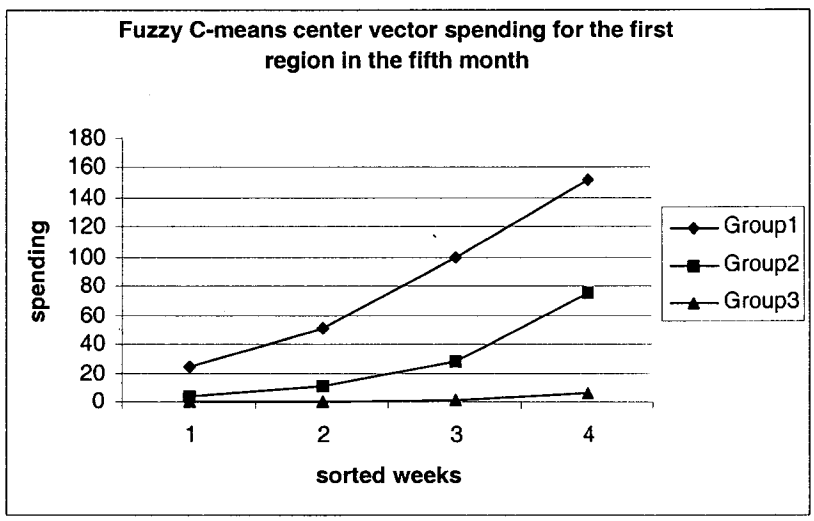


Figure 4.43: Fuzzy C-means center vector spending for the first region in the fifth month

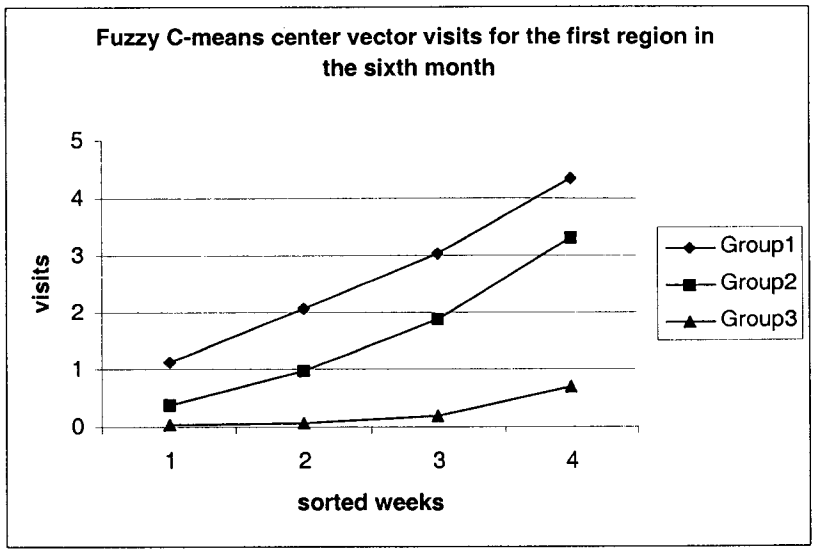


Figure 4.44: Fuzzy C-means center vector visits for the first region in the sixth month

