

# Applying Rough Set Concepts to Clustering

Pawan Lingras and Georg Peters

**Abstract** Clustering algorithms are probably the most commonly used methods in data mining. Applications can be found in virtually any domain; prominent areas of application are e.g. bioinformatics, engineering and marketing besides many others. In many applications the classic k-means clustering algorithm is applied. Its fuzzy version, Bezdek's fuzzy c-means has also gained tremendous attention. Another soft computing k-means algorithm based on rough set has been recently introduced by Lingras. This chapter describes how a core concept of rough sets, the lower and upper approximation of a set, can be used in clustering. Rough clusters are shown to be useful for representing groups of highway sections, web users, and supermarket customers.

## 1 Introduction

The goal of clustering is to group similar objects in one cluster and dissimilar objects in different clusters. Probably the most frequently used clustering algorithm is the classic k-means with applications in virtually any real life domain. The k-means clustering is characterized by non-overlapping, clearly separated ("crisp") clusters with bivalent memberships: an object either belongs to or does not belong to a cluster.

However, many real life applications are characterized by situations where overlapping clusters would be a more suitable representation.

---

P. Lingras (✉)

Department of Mathematics and Computer Science, Saint Mary's University, Halifax, Canada  
e-mail: [pawan.lingras@smu.ca](mailto:pawan.lingras@smu.ca)

P. Lingras

School of Computational Sciences, Swami Ramanand Teerth Marathwada University, Nanded, India

G. Peters

Department of Computer Science and Mathematics, Munich University of Applied Sciences, 80335 Munich, Germany  
e-mail: [georg.peters@cs.hm.edu](mailto:georg.peters@cs.hm.edu)

For example, consider seasons. Some days in the season winter might be undoubtedly belong to the “real” winter with snow and freezing temperatures and no indication of spring or even summer.

But at the end of the season winter, in March in the northern hemisphere, spring is “arriving”. Or in other words, many days of March are not really winter days any more. They are also not real spring days, these days in March are somehow in-between winter and spring. To address such common real life situations where clusters overlap, fuzzy c-means has been introduced [1].

Another example for the need for overlapping clusters is as follows. For diagnosis of a flu a general practitioner normally requires the temperature, of a patient and whether she has headache and cough (example inspired by Grzymała-Busse [2]). In such a case classic k-means with a bivalent classification (Flu=yes or Flu=no) is fully sufficient.

However, for some special cases these features might not be sufficient to decide if these patients suffer from flu or not (e.g. further, more detailed diagnoses are required). In such cases rough clustering is an appropriate method since it separates the objects that are definite members of a cluster from the objects that are only possible members of a cluster. In our case most patients are assigned to the clusters Flu=yes or Flu=no. But some, the tricky ones, are labeled as the “we do not know yet” patients.

Note the distinction with fuzzy clustering, where similarities are described by membership degrees while in rough clustering definite and possible members to a cluster are detected.

In contrast to original rough sets [3] which has its foundations in classic set theory, rough clustering is inspired by intervals. It utilizes the fundamental properties of original rough set theory, namely the concept of lower and upper approximations.

For almost a decade rough clustering is attracting increasing attention among researchers [4–6]. In particular, the rough k-means approach for clustering is of interest to several researchers. Lingras and West [7] provided rough k-means algorithm based on an extension of the k-means algorithm [8, 9]. Peters [10] discussed various refinements of Lingras and West’s original proposal. The rough k-means [7] and its various extensions [10–12] have been found to be effective in a number of practical applications of clustering.

This chapter first provides the essential concepts of rough k-means. In the following section the foundations of the rough k-means are presented. In Sect. 3 rough clustering is applied to highway, web users, and supermarket data to demonstrate the range of applications. The chapter concludes with a short summary in Sect. 4.

## 2 Foundations of Rough Clustering

### 2.1 Adaptation of Rough Set Theory for Clustering

Rough sets were originally proposed using equivalence relations with properties as specified by Pawlak [3, 13]. The core idea is to separate discernible from indis-

**Table 1** GP's diagnoses (categories)

Record	Symptoms		Diagnoses
	Temperature	Headache	Flu
1	high	no	yes
2	high	no	no
3	no	no	no

cernible objects and to assign objects to lower and upper approximations of a set ( $\underline{A}(X)$ ,  $\overline{A}(X)$ ).

Yao et al. [14, 15] described various generalizations of rough sets by relaxing the assumptions of an underlying equivalence relation. Such a trend toward generalization is also evident in rough mereology proposed by Polkowski and Skowron [16] and the use of information granules in a distributed environment by Skowron and Stepaniuk [17].

The present study uses such a generalized view of rough sets. If one adopts a more restrictive view of rough set theory, the rough sets developed in this paper may have to be looked upon as interval sets.

