Efficient Clustering Technique for University Admission Data

Abdul Fattah Mashat
Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Saudi Arabia

Mohammed M. Fouad
Faculty of Informatics and Computer Science
The British University in Egypt
Cairo, Egypt

Philip S. Yu
University of Illinois, Chicago, IL, USA
King Abdulaziz University
Jeddah, Saudi Arabia

Tarek F. Gharib
Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Saudi Arabia

ABSTRACT

Educational Data Mining (EDM) is the process of converting raw data from educational systems to useful information that can be used by educational software developers, students, teachers, parents, and other educational researchers. In this paper, we present an efficient clustering technique for King Abdulaziz University (KAU) admission data. The model uses K-Means algorithm. The clustering quality is evaluated using the DB internal measure. Experimental results show that K-Means achieves the minimum DB value that gives the best fits natural partitions. Additional analysis is also presented from the perspective of university admission office.

Keywords

Educational Data Mining (EDM); Data Clustering; University Admission Data; Clustering Evaluation.

1. INTRODUCTION

Data mining aims at the discovery of useful information from large collection of data. Recently, there are increasing research interests in using data mining in education. This new emerging field, called Educational Data Mining (EDM), concerns with developing methods that discover knowledge from data originating from educational environments [1].

Educational data mining (EDM) differs from knowledge discovery in other domains in several ways. One of them is the fact that it is difficult, or even impossible, to compare different methods or measures a posteriori and decide which is the best. Take the example of building a system to transform hand-written documents into printed documents. This system has to discover the printed letters behind the hand-written ones. It is possible to try several sets of measures or parameters and experiment what works best. Such an experimentation phase is difficult in the education field because the data is very dynamic, can vary a lot between samples and teachers just cannot afford the time and access to the expertise to do these tests on each sample, especially in real time. Therefore, one should care about the intuition of the measures, parameters or methods used in educational data mining [2].

Clustering algorithms attempt to organize unlabeled input vectors into clusters or “natural groups” such that points within a cluster are more similar to each other than vectors belonging to different clusters [3]. Clustering has been used in exploratory pattern-analysis, grouping, decision-making, and machine-learning situations, which include data mining, document retrieval, image segmentation, and pattern classification. The clustering methods are of five types: hierarchical clustering, partitioning clustering, density-based clustering, grid-based clustering, and model based clustering [4]. Each type has its advantages and disadvantages. Verma et al. Provided a comparative study of commonly used clustering algorithms in data mining field [5]. Their study compared the performance of six types of clustering techniques: K-Means, Hierarchical, DBScan, Density-based, OPTICS and EM algorithms. Their experimental results showed that K-Means were faster and achieved good clustering results than other algorithms but it was sensitive to noise (if exists).

In this paper, we present an efficient clustering model for King Abdulaziz University (KAU) admission data. The model uses K-Means algorithm and DB measure as internal clustering quality evaluation index.

The rest of this paper is organized into five sections. In section 2, the clustering model and algorithms are briefly reviewed. Section 3 presents the KAU admission system as a case study. In section 4, experimental results are presented and analyzed with respect to model results and admission system perspective. Finally, the conclusions of this work are presented in Section 5.

2. CLUSTERING MODEL

Clustering is one solution to the case of unsupervised learning, where class labeling information of the data is not available. Clustering is a method where data is divided into groups (clusters) which ‘seem’ to make sense. Clustering algorithms are usually fast and quite simple. They need no prior knowledge of the input data and form a solution by comparing the given samples to each other and to the clustering criterion. In this section we will discuss three popular clustering algorithms that are used in this paper.

2.1 K-Means Clustering

K-Means is one of the most commonly used clustering algorithms [6]. It uses squared Euclidean distance as a dissimilarity measure and tries to minimize within cluster distance and maximize between-cluster distance. For a given number of clusters K, k-Means searches for cluster centers C_i and assignment S that minimize the criterion shown in Eq. (1).

$$\min_S \sum_{a=1}^{K} \sum_{S(i)=a} \left\| x_i - C_a \right\|^2 \qquad (1)$$

The algorithm alternates between optimizing the cluster centers for the current assignment (by the current cluster means) and optimizing the cluster assignment for a given set of cluster centers (by assigning to the closest current center) until convergence (i.e. Cluster assignments do not change). It tends to find compact, spherical clusters and requires a priori the number of clusters K. The final cluster assignment can be sensitive to the choice of centers; a common method for initializing k-Means is to randomly choose K data points as initial centers.