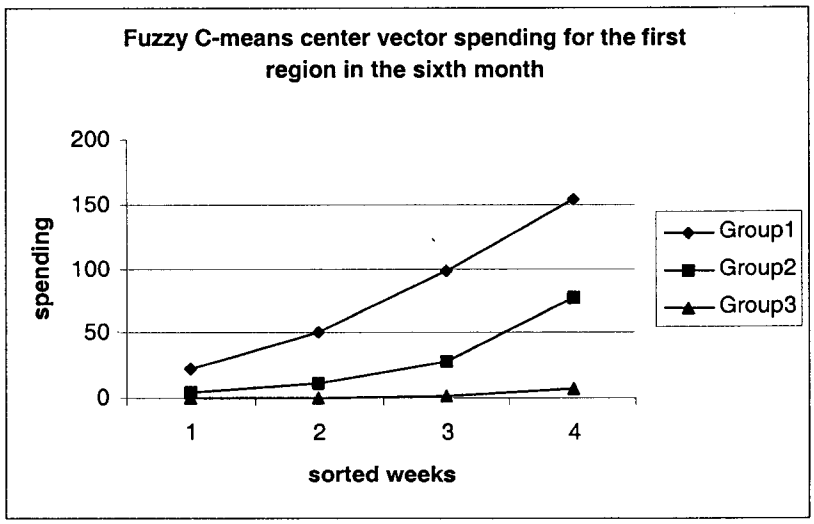
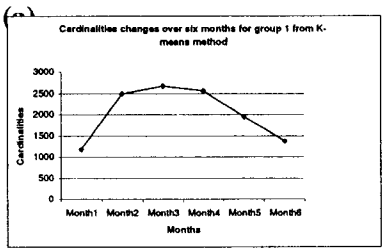
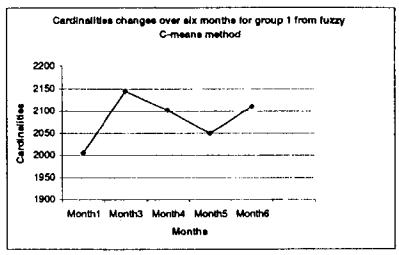


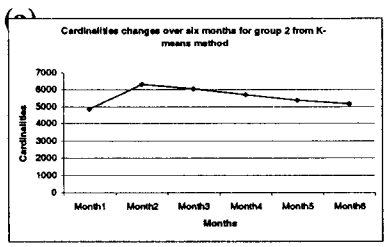
Figure 4.45: Fuzzy C-means center vector spending for the first region in the sixth month

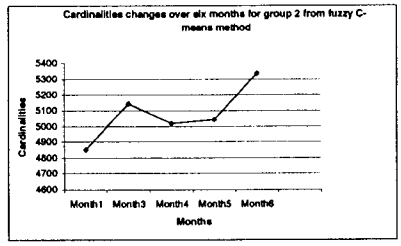




(K-means method) **(Fuzzy C-means method)**

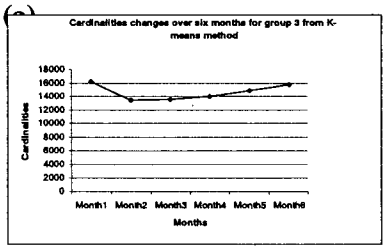
Figure 4.46: Cardinality changes over six months for group 1

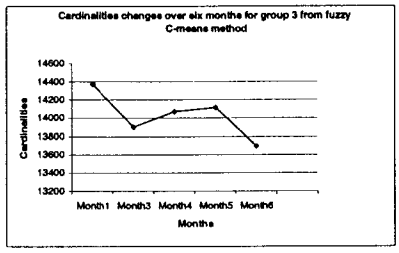




(K-means method) **(Fuzzy C-means method)**

Figure 4.47: Cardinality changes over six months for group 2





(K-means method) **(Fuzzy C-means method)**

Figure 4.48: Cardinality changes over six months for group 3

	KMs Group1	KMs Group2	KMs Group3
FCMs Group1	1180	826	0
FCMs Group2	0	3586	1265
FCMs Group3	0	0	14372

Table 4.26: Cardinality intersection between K-means and fuzzy C-means for first month

	KMs Group1	KMs Group2	KMs Group3
FCMs Group1	2007	6055	18190
FCMs Group2	6032	6140	19770
FCMs Group3	15553	19247	16184

Table 4.27: Cardinality union between K-means and fuzzy C-means for first month

Table 4.26 to 4.35 provide precise figures for the intersection and union among sets. As shown in table 4.26, 1180 customers are grouped into the loyal big spenders (Group 1) by fuzzy C-means and K-means methods for the first month (May), and there are 826 customers in the loyal big spenders (Group 1) by fuzzy C-means method, assigned to the semi-loyal spenders (Group 2) by K-means method. There is 0 customers in the intersection between fuzzy C-means method, the loyal big spenders (Group 1) and K-means method, the least loyal spenders (Group 3). This is reasonable because the similarity is much close between the loyal big spenders (Group 1) and the semi-loyal spenders (Group 2), compared with the loyal big spenders (Group 1) and the least loyal spenders (Group 3). Table 4.36 shows the intersection ratio for the fuzzy C-means and K-means methods for the first month. It can be easily seen that the highest value are found on the diagonal of the table. This can be found in the other months, (see Table 4.37 to 4.40).

