Interval Set Clustering of Web Users using Modified Kohonen Self-Organizing Maps based on the Properties of Rough Sets

Pawan Lingras*, Mofreh Hogo**, and Miroslav Snorek **

*Department of Math and Computer Science, Saint Mary's University, Halifax, Nova Scotia, Canada, B3H 3C3.

**Department of Computer Science and Engineering, Faculty of Electrical Engineering, Czech Technical University, Karlovo Nam. 13, 121 35 Prague 2, Czech Republic

Abstract

Web usage mining involves application of data mining techniques to discover usage patterns from the web data. Clustering is one of the important functions in web usage mining. The likelihood of bad or incomplete web usage data is higher than the conventional applications. The clusters and associations in web usage mining do not necessarily have crisp boundaries. Researchers have studied the possibility of using fuzzy sets in web mining clustering applications. Recent attempts have adapted the K-means clustering algorithm as well as genetic algorithms based on rough sets to find interval sets of clusters. The genetic algorithms based clustering may not be able to handle large amounts of data. The K-means algorithm does not lend itself well to adaptive clustering. This paper proposes an adaptation of Kohonen self-organizing maps based on the properties of rough sets, to find the interval sets of clusters. Experiments are used to create interval set representations of clusters of web visitors on three educational web sites.

Keywords: Clustering, Interval Sets, Kohonen Self-organizing Maps, Web Usage Mining, Rough Sets, Unsupervised Learning.

1. Introduction

Web mining can be broadly divided into three classes: content mining, usage mining, and structure mining [1]. Web usage mining applies data mining techniques to discover usage patterns from the Web data, in order to understand and better serve the needs of Web-based applications. Web usage mining consists of three phases, namely preprocessing, pattern discovery, and pattern analysis. While content mining and structure mining utilize the real or primary data on the web, web usage mining uses secondary data generated by the users' interaction with the web. Web usage data includes data from web server access logs, proxy server logs, browser logs, user profiles, registration files, user sessions or transactions, user queries, bookmark folders, mouse clicks and scrolls, and any other data generated by the interaction between users and the web. Logs of web access available on most servers are good examples of the data sets used in web usage mining. Web usage mining includes creation of user profiles, user access patterns, and navigation paths. The results of web usage mining are used by e-commerce companies for tracking customer behavior on their sites.

Clustering analysis is an important function in web usage mining, which 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 important characteristics of the web visitors. Due to the ease of movement from one portal to another, web users can be very mobile. If a particular web site doesn't satisfy the needs of a user in a relatively short period of time, the user will quickly move on to another web site. Therefore, it is very important to understand the needs and characteristics of web users. Clustering in web mining faces several additional challenges compared to

traditional applications [2], the clusters tend to have fuzzy or rough boundaries. The membership of an object in a cluster may not be precisely defined. There is 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, the possibility of the presence of outliers in the data set is high. Joshi and Krishnapuram [2] argued that the clustering operation in web mining involves modeling an unknown number of overlapping sets. They proposed the use of fuzzy clustering [3, 4, 5] for grouping the web users.

Lingras [8] described how a rough set theoretic clustering scheme could be represented using a rough set genome. The resulting genetic algorithms (GAs) were used to evolve groupings of highway sections represented as interval or rough sets. Lingras [9] applied the unsupervised rough set clustering based on GAs for grouping web users of a first year University course. He hypothesized that there are three types of visitors: studious, crammers, and workers. Studious visitors 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 lab and class assignments. Generally, the boundaries of these classes will not be precise. The preliminary experimentation by Lingras [9] illustrated the feasibility of rough set clustering for developing user profiles on the web. However, the clustering process based on GAs seemed computationally expensive for scaling to a larger data set. Lingras and West [11] provided a theoretical and experimental analysis of a modified K-means clustering based on the properties of rough sets. It was used to classify the visitors to an academic web site into upper and lower bounds of the three classes mentioned above. The modified K-means approach is suitable for large data sets. The Kohonen neural network or self-organizing map [14] is another popular clustering

technique. The Kohonen network is desirable in some applications due to its adaptive capabilities. This paper introduces the interval set clustering using a modification of the Kohonen self-organizing maps based on rough set theory. The proposed algorithm was used to find cluster intervals of web users. Three web sites that were used for the experimentation catered to two first year and one second year courses. The students used the web site for downloading class-notes and lab assignments; downloading, submitting and viewing class assignments; checking their current marks; as well as for accessing a discussion board. These web sites were accessed from a variety of locations. Only some of the web 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. This paper also provides a comparison of user behavior among first and second year students. The experiments show that the modified Kohonen network provides reasonable interval sets of clusters by adjusting to the changing user behaviour.

2. Review Of Literature

2.1 Rough Set Theory

The notion of rough set was proposed by Pawlak [6]. This section provides a brief summary of the concepts from rough set theory essential for introducing the Kohonen rough set theoretic 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 called 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 $E \subseteq U/R$, we say that *u* and *v* are indistinguishable. The equivalence classes of *R* are called the 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 $\varnothing$ is also considered a special composed set. *Com*(*A*) denotes the family of all composed sets. Since it is not possible to differentiate the elements within the same equivalence class, one may not be able to obtain a precise representation for an arbitrary set $X \subseteq U$ in terms of elementary sets in A. Instead, its lower and upper bounds may represent the set X. The lower bound $\underline{A}(X)$ is the union of all the elementary sets, which are subsets of X. The upper bound $\overline{A}(X)$ is the union of all the elementary sets that have a non-empty intersection with X.