In rough clustering we are not considering all the properties of the rough sets [3, 13]. However, the family of upper and lower approximations are required to follow some of the basic rough set properties such as:

1. An object  $\mathbf{v}$  can be part of at most one lower approximation. This implies that any two lower approximations do not overlap.
2. An object  $\mathbf{v}$  that is member of a lower approximation of a set is also part of its upper approximation ( $\mathbf{v} \in \underline{A}(\mathbf{x}_i) \rightarrow \mathbf{v} \in \overline{A}(\mathbf{x}_i)$ ). This implies that a lower approximation of a set is a subset of its corresponding upper approximation ( $\underline{A}(X_i) \subseteq \overline{A}(X_i)$ ).
3. If an object  $\mathbf{v}$  is not part of any lower approximation it belongs to two or more upper approximations. This implies that an object cannot belong to only a single boundary region.

Note that these basic properties are not necessarily independent or complete. However, enumerating them will be helpful in understanding the rough set adaptation of the k-means algorithm.

Let us consider the small decision table (Table 1) depicting three diagnoses of a GP (the example is inspired by Grzymała-Busse [2]).

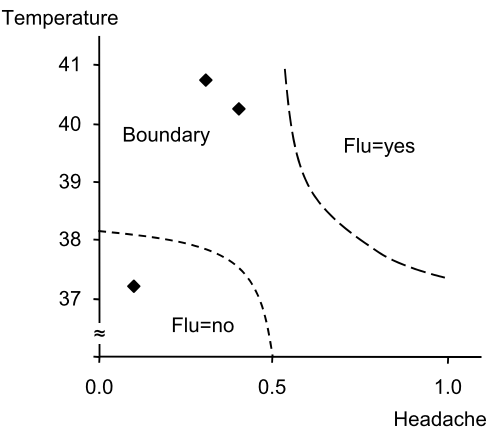
Although the symptoms of the patients 1 and 2 are indiscernible, patient 1 suffers from flu while patient 2 does not. To indicate this ambiguity, patients with these symptoms are assigned to the upper approximations of the sets Flu=yes and Flu=no. Since the diagnosis of patient 3 is clear this patient is assigned to the lower approximation of the set Flu=no.

Let us continue with our example. In rough k-means, the most widely used rough clustering algorithm, the symptoms are the features and the diagnoses correspond to the labels of the clusters. Hence, we need to map the symptoms on continues feature scales. While this is obvious for the feature temperature (in °C), let us assume that

**Table 2** GP’s diagnoses (numeric)

Record	Symptoms		Diagnoses
	Temperature	Headache	Flu
1	40.8	0.3	yes
2	40.3	0.4	no
3	37.1	0.1	no

**Fig. 1** Example: rough clustering results



we map the feature headache into an interval from 0 to 1 (0 = no headache, 1 = extremely strong headache). We may get the features as depicted in Table 2.

Figure 1 shows a pictorial representation of the results. Patient 3 is definitely a member of the cluster Flu=no (member of the lower approximation of the set Flu=no). However, the patients 1 and 2 are in “gray zone” (boundary region) between the clusters Flu=no and Flu=yes (members of the upper approximations of the sets Flu=yes and Flu=no).

2.2 Adaptation of *k*-Means to Rough Set Theory

**Classic k-Means** The most popular rough clustering approach has been derived from the classic k-means clustering approach [8, 9].

The name k-means originates from the means of the *k* clusters that are created from *n* objects. Let us assume that the objects are represented by *m*-dimensional vectors.

The objective is to assign these *n* objects to *k* clusters. Each of the clusters is also represented by an *m*-dimensional vector, which is the centroid or mean vector for that cluster. The process begins by randomly choosing *k* objects as the centroids of the *k* clusters. The objects are assigned to one of the *k* clusters based on the minimum value of the distance *d*(**v**, **x**) between the object vector

$\mathbf{v} = (v_1, \dots, v_j, \dots, v_m)$  and the cluster vector  $\mathbf{x} = (x_1, \dots, x_j, \dots, x_m)$ . The distance  $d(\mathbf{v}, \mathbf{x})$  is given as follows.

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

$$x_j = \frac{\sum_{\mathbf{v} \in \mathbf{x}} v_j}{\text{Size of cluster } \mathbf{x}} \\ \text{where } 1 \leq j \leq m.$$

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

**Rough k-Means** Incorporating rough sets into k-means clustering requires the addition of the concept of lower and upper approximations.

In particular

- (i) the calculation of the centroids needs to be adapted and
- (ii) it has to be decided whether an object is assigned to a lower or upper approximation of a cluster.

These items will be addressed in the following paragraphs in more detail.

(i) *Calculation of the Centroids.* Calculation of the centroids of clusters from conventional k-means needs to be modified to include the effects of lower as well as upper approximations.