Since fuzzy C-means method does not assign the customers to groups for the second month (June), we do not have the intersection and union sets for the second month. Further studies will be carried out with regard to this discrepancy.

	KMs Group1	KMs Group2	KMs Group3
FCMs Group1	2144	0	0
FCMs Group2	104	5037	104
FCMs Group3	0	379	13521

Table 4.28: Cardinality intersection between K-means and fuzzy C-means for third month

	KMs Group1	KMs Group2	KMs Group3
FCMs Group1	2660	8203	15665
FCMs Group2	7697	6163	18662
FCMs Group3	16560	19580	13900

Table 4.29: Cardinality union between K-means and fuzzy C-means for third month

	KMs Group1	KMs Group2	KMs Group3
FCMs Group1	2101	0	0
FCMs Group2	27	4992	0
FCMs Group3	0	96	13977

Table 4.30: Cardinality intersection between K-means and fuzzy C-means for fourth month

	KMs Group1	KMs Group2	KMs Group3
FCMs Group1	2554	7807	16081
FCMs Group2	7546	57336	18999
FCMs Group3	16627	19683	14076

Table 4.31: Cardinality union between K-means and fuzzy C-means for fourth month

	KMs Group1	KMs Group2	KMs Group3
FCMs Group1	1877	173	0
FCMs Group2	0	4810	230
FCMs Group3	0	0	14113

Table 4.32: Cardinality intersection between K-means and fuzzy C-means for fifth month

	KMs Group1	KMs Group2	KMs Group3
FCMs Group1	2129	7271	16940
FCMs Group2	6996	5624	19700
FCMs Group3	16069	19507	14890

Table 4.33: Cardinality union between K-means and fuzzy C-means for fifth month

	KMs Group1	KMs Group2	KMs Group3
FCMs Group1	1360	750	0
FCMs Group2	0	3908	1430
FCMs Group3	0	0	13691

Table 4.34: Cardinality intersection between K-means and fuzzy C-means for sixth month

	KMs Group1	KMs Group2	KMs Group3
FCMs Group1	2111	6508	17841
FCMs Group2	6699	6578	19639
FCMs Group3	15052	18839	15731

Table 4.35: Cardinality union between K-means and fuzzy C-means for sixth month

	KMs Group1	KMs Group2	KMs Group3
FCMs Group1	0.59	0.17	0.00
FCMs Group2	0.00	0.58	0.06
FCMs Group3	0.00	0.00	0.89

Table 4.36: Cardinality ratios between K-means and fuzzy C-means for first month

	KMs Group1	KMs Group2	KMs Group3
FCMs Group1	0.81	0.00	0.00
FCMs Group2	0.01	0.82	0.00
FCMs Group3	0.00	0.02	0.97

Table 4.37: Cardinality ratios between K-means and fuzzy C-means for third month

	KMs Group1	KMs Group2	KMs Group3
FCMs Group1	0.82	0.00	0.00
FCMs Group2	0.00	0.87	0.00
FCMs Group3	0.00	0.01	0.99

Table 4.38: Cardinality ratios between K-means and fuzzy C-means for fourth month

	KMs Group1	KMs Group2	KMs Group3
FCMs Group1	0.88	0.02	0.00
FCMs Group2	0.00	0.86	0.01
FCMs Group3	0.00	0.00	0.95

Table 4.39: Cardinality ratios between K-means and fuzzy C-means for fifth month

	KMs Group1	KMs Group2	KMs Group3
FCMs Group1	0.64	0.12	0.00
FCMs Group2	0.00	0.59	0.07
FCMs Group3	0.00	0.00	0.87