The pair $(\underline{A}(X), \overline{A}(X))$ is the representation of an ordinary set of X in the approximation space $A = (U, R)$, or simply the rough set of X. The elements in the lower bound of X definitely belong to X, while elements in the upper bound of X may or may not belong to X. Fig.1 illustrates the lower and upper approximation. It can be verified, that for any subsets $X, Y \subseteq U$, the following lemmas hold [6].

$$\underline{A}(X \cap Y) = \underline{A}(X) \cap \underline{A}(Y), \quad \text{(L1)}$$

$$\underline{A}(X \cup Y) \supseteq \underline{A}(X) \cup \underline{A}(Y), \quad \text{(L2)}$$

$$\overline{A}(X \cap Y) \subseteq \overline{A}(X) \cap \overline{A}(Y), \quad \text{(L3)}$$

$$\overline{A}(X \cup Y) = \overline{A}(X) \cup \overline{A}(Y), \tag{L4}$$

$$\overline{A}(-X) = -\underline{A}(X), \quad \underline{A}(-X) = -\overline{A}(X), \tag{L5}$$

$$X \supseteq Y \Rightarrow (\underline{A}(X) \supseteq \underline{A}(Y), \overline{A}(X) \supseteq \overline{A}(Y)), \tag{L6}$$

$$\underline{A}(U) = \overline{A}(U) = U, \tag{L7}$$

$$\underline{A}(\varnothing) = \overline{A}(\varnothing) = \varnothing, \tag{L8}$$

2.2. Kohonen Self-Organizing Maps

Fig. 2 illustrates the conventional Kohonen network architecture for the one-dimensional case. The unsupervised learning using the Kohonen rule [14] uses competitive learning approach. In competitive learning, the output neurons compete with each other. The winner output neuron has the output of 1, the rest of the output neurons have outputs of 0. The competitive learning is suitable for classifying a given pattern into exactly one of the mutually exclusive clusters. The network is used to group patterns represented by *m*-dimensional vectors into *k* groups. The network consists of two layers. The first layer is called the input layer and the second layer is called the Kohonen layer. The network receives the input vector for a given pattern. If the pattern belongs to the i^{th} group, then i^{th} neuron in the Kohonen layer has a output value of one and other Kohonen layer neurons have output values of zero. Each connection is assigned a weight w_i. Weights of all the connections to a Kohonen layer neuron make up an *m*-dimensional weight vector **w**. The weight vector **w** for a Kohonen layer neuron is the vector representation of the group corresponding to that neuron. For any input vector **v**, the network compares the input with the weight vector for a group using the measure such as $d(\mathbf{w}, \mathbf{v})$:

$$d(\mathbf{w},\mathbf{v}) = \frac{\sum_{j=1}^{m}(w_j - v_j)^2}{m} \quad (1)$$

The pattern **v** belongs to the group with minimum value for $d(\mathbf{w},\mathbf{v})$. The Kohonen neural network generates the clusters through a learning process as follows: Initially, the network connections are assigned somewhat arbitrary weights. The training set of input vectors is presented to the network several times. For each iteration, the weight vector **w** for a group that is closest to the pattern **v** is modified using the equation:

$$\mathbf{w}_{new} = \mathbf{w}_{old} + \alpha(t) \times (\mathbf{v} - \mathbf{w}_{old}), \quad (2)$$

where $\alpha(t)$ is a learning factor which starts with a high value at the beginning of the training process and is gradually reduced as a function of time.

3. Rough Set Based Kohonen Self Organizing Maps

Rough sets were 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 [6] are satisfied. Yao et al. [12] described various generalizations of rough sets by relaxing the assumptions of an underlying equivalence relation. Skowron and Stepaniuk [13] discuss a similar generalization of rough set theory.

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. Lingras [8] proposed the unsupervised rough set clustering based on genetic algorithms to create the interval sets of clusters for web users. Lingras and West [11] proposed an adaptation of the K-means algorithm based on rough set theory for interval set clustering of web users. This paper uses some of the concepts from Lingras and West [11] to create intervals of clusters using the

Kohonen self-organizing maps. Let us consider a hypothetical classification scheme $U/P = \{X_1, X_2, ..., X_k\}$, which partitions the set U based on certain criteria. The actual values of X_i are not known. The classification of web users is an example of such a hypothetical classification scheme. Depending on the predominant usage, a set of web visitors can be classified as crammers, workers, or studious. However, the actual sets corresponding to each one of these classes are not known. Let us assume that due to insufficient knowledge it is not possible to precisely describe the sets X_i, $1 \le i \le k$, in the partition. However, it is possible to define each set $X_i \in U/R$ using its lower and upper bounds ($\underline{A}(X)$, $\overline{A}(X)$) based on the available information. In this study, the available information consists of web access logs. Since vectors represent the objects and clusters in the Kohonen rough set clustering algorithm, we will use vector representations, **v** for an object and $\mathbf{x}_i$ for cluster X_i. We are considering the upper and lower bounds of only a few subsets of U. Therefore, it is not possible to verify all the properties of rough sets [6]. However, the family of upper and lower bounds of $\mathbf{x}_i \in U/R$ are required to follow some of the basic rough set properties such as:

- An object **v** can be part of at most one lower bound (P1)
- $\mathbf{v} \in \underline{A}(\mathbf{x}_i) \Rightarrow \mathbf{v} \in \overline{A}(\mathbf{x}_i)$ (P2)

- An object **v** is not part of any lower bound (P3)

$\Updownarrow$

v belongs to two or more upper bounds.