Basically the objects are weighted based on the different importance of the lower and upper approximations.

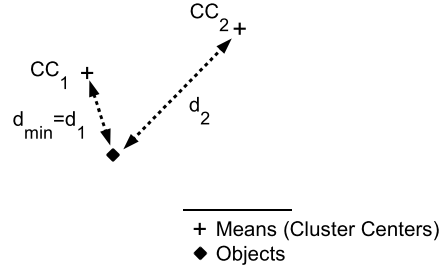
Then the modified centroid calculations for rough sets are given by:

$$\begin{aligned} &\text{IF} \quad [\underline{A}(\mathbf{x}) \neq \emptyset \text{ and } \overline{A}(\mathbf{x}) - \underline{A}(\mathbf{x}) = \emptyset] \\ &\text{THEN} \quad \left[ x_j = \frac{\sum_{\mathbf{v} \in \underline{A}(\mathbf{x})} v_j}{|\underline{A}(\mathbf{x})|} \right] \\ &\text{ELSE IF} \quad [\underline{A}(\mathbf{x}) = \emptyset \text{ and } \overline{A}(\mathbf{x}) - \underline{A}(\mathbf{x}) \neq \emptyset] \\ &\text{THEN} \quad \left[ x_j = \frac{\sum_{\mathbf{v} \in (\overline{A}(\mathbf{x}) - \underline{A}(\mathbf{x}))} v_j}{|\overline{A}(\mathbf{x}) - \underline{A}(\mathbf{x})|} \right] \\ &\text{ELSE} \quad \left[ x_j = w_{lower} \frac{\sum_{\mathbf{v} \in \underline{A}(\mathbf{x})} v_j}{|\underline{A}(\mathbf{x})|} + w_{upper} \frac{\sum_{\mathbf{v} \in (\overline{A}(\mathbf{x}) - \underline{A}(\mathbf{x}))} v_j}{|\overline{A}(\mathbf{x}) - \underline{A}(\mathbf{x})|} \right] \\ &\text{where } 1 \leq j \leq m \text{ and } w_{lower} + w_{upper} = 1. \end{aligned}$$

The parameters  $w_{lower}$  and  $w_{upper}$  correspond to the relative importance of lower and upper approximations. If the upper approximation of each cluster were equal to its lower approximation, the clusters would be conventional clusters. Thus, the first condition  $[\underline{A}(\mathbf{x}) \neq \emptyset \text{ and } \overline{A}(\mathbf{x}) - \underline{A}(\mathbf{x}) = \emptyset]$  always holds which is identical to conventional centroid calculations.

(ii) *Decide whether an object is assigned to a lower or upper approximation of a cluster.* The next step in the modification of the k-means algorithms for rough sets

**Fig. 2** Assigning an object to an approximation



is to design criteria to determine whether an object belongs to the upper or lower approximation of a cluster given as follows.

Basically, an object will be assigned to a lower approximation of a cluster when the distance between the object and the particular cluster center is much smaller than the distances to the remaining other cluster centers (see Fig. 2).

More formally, for each object vector,  $\mathbf{v}$ , let  $d(\mathbf{v}, \mathbf{x}_j)$  be the distance between itself and the centroid of cluster  $\mathbf{x}_j$ . Then we have two steps to determine the membership of an object:

1. Determine the nearest centroid:

$$d_{min} = d(\mathbf{v}, \mathbf{x}_i) = \min_{1 \leq j \leq k} d(\mathbf{v}, \mathbf{x}_j).$$

2. Check if further centroids are not significantly farther away than the closest one:  
Let  $T = \{j : d(\mathbf{v}, \mathbf{x}_j)/d(\mathbf{v}, \mathbf{x}_i) \leq \text{threshold and } i \neq j\}$ . Then we get:
  - If  $T \neq \emptyset$  then at least one other centroid is similarly close to the object.
  - If  $T = \emptyset$  then no other centroids are similarly close to the object.

Hence, we get the following rule for the assignment of the objects to the approximations:

IF  $[T \neq \emptyset]$   
 THEN  $[\mathbf{v} \in \overline{A}(\mathbf{x}_i) \text{ and } \mathbf{v} \in \overline{A}(\mathbf{x}_j), \forall j \in T]$   
 ELSE  $[\mathbf{v} \in \overline{A}(\mathbf{x}_i) \text{ and } \mathbf{v} \in \underline{A}(\mathbf{x}_i)]$

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 approximations are constructed based on the criteria described above.

While this chapter only describes the rough k-means algorithm, a number of other alternatives based on Kohonen self-organizing maps [18], evolutionary partitive approach [11], and evolutionary k-medoids [19] are also available.