Table 4.40: Cardinality ratios between K-means and fuzzy C-means for sixth month

4.3 Summary and Conclusions

Customer classification in supermarket data mining may not necessarily be precise. The behavior of some of the customers may correspond to more than one category. Therefore, unsupervised clustering should aim to model overlapping clusters. This research uses the conventional K-means clustering method, a rough K-means algorithm, and a fuzzy C-means algorithm to develop interval clusters of supermarket customers. The rough K-means approach is based on the results of rough set theory. In order to develop interval clusters, the K-means algorithm has been modified based on the concept of lower and upper bounds. The fuzzy C-means approach calculates the memberships for each group. Based on the membership values, the customers are assigned to the designated groups.

Customer data from twelve supermarket stores concentrated in a rural setting were used to create interval clusters based on the three techniques. The data include information on spending, visits, shopping caregories, and other transactional data. Time-series values for spending and visits over a 26-week period are used to represent the customers. In order to eliminate artificial distinctions introduced by the timing of purchases, the time-series data were sorted. The value of groceries purchased provides an indication of spending potential, while the number of visits is a reasonable representation of customer loyalty. Therefore, it was hoped that the three methods would yield clusters such as: loyal big spenders, loyal moderate spenders, semi-loyal potentially big spenders, potentially moderate to big spenders with limited loyalty, and infrequent customers.

The experiment resulted in a meaningful clustering of customers using the three methods discussed in Chapter 2. A study of the variables used for clustering made it possible to label the five clusters as described above. The three techniques were compared by analyzing the overlap among the clusters. The higher values appearing along the diagonal of the comparison tables demonstrate the validity of the clustering analysis. The membership for each cluster was analyzed using the fuzzy C-means method as well as the K-means method. It was found that the fuzzy C-means method is more sensitive with regard to the number of visits. In the monthly analysis, the K-means and fuzzy C-means methods were employed in order to discover customer migration over a period of six months. The K-means method indicates a migration of customers from the less loyal group to the more loyal group in the first 2 to 3 months, and a migration back to the less loyal group in the remaining months. The fuzzy C-means method yields the same findings for the first 2 to 3 months however, in addition, it detects an increasing trend in the last month. A more detailed analysis should be made of this phenomenon in the future analysis.

In the current research, 26-dimensional vectors are used. The large number of dimensions makes the research difficult to visualize. Reducing the number of dimensions is a possible approach for studying customer behavior. For example, reducing the study period from six months to two months is a possible solution. Another possibility is to study the 50^{th}, 85^{th}, and 95^{th} percentiles for spending and visits rather than weekly spending and visits. This approach will be investigated in future work.

Chapter 5

Concluding remarks

5.1 Conclusions

This research investigates the spatio-temporal variations in cluster memberships of supermarket customers and the temporal variations in cluster characteristics of web users. The study analyzes the objects in data sets and assigns them to the designated groups using conventional K-means, fuzzy C-means and rough K-means methods. This research shows that the three methods successfully identify the clusters in the different data sets. The conventional K-means technique assigns each object to precisely one group. The fuzzy method calculates the degree of membership for each object in each cluster and the rough method assigns objects to lower and upper bounds, making it possible to provide a rough or unclear boundary for each cluster. Clusters cardinalities from the three methods are compared, and cluster characteristics are analyzed.

In the web user analysis, data from visits to three university course websites are used in the experiments. The first two courses are for first-year students and the third course is for second-year students. The students in the third course are core computing science students. The attitudes of students in the three courses are quite different. It was expected that the visitors to the websites would be classified as studious, crammers, or workers. Since some of the visitors to the websites may not precisely belong to one of the groups, the visitors were represented using fuzzy membership functions and rough sets. The experiments produced meaningful clustering of the web visitors using all three clustering techniques. Analysis of the variables used for clustering permitted clear identification of the three clusters as studious users, workers, and crammers. Many similarities and a few differences among the characteristics of the conventional clusters, fuzzy clusters and rough sets for the three websites were found.