Properties (P1)-(P3) can be obtained from (L1)-(L8) and the fact that $X_i \cap X_j = \varnothing, i \neq j$.

It is important to note that, (P1)-(P3) are not necessarily independent or complete. However, enumerating them will be helpful in understanding the rough set adaptation of the Kohonen neural networks.

Incorporating rough sets into the Kohonen algorithm requires an addition of the concept of lower and upper bounds in the equations, which are used for updating the weights of the winners. The Kohonen rough set architecture is similar to the conventional Kohonen architecture. It consists of two layers, an input layer and the Kohonen rough set layer (rough set output layer). These two layers are fully connected. Each input layer neuron has a feed forward connection to each output layer neuron. Fig.3 illustrates the Kohonen rough set neural network architecture for one-dimensional case. A neuron in the Kohonen layer consists of two parts, a lower neuron and an upper neuron. The lower neuron has an output of 1, if an object belongs to the lower bound of the cluster. Similarly, a membership in the upper bound of the cluster will result in an output of 1 from the upper neuron. Since an object belonging to the lower bound of a cluster also belongs to its upper bound, when lower neuron has an output of 1, the upper neuron also has an out of 1. However, membership in the upper bound of a cluster does not necessarily imply the membership in its lower bound. Therefore, the upper neuron contains the lower neuron.

Figs. 4 and 5 provide some cases to explains outputs from the Kohonen rough set neural

network works based on properties (P1), (P2), and (P3). Fig. 4 shows some of the possible outputs, while Fig. 5 shows some of the invalid outputs from the network. Fig. 4(a) shows a case where an object belongs to lower bound of cluster $\mathbf{x}_2$. Based on the property (P2), it also belongs to the upper bound of $\mathbf{x}_2$. Fig. 4(b) shows a situation where an object belongs to the upper bounds of clusters $\mathbf{x}_1$ and $\mathbf{x}_2$. The object in Fig. 4(c) belongs to the upper bounds of clusters $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$. Fig. 5(a) shows an invalid situation where an object belongs *only* to the upper bound of the cluster $\mathbf{x}_3$. This is a violation of the property (P3). Fig. 5(b) shows a violation of property (P1), where an object belongs to lower bound of $\mathbf{x}_1$ as well as the upper bound of $\mathbf{x}_2$. Similarly, a violation of property (P2) can be seen in an invalid case shown in Fig. 5(c). Here the object *only* belongs to the lower bound of cluster $\mathbf{x}_3$ and not its upper bound. The modification of the Kohonen algorithm must make sure that the properties (P1)-(P3) are obeyed by avoiding cases such as the ones shown in Fig. 5. The interval clustering provides good results, if initial weights are obtained by running the conventional Kohonen learning. The next step in the modification of the Kohonen algorithm for obtaining rough sets is to design criteria to determine whether an object belongs to the upper or lower bounds of a cluster. For each object vector, **v**, let $d(\mathbf{v}, \mathbf{x}_i)$ be the distance between itself and the weight vector $\mathbf{x}_i$ of cluster X_i. The ratios $\frac{d(\mathbf{v}, \mathbf{x}_i)}{d(\mathbf{v}, \mathbf{x}_j)}, 1 \le i, j \le k$, were used to determine the membership of **v** as follows:

1. If $d(\mathbf{v}, \mathbf{x}_i)$ is the minimum for $1 \le i \le k$ and $\frac{d(\mathbf{v}, \mathbf{x}_i)}{d(\mathbf{v}, \mathbf{x}_j)} \ge$ *threshold* for any pair (i, j),

then $\mathbf{v} \in \overline{A}(\mathbf{x}_i)$ and $\mathbf{v} \in \overline{A}(\mathbf{x}_j)$. Furthermore, **v** is not part of any lower bound. The above criterion guarantees that property (P3) is satisfied. The weight vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ are modified as:

$\mathbf{x}_i^{new} = \mathbf{x}_i^{old} + \alpha_{upper}(t) \times (\mathbf{v} - \mathbf{x}_i^{old})$, and

$\mathbf{x}_j^{new} = \mathbf{x}_j^{old} + \alpha_{upper}(t) \times (\mathbf{v} - \mathbf{x}_j^{old})$.

2. Otherwise, $\mathbf{v} \in \underline{A}(\mathbf{x}_i)$ such that $d(\mathbf{v}, \mathbf{x}_i)$ is the minimum for $1 \le i \le k$. In addition, by property (P2), $\mathbf{v} \in \overline{A}(\mathbf{x}_i)$. The weight vector $\mathbf{x}_i$ is modified as:

$\mathbf{x}_i^{new} = \mathbf{x}_i^{old} + \alpha_{lower}(t) \times (\mathbf{v} - \mathbf{x}_i^{old})$.

Usually, $\alpha_{lower}(t) > \alpha_{upper}(t)$. It can be easily verified that the above algorithm preserves properties (P1)-(P3). The following section describes experiments with web logs on three web sites, which suggest that the proposed modification of the Kohonen neural networks provide reasonable interval set representations of clusters.