### 3 Applications of Rough Clustering

The rough k-means has already been applied to real life data in several domains. In this section we particularly address applications in the areas of

- Traffic data (Sect. 3.1),
- Web user data (Sect. 3.2) and
- Supermarket data (Sect. 3.3)

in detail.

### ***3.1 Rough Clustering Highway Sections***

Seasonal and permanent traffic counters scattered across a highway network are the major sources of traffic data. These traffic counters measure the traffic volume—the number of vehicles that have passed through a particular section of a lane or highway in a given time period. Traffic volumes can be expressed in terms of hourly or daily traffic. More sophisticated traffic counters record additional information such as the speed, length and weight of the vehicle. Highway agencies generally have records from traffic counters collected over a number of years. In addition to obtaining data from traffic counters, traffic engineers also conduct occasional surveys of road users to get more information.

The permanent traffic counter (PTC) sites are grouped to form various road categories. These categories are used to develop guidelines for the construction, maintenance and upgrading of highway sections. In one commonly used system, roads are categorized on the basis of trip purpose and trip length characteristics [20].

Examples of resulting categories are commuter, business, long distance, and recreational highways. The trip purpose provides information about the road users, an important criterion in a variety of traffic engineering analyzes. Trip purpose information can be obtained directly from the road users, but since all users cannot be surveyed, traffic engineers study various traffic patterns obtained from seasonal and permanent traffic counters and sample surveys of a few road users.

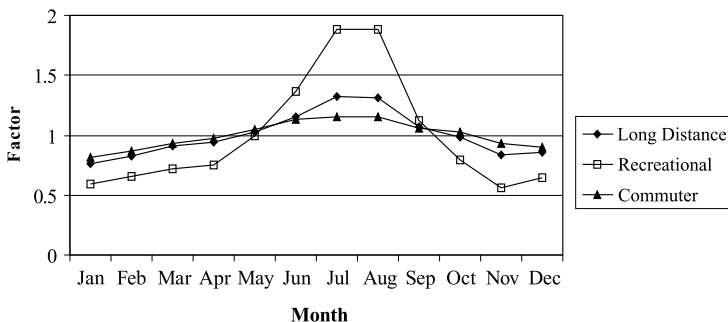
The present study is based on a sample of 264 monthly traffic patterns—variation of monthly average daily traffic volume in a given year—recorded between 1987 and 1991 on Alberta highways. The distribution of PTCs in various regions are determined based on the traffic flow through the provincial highway networks. The patterns obtained from these PTCs represent traffic from all major regions in the province.

The rough set genomes used in the experiment consisted of 264 genes, one gene per pattern. The hypothetical clustering scheme consisted of three categories:

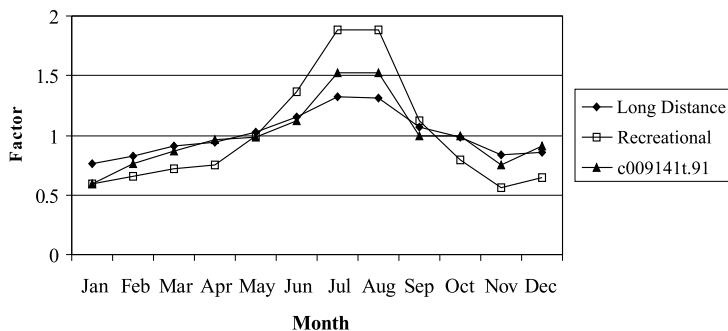
1. Commuter/business,
2. Long distance, and
3. Recreational.

The rough set clustering scheme was expected to specify lower and upper approximations of these categories.

The upper and lower approximations of the commuter/business, long distance, and recreational clusters were also checked against the geography of Alberta highway networks. More details of the experiment can be found in [21].



**Fig. 3** Monthly patterns for the lower approximations



**Fig. 4** Monthly pattern that may be long distance or recreational

Figure 3 shows the monthly patterns for the lower approximations of the three groups: commuter/business, long distance, and recreational. The average pattern for the lower approximation of commuter/business cluster has the least variation over the year. The recreational cluster, conversely, has the most variation. The variation for long distance cluster is less than the recreational but more than the commuter/business cluster. Figure 4 shows one of the highway sections near counter number C013201 that may have been Commuter/Business or Long Distance in 1985. It is clear that the monthly pattern for the highway section falls in between the two clusters. The counter C013201 is located on highway 13, 20 km. west of Alberta–Saskatchewan border. It is an alternate route for travel from the city of Saskatoon and surrounding townships to townships surrounding the city of Edmonton. A similar observation can be made in Fig. 5 for highway section C009141 that may have been Long Distance or Recreational in 1991. The counter C009141 is located on highway 9, 141 km. west of Alberta–Saskatchewan border. The traffic on that particular road seems to have higher seasonal variation than a long distance road. Rough set representation of clusters enables us to identify such intermediate patterns.



