Fuzzy C-Means Clustering of Web Users for Educational Sites

Pawan Lingras, Rui Yan, and Chad West

Department of Mathematics and Computing Science
Saint Mary's University, Halifax, Nova Scotia, Canada, B3H 3C3

Abstract. Characterization of users is an important issue in the design and maintenance of websites. Analysis of the data from the World Wide Web faces certain challenges that are not commonly observed in conventional data analysis. The likelihood of bad or incomplete web usage data is higher than in conventional applications. The clusters and associations in web mining do not necessarily have crisp boundaries. Researchers have studied the possibility of using fuzzy sets for clustering of web resources. This paper presents clustering using a fuzzy c-means algorithm, on secondary data consisting of access logs from the World Wide Web. This type of analysis is called web usage mining, which involves applying data mining techniques to discover usage patterns from web data. The fuzzy c-means clustering was applied to the web visitors to three educational websites. The analysis shows the ability of the fuzzy c-means clustering to distinguish different user characteristics of these sites.

Keywords: Fuzzy C-means, Clustering, Web Usage mining, Unsupervised Learning.

1 Introduction

Clustering analysis is an important function in web usage mining, which groups together users or data items with similar characteristics. Clusters tend to have fuzzy or rough boundaries. Joshi and Krishnapuram [1] argued that the clustering operation in web mining involves modeling an unknown number of overlapping sets. They used fuzzy clustering to cluster web documents. Lingras [4] 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 clusters will not be precise. The present study applies the concept of fuzzy c-means [2,3] to the three educational websites analyzed earlier by Lingras *et al.* [6]. The resulting fuzzy clusters also provide a reasonable representation of user behaviours for the three websites.

Y. Xiang and B. Chaib-draa (Eds.): AI 2003, LNAI 2671, pp. 557-562, 2003.

2 Fuzzy C-Means

Cannon *et al.* [2] described an efficient implementation of an unsupervised clustering mechanism that generates the fuzzy membership of objects to 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 pattern vectors (attributes). Each cluster is representated by the cluster center vector *V*. The FCM algorithm minimizes the weighted within group sum of the squared error objective function *J(U,V)*:

$$J(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^{m} d_{ik}^{2}. \qquad \sum_{k=1}^{c} u_{ik} = 1. \quad 0 < \sum_{i=1}^{n} u_{ik} < n. \tag{1}$$

where U represent*s* the membership function matrix; u_{ik} is the elements of U ($u_{ik} \in [0,1]$, $i = 1,...n$, $k = 1,...c.$); V is the cluster center vector, $V = \{v_1, v_2, ...v_c\}$; n is the number of pattern; *c* is the number of clusters; d_{ik} represents the distance between x_i and v_k; *m* is the exponent of u_{ik} that controls fuzziness or amount of cluster overlap. Gao *et al.* [7] suggested the use of $m = 2$ in the experiments. The FCM algorithm is as follows :

Step 1: Given the cluster number *c*, randomly choose the initial cluster center V^0. Set $m = 2$, *s*, the index of the calculations, as 0, and the threshold ε, as a small positive constant.

Step 2: Based on *V*, the membership of each object U^s is calculated as:

$$u_{ik} = 1 \bigg/ \sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}}\right)^{\frac{2}{(m-1)}}, \ i = 1,...n, \ k = 1,...c. \ d_{ik} = \|x_k - v_i\| > 0, \forall i, k. \tag{2}$$

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

Step 3: Increment *s* by one. Calculate the new cluster center vector V^s as :

$$v_i = \sum_{k=1}^{n} (u_{ik})^m x_k \bigg/ \sum_{k=1}^{n} (u_{ik})^m, \ \forall i, \ i = 1,...n. \tag{3}$$

Step 4: Compute the new membership U^s using the equation (2) in step 2.

Step 5: If $\left\| U^s - U^{(s-1)} \right\| < \varepsilon$, then stop, otherwise repeat step 3, 4, and 5.

3 Study Data and Design of the Experiment

The study data was obtained from web access logs of three courses. These courses represent a sequence of required courses for the computing science programme at Saint Mary's University. The first and second courses were for first year students. The third course was for second year students. Lingras [4] and Lingras and West [5] showed that visits from students attending the 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 fuzzy c-means algorithm was expected to provide the membership of each visitor to the three clusters mentioned above. Data cleaning involved removing hits from various search engines and other robots. Some of the outliers with 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. The details about the data can be found in Table 1. Five attributes are used for representing each visitor [4]:

1. On campus/Off campus access. (Binary value)
2. Day time/Night time access: 8 a.m. to 8 p.m. were considered to be the daytime. (Binary value)
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. (Binary value)
4. Number of hits. (Normalized in the range [0,10])
5. Number of class-notes downloads. (Normalized in the range [0,20])

4 Results and Discussion

Table 2 shows the fuzzy center vectors for the three data sets. It was possible to classify the three clusters as studious, workers, and crammers, from the results obtained using the fuzzy c-means clustering. 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 studious visitors downloaded the second highest number of notes. The distinction between workers and studious visitors for the second course was based on other attributes. It is also interesting to note that the crammers had a higher ratio of document requests to hits. The workers, on the other hand, had the lowest ratio of document requests to hits.

Table 1. Description of the Data Sets

Data Set	Hits	Hits after cleaning	Visits	Visits after cleaning
First	361609	343000	23754	7619
Second	265365	256012	16255	6048
Third	40152	36005	4248	1274

Table 2. Fuzzy 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	0.59	0.74	0.15	0.68	0.57
	Crammers	0.63	0.73	0.33	2.34	3.07
	Workers	0.82	0.86	0.71	0.64	0.49
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. Visitors with Fuzzy Memberships Greater than 0.6

Course	Cluster Name	Number of Visitors with Memberships > 0.6
First	Studious	1382
	Crammers	414
	Workers	4354
Second	Studious	1419
	Crammers	317
	Workers	1360
Third	Studious	265
	Crammers	84
	Workers	717

Table 3 shows the cardinalities of sets with fuzzy memberships greater than 0.6. The choice of 0.6 is somewhat arbitrary. However, a membership of 0.6 (or above) for a cluster indicates a stronger tendency towards the cluster. 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. Similarly, the third course had significantly more studious visitors than workers. Crammers constituted less than 10% of the visitors.

The characteristics of the first two sites 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 show many similarities between the fuzzy c-means clustering for the three sites. The differences between the results can be easily explained based on further analysis of the websites. It is interesting to see that the fuzzy c-means clustering captured the subtle differences between the websites in the resulting clustering schemes. The clustering process can be individually fine-tuned for each website to obtain even more meaningful clustering schemes.

5 Summary and Conclusions

This paper described an experiment for clustering web users, including data collection, data cleaning, data preparation, and the fuzzy c-means clustering process. Web visitors for three courses 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 classes, the clusters were represented using fuzzy membership functions. 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 clusters for the three websites. These similarities and differences indicate the ability of the fuzzy c-means clustering to incorporate subtle differences between the usages of different websites.

Acknowledgment

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

References

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

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

[3] 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 (1995) 2289-2295.

[4] P. Lingras: Rough Set Clustering for Web Mining. In the Proceedings of 2002 IEEE International Conference on Fuzzy Systems (2002).

[5] Lingras, and C. West: Interval Set Clustering of Web Users with Rough K-means. Submitted to Journal of Intelligent Information Systems (2002).

[6] 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).

[7] 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 (2000) 1457-1461.