In the supermarket customer shopping behavior analysis, the experiment is designed to analyze the customers of twelve supermarkets located in a rural setting. The analysis is used to create interval clusters based on the above three algorithms. Comparisons are made among the three methods. The target supermarkets are part of a national chain. The data were collected over a 26-week period beginning in May, 2001. The data collected include information on spending, visits, shopping categories, and other transactions. In order to test the validity of the results for different regions, data sets from three regions are used. The first region has only one store. The second region has five stores. The third region has six stores. The third region also differs from the other two regions in

terms of geographical characteristics. The first two regions are both small towns, which are tourist destinations. Many tourists come to the two regions during the summer season. In addition, there are no competitive grocery stores in the first two regions. The three methods assigned the customers to five groups: loyal big spenders, loyal moderate spenders, semi-loyal potentially big spenders, potentially moderate to big spenders with limited loyalty, and infrequent customers. Because of the geographical differences between the third region and the first two regions, the membership of the least loyal group in the third region is lower than that for the first two regions. A monthly analysis of customer shopping behavior is also performed. Since it was found in an initial study that the use of the five groups did not result in clear distinctions among the groups in a monthly analysis, three groups are identified: loyal big spenders, semi-loyal spenders and least loyal spenders. It was found that customers join a more loyal group when summer comes (May to July) and start to flow back to a less loyal group after July. This makes sense, since customers do not want to spend so much time on price comparisons when summer comes, so that they may have more time for outdoor activities.

Both experiments show that the rough and fuzzy methods are more subtle and accurate in capturing the subtle differences among clusters. Web users and supermarket customers tend to change their characteristics over a period of time. These changes may be temporary or permanent. Monthly analyses of supermarket customers indicate the migration of customers over a period of six months. Such findings can help managers to implement directed marketing strategies.

5.2 Future Work

There are three aspects which can be investigated in future work.

The initial experiments show that the improved rough K-means method presented in Section 3.3 provides more robust results. Due to insufficient time, it was not possible to apply this algorithm to supermarket data mining. In future work, the improved rough K-means method will also be applied to supermarket data mining. Moreover, some special cases may be considered, in order to investigate the robustness of the improved algorithm.

In this study, 26-dimensional vectors are used in the supermarket data analysis. The large number of dimensions makes it difficult to visualize the clustering behavior. Reducing the number of dimensions and the length of the study periods may prove helpful for visualization and analysis. For example, instead of studying the weekly spending and visits for 26 weeks, the 50^{th}, 85^{th}, and 95^{th} percentile of spending and visits could be studied. The use of percentiles also make it possible to keep the same dimensionality for different period of study.

In the web data mining, the weighting for the first three attributes is [0,1]. A suggestion for future work is to use different weighting. Different weighting schemes could balance the weighting for the five attributes. Such modifications may produce different clustering results.

Finally, the clustering results for web and supermarket datasets need further semantic analysis.

Bibliography

[1] W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus. *Knowledge Discovery in Databases: An Overview.* AI Magazine, Fall 1992, pp 213-228.

[2] IBM. *IBM's Intelligent Miner Helps Safeway Focus on the Individual Shopper.* "http://www-3.ibm.com/software/data/solutions/customer/safeway/safeway.pdf", 2000.

[3] X.D. Chen, and I. Petrounias. *An Architecture for Temporal Data Mining.* Knowledge Discovery and Data Mining (Digest No. 1998/310), IEE Colloquium on, pp 8/1 - 8/4, 7 May 1998.

[4] M. Saraee, and B. Theodoulidis. *Knowledge Discovery in Temporal Databases.* Proceedings of IEE Colloquium on Knowledge Discovery in Databases, pp. 1-4, 1995.

[5] S. Weiss, and N. Indurkhya. *Predictive Data Mining.* Morgan Kaufmann Publishers, San Francisco, USA, 1998.