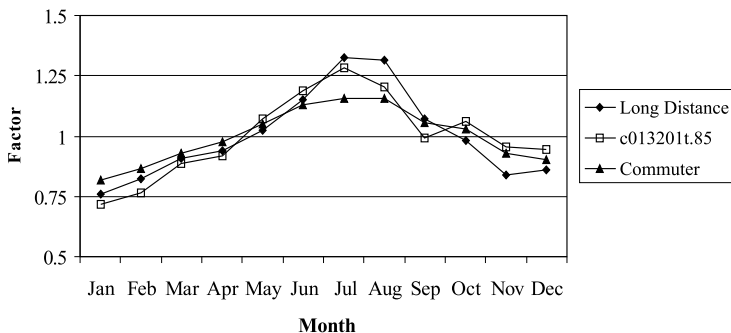


Fig. 5 Monthly pattern that may be commuter/business or long distance

### 3.2 Rough Clustering Web Users

The study data was obtained from the Web access logs of the first three courses in computing science at Saint Mary's University over a sixteen-week period. Students' attitudes toward the course vary a great deal. It was hoped that the profile of visits would reflect some of the distinctions between the students. For the initial analysis, it was assumed that the visitors could fall into one of the following three categories:

1. **Studious:** These visitors download the current set of notes. Since they download a limited/current set of notes, they probably study classnotes on a regular basis.
2. **Crammers:** These visitors download a large set of notes. This indicates that they have stayed away from the classnotes for a long period of time. They are planning for pre-test cramming.
3. **Workers:** These visitors are mostly working on class or lab assignments or accessing the discussion board.

The rough set clustering scheme was expected to specify lower and upper approximations for these categories.

It was hoped that the variety of user behaviors mentioned above would be identifiable based on the number of Web accesses, types of documents downloaded, and time of day. Certain areas of the Web site were protected and the users could only access them using their IDs and passwords. The activities in the restricted parts of the Web site consisted of submitting a user profile, changing a password, submission of assignments, viewing the submissions, accessing the discussion board, and viewing current class marks. The rest of the Web site was public. The public portion consisted of viewing course information, the lab manual, classnotes, class assignments, and lab assignments.

If the users only accessed the public Web site, their IDs would be unknown. Therefore, the Web users were identified based on their IP address. This also assured that the user privacy was protected. A visit from an IP address started when the first request was made from the IP address. The visit continued as long as the consecutive requests from the IP address had sufficiently small delay.

**Table 3** Cardinalities of the clusters for three techniques

Course	Cluster	Lower approximation	Conventional clusters
First	Studious	1412	1814
	Crammers	288	406
	Workers	5350	5399
Second	Studious	1197	1699
	Crammers	443	634
	Workers	1677	3697
Third	Studious	223	318
	Crammers	69	89
	Workers	906	867

The Web logs were preprocessed to create an appropriate representation of each user corresponding to a visit. The abstract representation of a Web user is a critical step that requires a good knowledge of the application domain. Previous personal experience with the students in the course suggested that some of the students print preliminary notes just before a class and an updated copy after the class. Some students view the notes on-line on a regular basis. Some students print all the notes around important days such as midterm and final examinations. In addition, there are many accesses on Tuesdays and Thursdays, when the in-laboratory assignments are due. On and off campus points of access can also provide some indication of a user's objectives for the visit. Based on some of these observations, it was decided to use the following attributes for representing each visitor:

1. On campus/Off campus access.
2. Day time/Night time access: 8 a.m. to 8 p.m. was considered to be the day time.
3. Access during lab/class days or non-lab/class days: All the labs and classes were held on Tuesday and Thursday. The visitors on these days are more likely to be workers.
4. Number of hits.
5. Number of classnotes downloaded.

The first three attributes had binary values of 0 or 1. The last two values were normalized. Since the classnotes were the focus of the clustering, the last variable was assigned higher importance. The resulting rough set clustering schemes were subjectively analyzed. The results were compared with conventional clustering. More details about the experiments can be found in [22].

Table 3 shows the cardinalities of conventional clusters, the modified k-means based on rough set theory. The actual numbers in each cluster vary based on the characteristics of each course. For example, the first term course had significantly more workers than studious visitors, while the second term course had more studious visitors than workers. The increase in the percentage of studious visitors in the second term seems to be a natural progression. Interestingly, the second year

**Table 4** The conventional k-means cluster center vectors

Course	Cluster	Campus access	Day night	Lab day	Hits	Doc req
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

course had significantly large number of workers than studious visitors. This seems to be counter-intuitive. However, it can be explained based on the structure of the web sites. Unlike the two first year courses, the second year course did not post the classnotes on the Web. The notes downloaded by these students were usually sample programs that were essential during their laboratory work.