4. Study Data and Design of the Experiment

4. 1 Data Description

The study data was obtained from web access logs of three courses. These courses represent a sequence of required courses for computing science programme at Saint Mary's University. Two courses were for the first year students. The third course was for the second year students. The first course is “Introduction to Computing Science and Programming” offered in the first term of first year. The initial number of students in the course was 180. The number reduced over the course of the semester to 130 to 140

students. The students in the course come from a wide variety of backgrounds, such as 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, students' attitudes towards the course also vary a great deal. The second course is “Intermediate Programming and Problem Solving” offered in the second term of the first year. The initial number of students in the course was around 100. The number reduced over the course of the semester to 90 students. The students have similar backgrounds and motivations as the first course. However, the student population is less susceptible to attrition. It was hoped that these subtle changes between the two courses would be reflected in the interval set clustering. These results were also compared with the third course (data structures) offered in the second year. This course consisted of core computing science students. The number of students in this course was around 23 students. It was hoped that the profile of visits would reflect some of the distinctions between the students. Lingras [8] and Lingras and West [11] showed that visits from students attending first course 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 class-notes on a regular basis.
2. Crammers: These visitors download a large set of notes. This indicates that they have stayed away from the class-notes for a long period of time. They are planning for pretest cramming.
3. Workers: These visitors are mostly working on class or lab assignments or accessing the discussion board.

The modified Kohonen algorithm was expected to specify the interval set clustering (lower and upper bounds for these classes).

4.2 Data Preparation

Data quality is one of the fundamental issues in data mining. Poor quality of data always leads to poor quality of results. Data preparation is an important step before applying data mining algorithms. The data preparation in this paper consisted of two phases: data cleaning and data transformation.

Data Cleaning involved removing hits from various search engines and other robots. This reduced the first data set by 5%. The second and third data sets were reduced by 3.5% and 10%, respectively. The details about the data can be found in Table 1.

The data transformation required the identification of web visits. 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, a lab manual, class-notes, 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 made sure 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.

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 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. were considered to be the daytime.
3. Access during lab/class days or non-lab/class days: All the labs and classes were held on Tuesdays and Thursdays. The visitors on these days are more likely to be workers.
4. Number of hits.
5. Number of class-notes downloads.

The first three attributes had binary values of 0 or 1. The last two values were normalized. The distribution of the number of hits and the number of class-notes was analyzed for determining appropriate weight factors. The numbers of hits were set to be in the range [0,10]. Since the class-notes were the focus of the clustering, the last variable was assigned higher importance, where the values ranged from 0 to 20. Even though the

weight for class-notes seems high, the study of actual distributions showed that 99% of visits had values of less than 5 for the first data set, less than 3 for the second data set, and less than 10 for the third data set.

Total visits were 23,754 for the first data set, 16,255 for the second data set, and 4,248 for the third data set. The visits that didn't download any class-notes were eliminated, since these visits correspond to either casual visitors or workers. The modified Kohonen clustering was applied to the remaining visits: 7,673 for the first data set, 6,056 for the second data set, and 1,287 for the third data set as shown in Table 1.

After experimenting with a range of values, the *threshold* was set at 0.7, $\alpha_{lower}(t)$ was chosen to be 0.01, 0.005 was used as the value of $\alpha_{upper}(t)$, and 1000 iterations were used for the training phase of each data set.

5. Results and Discussion

Table 3 shows the results for the first data set. Tables 4 and 5, show the results for the second and third data sets, respectively. It was possible to classify the three clusters as studious, workers, and crammers, from the results obtained using the modified Kohonen self-organizing maps. The crammers had the highest number of hits and class-notes in every data set. The average numbers of notes downloaded by crammers varied from one set to another. The significantly large number of class-notes downloaded by crammers in the third data set can be explained by further analysis. The third course had only 11 visitors in the crammers cluster. In addition, the distribution of notes downloaded in the third data set was more uniform than the two first year courses. As mentioned before, 99% of visitors in the first data set had values of less than 5, and less than 3 for the

second data set. However, the same number was 10 for the third data set. Because of a more uniform distribution, the number of class-notes was a good distinguishing attribute for the third data set. The studious visitors downloaded the second highest number of notes. The workers in the third data set downloaded the smallest number of notes. The distinctions between workers and studious visitors for the two first year courses were based on other attributes. For example, in the first data set, the on/off campus access was the most distinguishable attribute, followed by the lab day. Studious visitors exclusively came from off campus, while the workers exclusively came from campus locations. Workers were more prone to come on lab days than studious visitors. The distinguishable attributes for the second data set were again day and place of the visit. However, in contrast to the first data set, the day was the most distinguishable attribute. The workers exclusively came on lab days, and studious visitors always avoided the lab days. The workers were more prone to work from campus than the studious visitors. The profiles of upper bounds of workers and studious clusters were closer to each other than their lower bounds. It is interesting to note the similarity of the boundary regions of studious and workers for all the three data sets. The last two observations about the upper bounds and boundary regions suggest that there is a large overlap between upper bounds of studious and workers clusters. Fig. 6 gives a complete picture of the memberships from the interval clustering for the three data sets. For all the three data sets, there is more overlap between the upper bounds of studious and workers clusters than any other pair. The actual numbers in each cluster vary based on the characteristics of each course. For example, the first term course had 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 course had significantly large number of workers than studious visitors. This seems counter-intuitive. However, it should be noted that the lower bounds of studious and crammers were significantly smaller than the workers. That means most visitors in the third data set (second year course) had more uniform profiles. Moreover, unlike the two first year courses, the second year course did not post the class-notes on the web. The notes downloaded by these students were usually sample programs that were essential during their laboratory work.

The experiments used exactly the same setup for all the three web sites. The characteristics of the first two sites were similar. The third web site was somewhat different in terms of the site contents, course size, and types of students. The results discussed in this section show many similarities between the interval set clustering for the three sites. The differences between the results can be easily explained based on further analysis of the web sites. It is interesting to see that the proposed adaptation of the Kohonen networks captured the subtle differences between the web sites into the resulting clustering schemes. The clustering process can be individually fine-tuned for each web site to obtain even more meaningful interval set clustering scheme.