[6] W. H. Cheng, C. M. Tsui, and H. N. Shiu. *Survey on Spatial Data Mining of Large Spatial Database.*

[7] M. H. Dunham. *Data Mining: Introductory and Advanced Topics.* Pearson Eduction INC, 2003.

[8] K. Koperski, J. W. Han, and N. Stefanovic. *An Efficient Two-Step for Classification of Spatial Data.* Proc. Symposium on Spatial Data Handling (SDH '98), Vancouver, Canada, 1998.

[9] U. M. Fayyad, S. G. Djorgovski, and N. Weir. *Automating the Analysis and Cataloging of Sky Surveys.* Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, Menlo Park, CA, 1996.

[10] M. Ester, H. P. Kriegel, and J.Sander. *Spatial Data Mining: A Database Approach.* Proc. Int. Symp. on Large Spatial Databases (SSD'97), Berlin, Germany, pp.47-66, August, 1997.

[11] K. Koerski, J. W. Han. *Discovery of Spatial Association Rules in Geographic Information Databases.* In Advances in Spatial Databases, Proc. of 4th International Symposium, SSD'95, pp 47–66, Portland, Maine, USA, 1995.

[12] J. Sander, M. Ester, H. P. Kriegel, and X. W. Xu. *Density-based Clustering in Spatial Databases: the Algorithm GDBSCAN and its Application.* Data Mining and Knowledge Discovery , An International Journal 2(2), pp 169-194, June 1998. Kluwer Academic Publishers, Norwell, MA.

[13] M. Ester, H. P. Kriegel, X. W. Xu. *Knowledge Discovery in Large Spatial Databases: Focusing Techniques for Efficient Class Identification.* In Advances in Spatial Databases, Pro- ceedings of th Symposium, SSD'95 (Aug. 69, Portland, Maine). Springer-Verlag, Berlin, pp 67-82, 1995.

[14] R. T. Ng, and J. W. Han. *Efficient and Effective Clustering Methods for Spatial Data Mining.* 20th International Conference on Very Large Data Bases, Santiago, Chile proceedings, 12-15 September, 1994.

[15] L. Kaufman and P. J.Rousseeuw. *Finding Groups in Data: an Introduction to Cluster Analysis.* John Wiley & Sons, 1990.

[16] A. Joshi and R. Krishnapuram. *Robust Fuzzy Clustering Methods to Support Web Mining.* Proceedings of the workshop on Data Mining and Knowledge Discovery, SIGMOD'98", pp 15/1-15/8, 1998.

[17] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function.* Plenum Press, 1981.

[18] F. C. H. Rhee and C. Hwang. *A Type-2 Fuzzy C-Means Clustering Algorithm.* IFSA World Congress and 20th NAFIPS International Conference, Joint 9th , Vol. 4, pp 1926 - 1929, 25-28 July, 2001.

[19] M. C. Hung, and D. L. Yang. *An Efficient Fuzzy C-Means Clustering Algorithm.* ICDM 2001, Proceedings IEEE International Conference on, pp 225-232, 29 Nov.-2 Dec. 2001.

[20] T. W. Cheng, D. B. Goldgof, and L. O. Hall. *Fast Clustering with Application to Fuzzy Rule Generation.* International Joint Conference of the Fourth IEEE International Conference on Fuzzy Systems and The Second International Fuzzy Engineering Symposium., Proceedings of 1995 IEEE International Conference on , Vol. 4, 2, pp 2289 - 2295, 20-24 March, 1995.

[21] J. Srivastava, R. Cooley, M. Deshpande, and P. Tan. *Web Usage Mining.* Discovery and Applications of Usage Patterns from Web Data, in SIGKDD Explorations, Vol. 1, Issue 2 1-12, 2000.

[22] R. Hathaway and J. Bezdek. *Switching Regression Models and Fuzzy Clustering.* IEEE Transactions of Fuzzy Systems, Vol. 1, No. 3, pp 195-204, 1993.