Table 4 shows cluster center vectors from the conventional k-means. It was possible to identify the three clusters as studious, workers, and crammers, from the results obtained using the conventional k-means algorithm. The crammers had the highest number of hits and classnotes in every data set. The average number of notes downloaded by crammers varied from one set to another. The studious visitors downloaded the second highest number of notes. The distinction between workers and studious visitors for the second course was also based on other attributes. For example, in the second data set, the workers were more prone to come on lab days, access 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. Table 5 shows the modified k-means center vectors. These center vectors are comparable to the conventional centroid vectors. For the second data set, the modified k-means is more sensitive to the differences between studious and crammers in the first three attributes than the conventional k-means.

### 3.3 Rough Clustering Supermarket Customers

The data used in the study was supplied by a supermarket chain. The data consisted of transactional records from three regions. The first region, S1, consisted of one store in a rural setting. The second rural region (S2) was served by five stores, while the third region was an urban area with six stores. The data was collected over a twenty-six week period: October 22, 2000–April 21, 2001. Lingras and Adams [23] used data on the spending and visits of supermarket customers for clustering

**Table 5** The modified k-means cluster center vectors

Course	Cluster	Campus access	Day night	Lab day	Hits	Doc req
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

those customers. The use of average values of these variables may hide some of the important information present in the temporal patterns. Therefore, Lingras and Adams [23] used the weekly time series values. It is possible that customers with similar profiles may spend different amounts in a given week. However, if the values were sorted, the differences between these customers may vanish. For example, three weeks spending of customer A may be CAD 10, CAD 30, and CAD 20. Customer B may spend CAD 20, CAD 10, and CAD 30 in those three weeks. If the two time-series were compared with each other, the two customers may seem to have completely different profiles. However, if the time-series values were sorted, the two customers would have identical patterns. Therefore, the values of these variables for 26 weeks were sorted, resulting in a total of 52 variables. A variety of values for  $k$  (number of clusters) were used in the initial experiments. A setting of  $k = 5$  seemed to provide a reasonable clustering.

Figure 6 shows the average spending and visit patterns for the lower approximations of the five clusters. The patterns enable us to distinguish between the five types of customers as:

- Loyal big spenders (G1)
- Loyal moderate spenders (G2)
- Semi-loyal potentially big spenders (G3)
- Potentially moderate to big spenders with limited loyalty (G4)
- Infrequent customers (G5)

The patterns of these clusters for the three regions were mostly similar. However, there was an interesting difference in S1 region. Even though for most weeks *loyal moderate spenders* (G2) had higher spending than *semi-loyal potentially big spenders* (G3), the highest spending of G3 was higher than G2. The region has only one store and hence it is likely that *semi-loyal potentially big spenders* do not find it convenient to shop at the supermarket on a regular basis.

While the lower approximations tend to provide distinguishing characteristics of various clusters, the boundary regions of the clusters tend to fall between the lower approximations of two regions. This fact is illustrated in Fig. 7(a).