6. Summary and Conclusions

This paper proposed an adaptation of the Kohonen self-organizing maps to develop interval clusters using rough set theory. The paper also described an experiment for clustering web users including data collection, data cleaning, data preparation and the clustering process. Web visitors for three courses were used in the experiments to test the feasibility of the proposed adaptation. 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 classes, the clusters were represented using interval sets. In order to develop interval clusters the Kohonen algorithm was modified based on the concept of lower and upper bounds, and tested with the three data sets. The experiments produced meaningful clustering of web visitors. The study of variables used for clustering made it possible to clearly identify the three clusters as studious, workers, and crammers. There were many similarities and a few differences between the characteristics of interval clusters for the three web sites. These similarities and differences indicate the ability of the proposed modification of Kohonen networks to incorporate subtle differences between the usages of different web sites.

Acknowledgment

The authors would like to thank Natural Sciences and Engineering Research Council of Canada, and Government of Egypt for their financial support.

References

1. 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, 2000, 1-12.
2. A. Joshi and R. Krishnapuram, Robust Fuzzy Clustering Methods to Support Web Mining, in Proceedings of the workshop on Data Mining and Knowledge Discovery, SIGMOD '98, 1998, 15/1-15/8.
3. R. Hathaway and J. Bezdek, Switching Regression Models and Fuzzy Clustering, IEEE Transactions of Fuzzy Systems, **vol. 1**, no. 3, 1993, 195-204.

4. 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, 1995, 29-60.
5. R. Krishnapuram and J. Keller, A Possibilistic Approach to Clustering, IEEE Transactions, **vol. 1**, no.2, 1993, 98-110.
6. Z. Pawlak, Rough Sets, International Journal of Information and Computer Sciences, **vol. 11**, 1982, 145-172.
7. Z. Pawlak, Rough Classification, International Journal of Man-Machine Studies, **vol. 20**, 1984, 469-483.
8. P. Lingras, Unsupervised Rough Set Classification Using Gas, Journal of Intelligent Information Systems, **vol. 16**, no. 3, 2001, 215-228.
9. P. Lingras, Rough Set Clustering For Web Mining, in Proceedings of 2002 IEEE International Conference on Fuzzy Systems, 2002.
10. P. Lingras and X. Huang, Statistical, Evolutionary, and Neurocomputing Clustering Techniques: Cluster-Based Versus Object-Based Approaches, submitted to the Artificial Intelligence Review, 2002.
11. P. Lingras, and C. West, Interval Set Clustering of Web Users with Rough K-means, submitted to Journal of Intelligent Information Systems, 2002.
12. Y.Y. Yao, X. Li, T.Y. Lin and Q. Liu, Representation and Classification of Rough Set Models, in Proceeding of Third International Workshop on Rough Sets and Soft Computing, 1994, 630-637.

13. 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, 1999, 357-365.
14. T. Kohonen, Self-Organization and Associative Memory, Springer Verlag, Berlin, 1988.

List of Tables

Data Set	Hits	Hits after cleaning	Visits	Visits after cleaning
First course	361609	343000	23754	7673
Second course	265365	256012	16255	6056
Third course	40152	36005	4248	1287

TABLE 1. Description of the data sets

Group Name	Campus	Time	Lab	Hits	Req.	Cardinality
$\underline{A}$ (Studious)	0.000	0.596	0.224	0.379	0.408	1704
$\overline{A}$ (Studious)	0.475	0.680	0.406	0.530	0.489	4542
BND (Studious)	0.760	0.731	0.515	0.621	0.539	2838
$\underline{A}$ (Worker)	1.000	0.862	0.594	0.872	0.921	2633
$\overline{A}$ (Worker)	0.876	0.792	0.551	0.758	0.757	5566
BND (Worker)	0.764	0.730	0.512	0.655	0.609	2933
$\underline{A}$ (Crammers)	0.598	0.732	0.305	2.109	5.030	403
$\overline{A}$ (Crammers)	0.595	0.712	0.329	2.106	4.306	563
BND (Crammers)	0.588	0.663	0.388	2.098	2.482	160

TABLE 2. Results of interval clustering for the first data set

Group Name	Campus	Time	Lab	Hits	Req.	Cardinality
$\underline{A}$ (Studious)	0.62	0.73	0.00	0.33	0.32	3490
$\overline{A}$ (Studious)	0.62	0.72	0.02	0.38	0.38	3662
BND (Studious)	0.55	0.66	0.34	1.42	1.63	172
$\underline{A}$ (Worker)	0.78	0.84	1.00	0.33	0.29	2317
$\overline{A}$ (Worker)	0.77	0.83	0.96	0.40	0.38	2489
BND (Worker)	0.55	0.67	0.35	1.41	1.63	172
$\underline{A}$ (Crammers)	0.53	0.69	0.27	2.32	4.86	75
$\overline{A}$ (Crammers)	0.57	0.69	0.32	2.07	4.04	111
BND (Crammers)	0.64	0.69	0.42	1.54	2.32	36

TABLE 3 Results of interval clustering for the second data set

Group Name	Campus	Time	Lab	Hits	Req.	Cardinality
$\underline{A}$ (Studious)	0.54	0.74	0.44	2.45	4.14	143
$\overline{A}$ (Studious)	0.54	0.70	0.42	2.40	4.84	182
BND (Studious)	0.57	0.55	0.45	2.24	2.75	39
$\underline{A}$ (Worker)	0.54	0.75	0.51	0.90	0.58	1094
$\overline{A}$ (Worker)	0.54	0.74	0.51	0.94	0.74	1130
BND (Worker)	0.54	0.58	0.49	2.09	2.18	36
$\underline{A}$ (Crammers)	0.27	0.45	0.27	7.05	15.52	11
$\overline{A}$ (Crammers)	0.44	0.44	0.21	5.48	14.24	14
BND (Crammers)	1.00	0.44	0.00	4.88	9.52	3