[23] R. Krishnapuram, H. Frigui, and O. Nasraoui. *Fuzzy and Possibilistic Shell Clustering Algorithms and their Application to Boundary Detection and Surface Approximation, Parts I and II.* IEEE Transactions on Fuzzy Systems, Vol. 3, No. 1, pp 29-60, 1995.

[24] R. Krishnapuram and J. Keller. *A Possibilistic Approach to Clustering.* IEEE Transactions on Fuzzy Systems, Vol. 1, No. 2, pp 98-110, 1993.

[25] R. Cannon, J. Dave, and J. Bezdek. *Efficient Implementation of the Fuzzy C-Means Clustering Algorithms.* IEEE Trans. PAMI, Vol. 8, pp 248-255, 1986.

[26] T. Cheng, D.B. Goldgof, and L.O. Hall. *Fast Clustering with Application to Fuzzy Rule Generation.* In the proceedings of 1995 IEEE International Conference on Fuzzy Systems, Vol. 4, pp 2289-2295, 1995.

[27] P. Lingras. *Unsupervised Rough Set Classification using GAs. Journal of Intelligent Information Systems.* Vol. 16, No. 3, pp 215-228, 2001.

[28] P. Lingras. *Rough Set Clustering for Web Mining.* In the Proceedings of 2002 IEEE International Conference on Fuzzy Systems, 2002.

[29] P. Lingras, and C. West. *Interval Set Clustering of Web Users with Rough K-means.* Submitted to Journal of Intelligent Information Systems, 2002.

[30] T. Kohonen. *Self-Organization and Associative Memory.* Springer Verlag, Berlin, 1988.

[31] P. Lingras, M. Hogo and M. Snorek. *Interval Set Clustering of Web Users using Modified Kohonen Self-Organization Maps based on the Properties of Rough Sets.* Submitted to Web Intelligence and Agent Systems: an International Journal, 2002.

[32] X. Gao, J. Li, and W. Xie. *Parameter Optimization in FCM Clustering Algorithms.* In the Proceedings of 2000 IEEE 5th International Conference on Signal Processing, Vol. 3, pp 1457-1461, 2000.

[33] P. Lingras, R. Yan and C. West. *Fuzzy C-Means Clustering of Web Users for Educational Sites.* The 16^{th} Canadian Conference on Aritificial Intelligence, **AI' 2003**, Halifax, Nova Scotia, Canada, 2003.

[34] Z. Pawlak. *Rough Sets.* International Journal of Information and Computer Sciences, Vol. 11, pp 145-172, 1982.

[35] Y. Y. Yao, X. Li, T. Y. Lin and Q. Liu. *Representation and Classification of Rough Set Models.* in the proceedings of third International Workshop on Rough Sets and Soft Computing, pp 630-637, 1994.

[36] P. Lingras, and G. Adams. *Selection of Time-Series for Clustering Supermarket Customers, Technical report.* Dept. of Mathematics and Computer Science, Saint Mary's University, Halifax, Nova Scotia, Canada. "http://cs.smu.ca/tech_reports/txt2002_006.doc", 2002.

[37] A. Skowron and J. Stepaniuk. *Information Granules in Distributed Environment.* in New Directions in Rough Sets, Data Mining, and Granular-Soft Computing, Setsuo Ohsuga, Ning Zhong, Andrzej Skowron, Ed., Springer-Verlag, Lecture notes in Artificial Intelligence 1711, Tokyo, pp 357-365, 1999.

[38] P. Lingras and L. Young. *Multi-criteria Time-Series based Clustering of Supermarket Customers using Kohonen Networks.* in the proceedings of the 2001 International Conference on Artificial Intelligence (IC-AI'2001), June 25-28, Las Vegas, Nevada, USA, Vol 1, pp 158-164, 2001.

[39] P. Lingras and X. D. Huang. *Statistical, Ecolutional, and Neurocomputing Clustering Tehcniques: Cluster-based Versus Object-based Approaches.* Submitted to AI Review in July 2002.

[40] "http://relocatecanada.com/saintjohn", 2004.

[41] "http://www.littletechshoppe.com/ns1625/fed1997.html", 1997.