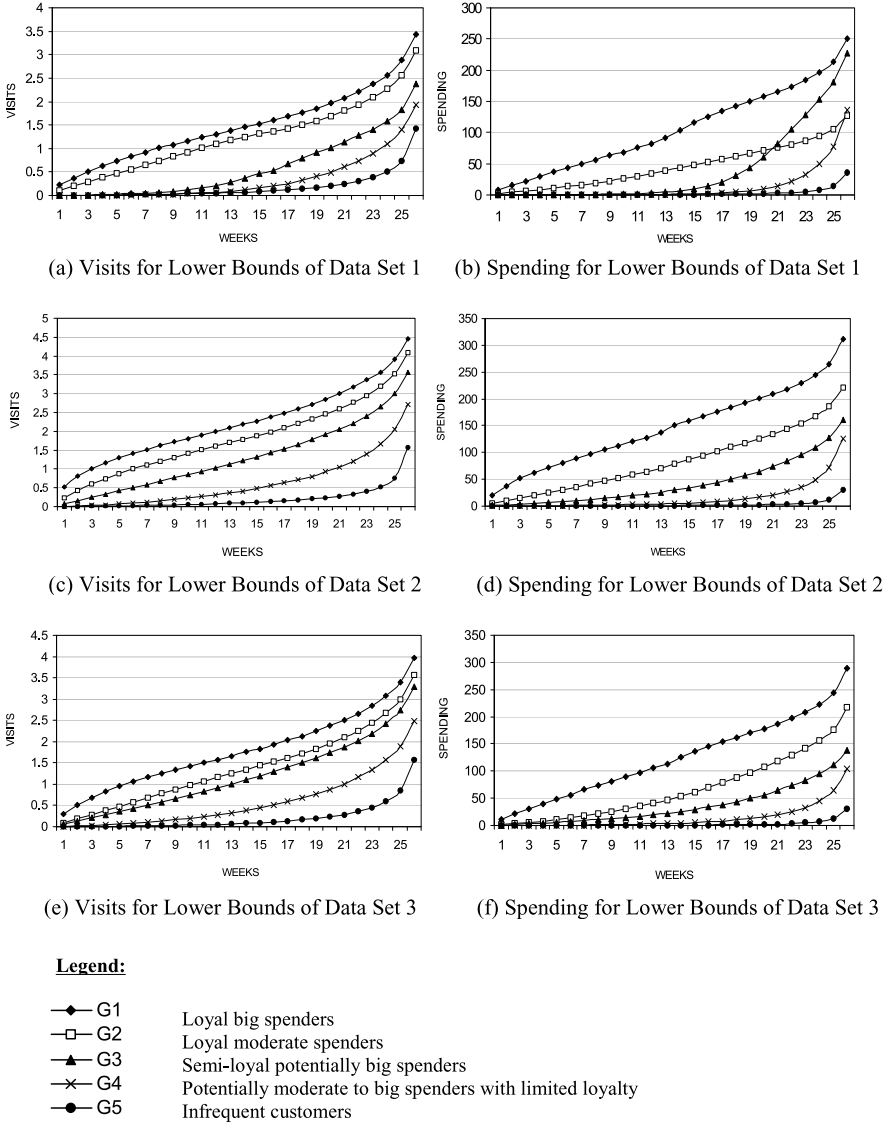
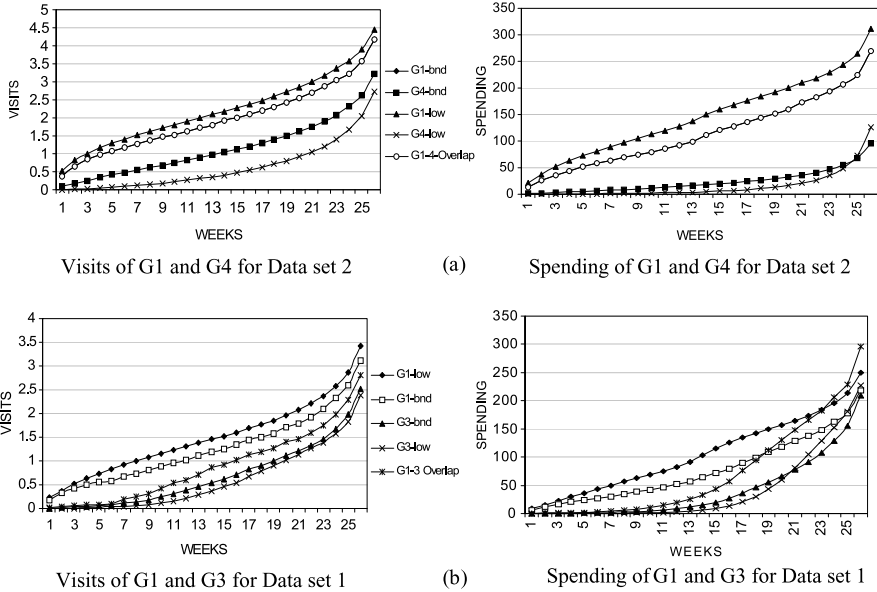


Fig. 6 Patterns for visits and spending over 26 weeks

There is a large difference between the lower approximations of groups *loyal big spenders* (G1) and *potentially moderate to big spenders with limited loyalty* (G4). However, their boundary regions seem to be less distinct. The boundary regions of G1 and G4 fall between the lower approximations of those groups. The figure also shows the patterns for the overlap of the two groups. Figure 7(b) shows a similar comparison for loyal big spenders (G1) and semi-loyal potentially big spenders (G3).



**Fig. 7** Comparison of two interval clusters

## 4 Conclusion

In this chapter a brief introduction on rough clustering was given. In particular the rough k-means was described. It was applied to highway section, web user, and supermarket customer data.

Rough sets may provide representation of clusters, where it is possible for an object to belong to more than one cluster. This is of particular interest when buffer zones between clusters are immanent or a “buffer zone” is required to diminish the clustering mistakes. The objects in such a buffer zone need a second look (further information, an expert opinion etc.) before they can eventually be assigned to a cluster.

Hence, the rough k-means has proven to be an important enrichment to clustering approaches, particularly in the direction of soft computing methods.

**Acknowledgements** The authors would like to thank the Natural Sciences and Engineering Research Council of Canada and the Faculty of Graduate Studies and Research, Saint Mary’s University for funding. Data from Alberta Highways, Saint Mary’s University, and the supermarket is also appreciated.