TABLE 4. Results of interval clustering for the third data set

List of Figures

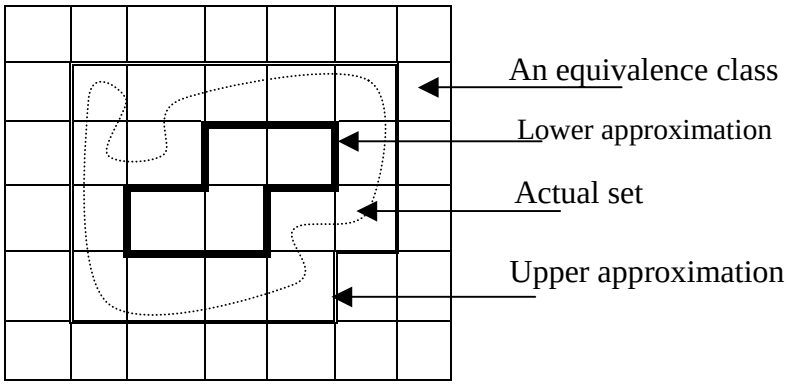


Fig.1 Rough Set Approximation

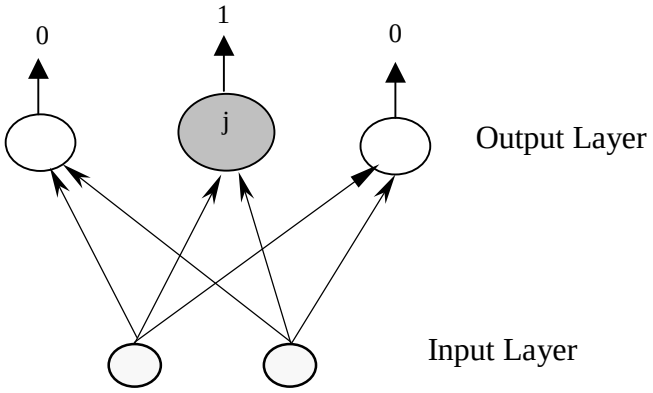


Fig. 2. Kohonen Neural Network

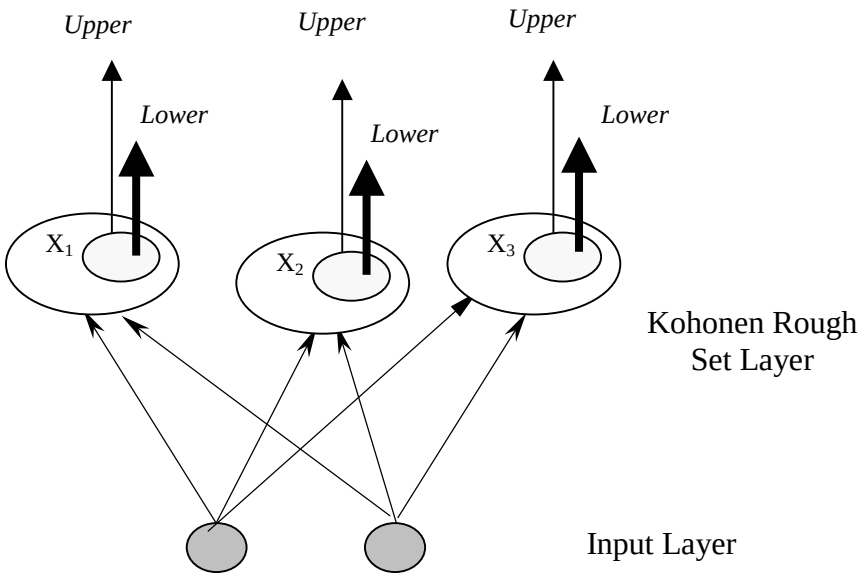


Fig. 3. Modified Kohonen Neural Network based on rough set theory

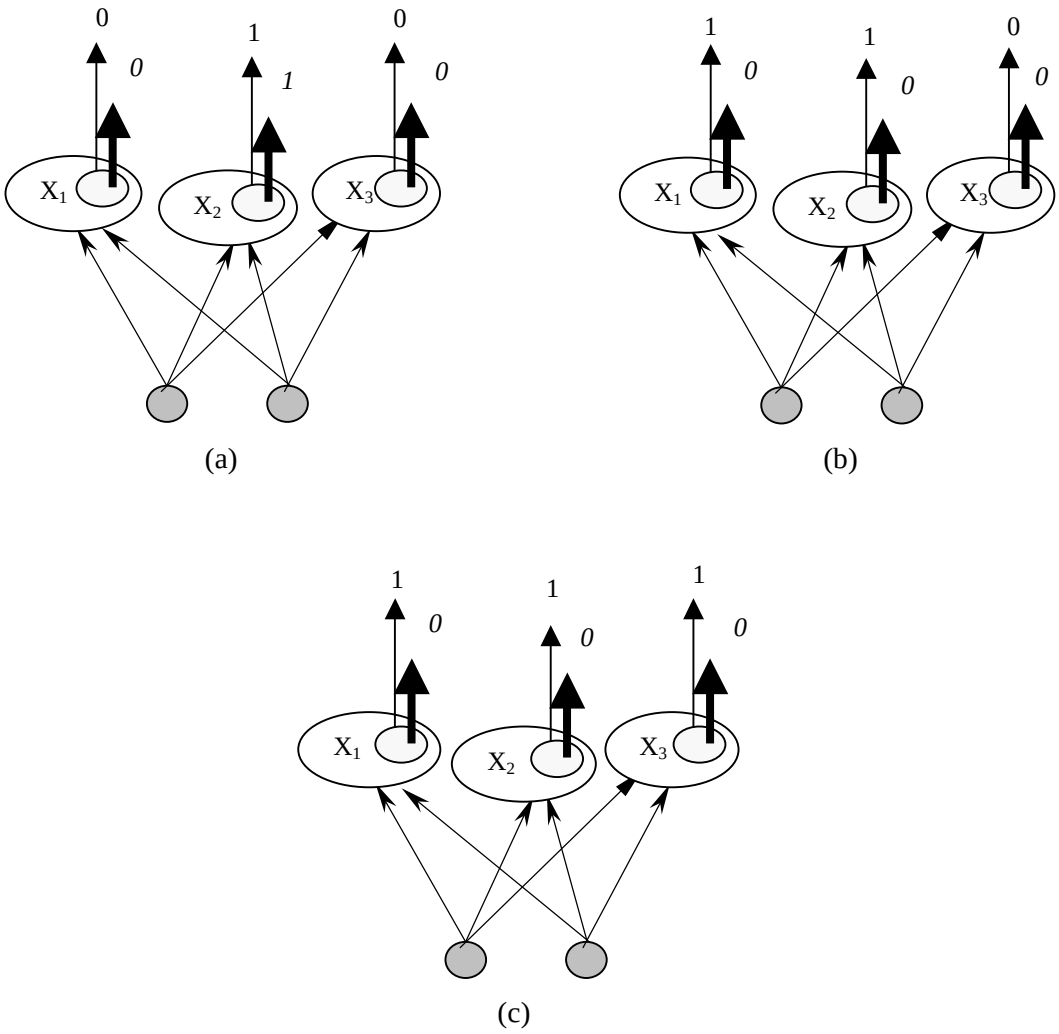


Fig. 4. Valid outputs from a Kohonen rough set layer

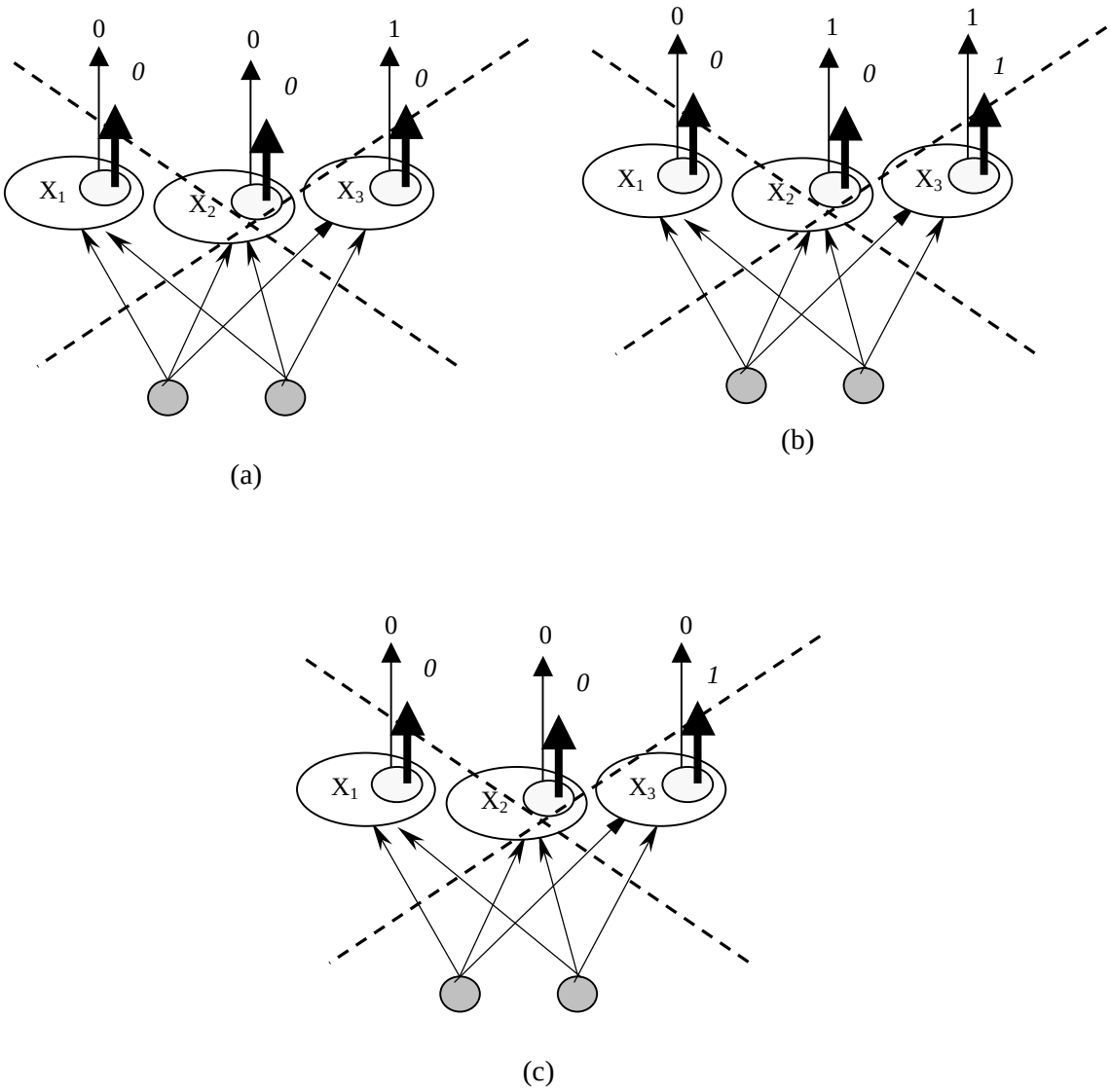
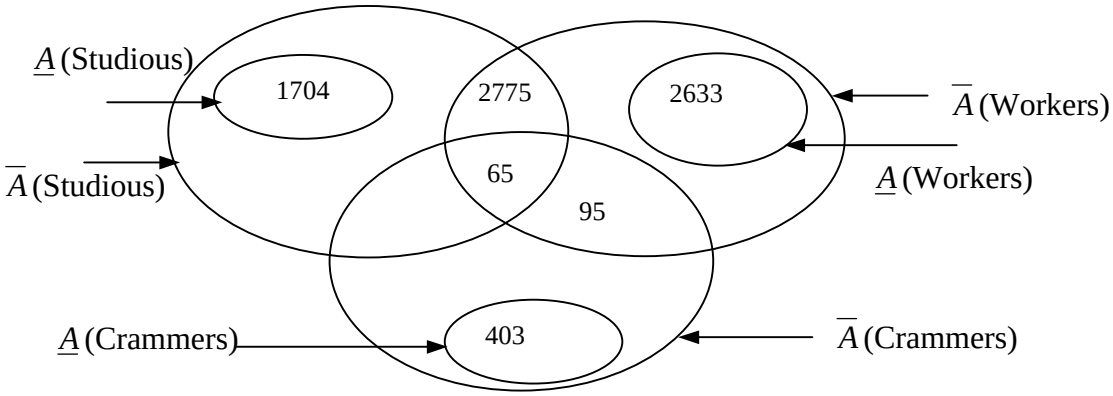