## References

1. Bezdek, J.: Pattern Recognition with Fuzzy Objective Algorithms. Plenum Press, New York (1981)

2. Grzymala-Busse, J.W.: Rough set theory with applications to data mining. In: Negoita, M.G., Reusch, B. (eds.) *Real World Applications of Computational Intelligence*, pp. 221–244. Springer, Berlin (2005)
3. Pawlak, Z.: *Rough Sets: Theoretical Aspects of Reasoning About Data*. Kluwer Academic, Dordrecht (1991)
4. Hirano, S., Tsumoto, S.: Rough clustering and its application to medicine. *J. Inf. Sci.* **124**, 125–137 (2000)
5. Peters, J.F., Skowron, A., Suraj, Z., Rzasa, W., Borkowski, M.: Clustering: a rough set approach to constructing information granules. In: *Proceedings of 6th International Conference on Soft Computing and Distributed Processing (SCDP 2002)*, pp. 57–61 (2002)
6. Voges, K.E., Pope, N.K., Brown, M.R.: A rough cluster analysis of shopping orientation data. In: *Proceedings Australian and New Zealand Marketing Academy Conference*, Adelaide, pp. 1625–1631 (2003)
7. Lingras, P., West, C.: Interval set clustering of web users with rough k-means. *J. Intell. Inf. Syst.* **23**(1), 5–16 (2004)
8. Hartigan, J.A., Wong, M.A.: Algorithm as136: a k-means clustering algorithm. *Appl. Stat.* **28**, 100–108 (1979)
9. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 281–297 (1967)
10. Peters, G.: Some refinements of rough k-means. *Pattern Recognit.* **39**, 1481–1491 (2006)
11. Mitra, S.: An evolutionary rough partitive clustering. *Pattern Recognit. Lett.* **25**(12), 1439–1449 (2004)
12. Peters, G.: Outliers in rough k-means clustering. In: *Pattern Recognition and Machine Intelligence (PReMI 2005)*. LNCS, vol. 3776, pp. 702–707. Springer, Berlin (2005). [ISI SCI] [SCOPUS]
13. Pawlak, Z.: Rough sets. *Int. J. Inf. Comput. Sci.* **11**, 145–172 (1982)
14. Yao, Y.Y., Lin, T.Y.: Generalization of rough sets using modal logic. *Intell. Autom. Soft Comput.* **2**(2), 103–120 (1996)
15. Yao, Y.Y.: Constructive and algebraic methods of the theory of rough sets. *Inf. Sci.* **109**, 21–47 (1998)
16. Polkowski, L., Skowron, A.: Rough mereology: a new paradigm for approximate reasoning. *Int. J. Approx. Reason.* **15**(4), 333–365 (1996)
17. Skowron, A., Stepaniuk, J.: Information granules in distributed environment. In: Zhong, N., Skowron, A., Ohsuga, S. (eds.) *New Directions in Rough Sets, Data Mining, and Granular-soft Computing*. Lecture Notes in Artificial Intelligence, vol. 1711, pp. 357–365. Springer, Tokyo (1999)
18. Lingras, P., Hogo, M., Snorek, M., Leonard, B.: Clustering supermarket customers using rough set based Kohonen networks. In: *Proceedings of the Fourteenth International Symposium on Methodologies for Intelligent Systems*. Lecture Notes in Artificial Intelligence, vol. 2871, pp. 169–173. Springer, Tokyo (2003)
19. Peters, G., Lampart, M., Weber, R.: Evolutionary rough k-medoids clustering. *Trans. Rough Sets* **VIII**, 289–306 (2008). LNCS 5084
20. Sharma, S.C., Werner, A.: Improved method of grouping provincewide permanent traffic counters. *Transp. Res. Rec.* **815**, 13–18 (1981)
21. Lingras, P.: Unsupervised rough set classification using gas. *J. Intell. Inf. Syst.* **16**(3), 215–228 (2001)
22. Lingras, P., Yan, R., West, C.: Fuzzy c-means clustering of web users for educational sites. In: *Proceedings of the Sixteenth Conference of the Canadian Society of Computational Studies of the Intelligence*. Advances in Artificial Intelligence, vol. 2671, pp. 557–562. Springer, Tokyo (2003)
23. Lingras, P., Adams, G.: Selection of time-series for clustering supermarket customers. Technical report 2002-006, Department of Mathematics and Computing Science, Saint Mary's University, Halifax, NS, Canada (2002)

Rough Sets: Selected Methods and Applications in  
Management and Engineering

Peters, G.; Lingras, P.; Slezak, D.; Yao, Y. (Eds.)

2012, X, 214 p., Hardcover

ISBN: 978-1-4471-2759-8