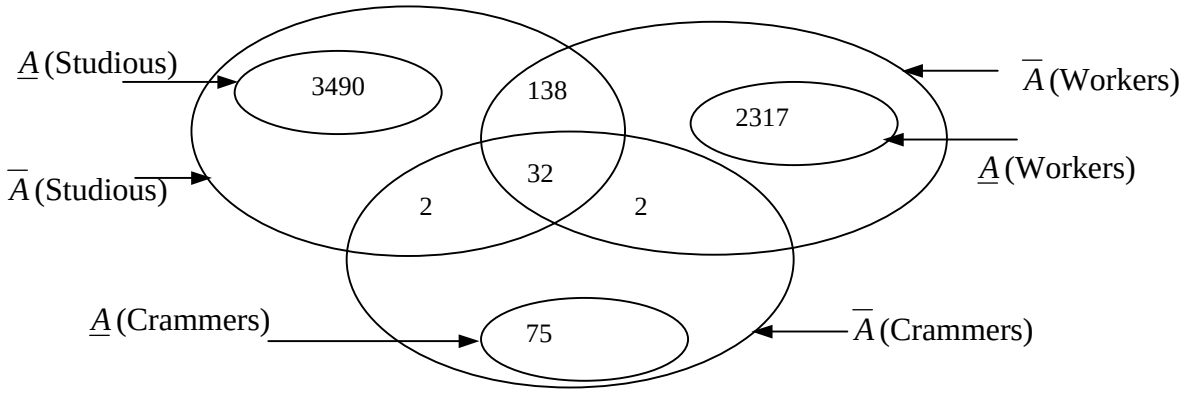


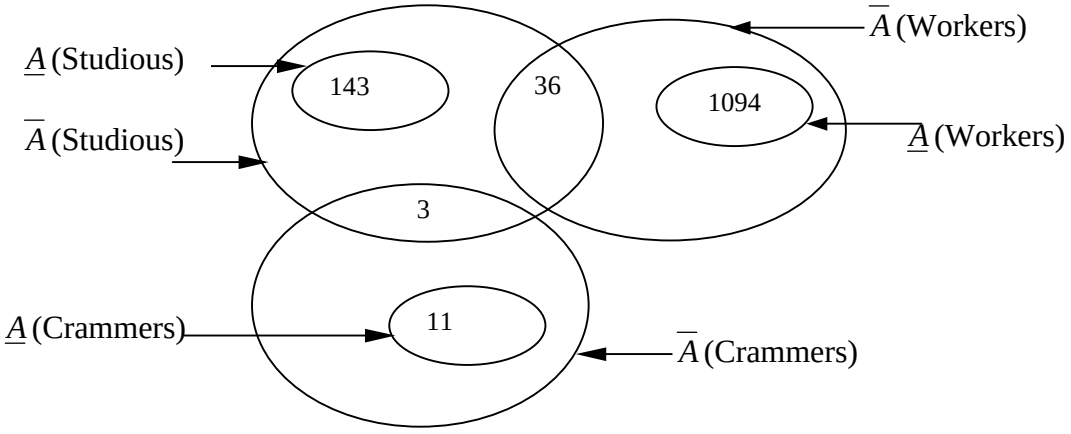
Fig. 5. Invalid outputs from a Kohonen rough set layer



First data set



Second data set



Third data set

Fig.6. Memberships in the interval sets of clusters