2.2 Self-Organizing Maps (SOM)

The SOM consists of M neurons organized, usually, in a two dimensional grid. The SOM network uses a competitive learning process in which the neurons in a neural network gradually become sensitive to different input categories, sets of samples in a specific domain of the input space. A division of neural nodes emerges in the network to represent different patterns of the inputs after training [7].

The basic SOM algorithm is iterative. Each neuron has a feature vector $w_i=[w_{i1},\ldots,w_{id}]$ with d-dimension. At each training step t, a sample data vector x(t) is randomly chosen from the training set. Distances between x(t) and all the feature vectors are computed. The winning neuron, denoted by c, is the neuron with the feature vector closest to x(t):

$$c = \arg\min_i \left\| x(t) - w_i \right\|, i \in \{1,\ldots,M\} \qquad (2)$$

After completion of training, each neuron is attached to a feature vector of the same dimension as the input space. By assigning each input vector to the neuron with the nearest feature vector, the SOM is able to divide the input space into regions with common nearest feature vectors.

2.3 Fuzzy Clustering

Fuzzy c-means is a data clustering technique wherein each data point belongs to a cluster to some degree that is specified by a membership measure while other clustering algorithms assign each data point to exactly one cluster. This technique was originally introduced by Bezdec [8] as an improvement on earlier clustering methods. It provides a method that shows how to group data points that populate some multidimensional space into a specific number of different clusters. It is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} U_{ij}^m \left\| x_i - c_j \right\|^2 \ , \ m \geq 1 \qquad (3)$$

where *m* is any real number greater than 1, U_{ij} is the degree of membership of x_i in the cluster *j*, x_i is the *i*th of d-dimensional measured data, c_j is the center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center which is Euclidean distance in our implementation.

Fuzzy partitioning is carried out through an iterative optimization of the objective function shown in Eq. (3), with the update of membership U_{ij} and the cluster centers c_j as in Eq. (4).

$$U_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|} \right)^{\frac{2}{m-1}}} , \ c_j = \frac{\sum_{i=1}^{N} U_{ij}^m x_i}{\sum_{i=1}^{N} U_{ij}^m} \qquad (4)$$

2.4 Cluster Validation Index (CVI)

The problem of detecting clusters of points in data is challenging when the clusters are of different size, density and shape. Many of these issues become even more significant when the data is of very high dimensionality and when it includes noise and outliers. Cluster validation is a technique for finding a set of clusters that best fits natural partitions (of given datasets) without the benefit of any a priori class information. A good clustering algorithm will have small intra-cluster distances and large inter-cluster distances [9].

Most CVIs are usually defined by combining the following pair of evaluation criteria [10]:

1. ***Compactness***: This measure shows how close the data points in the same cluster. Variance is one example in this criterion that shows how the cluster members are different from cluster center (mean). The lower value of variance indicates better compactness.
2. ***Separability***: This measure computes the distance between adjacent clusters to show how distinct they are. This distance can be computed between the centers of the two clusters. If the clusters are two adjacent they can be merged into one cluster if its compactness is high.

In this paper, we use the Davies-Bouldin's (DB) validity index proposed by Davis et al., [11]. DB index is the ratio of cluster scatter S_i of cluster D_i to cluster separation. For a given dataset that is clustered into n clusters ($D_1,\ldots D_n$) defined by n centers ($c_1,\ldots,c_n$), DB index is calculated as in Eq. (5).

$$\mathrm{DB}(n) = \frac{1}{n} \sum_{i=1}^{n} \max_{k,k \neq i} \left(\frac{S_i + S_k}{d(c_i, c_k)} \right)$$
$$\text{where } S_i = \frac{1}{|D_i|} \sum_{x \in D_i} d(x, c_i) \qquad (5)$$

An individual cluster index is taken as the maximum pairwise comparison computed as the ratio of the sum of within cluster dispersions from the two partitions divided by a measure of the between cluster separation. Smaller values of the DB index correspond to good clusters. The number of clusters that minimizes DB index is the optimal number of clusters [12, 13].

3. KAU ADMISSION SYSTEM – CASE STUDY

King Abdulaziz University (KAU) admission system in the Kingdom of Saudi Arabia (KSA) is a complex decision process that goes beyond simply matching test scores and admission requirements because of many reasons. First, the university has many branches in KSA for both division male and female students. Second, the number of applicants in each year is a huge which needs a complex selection criterion that depends on high school grades and applicant region/city.

In this paper, we used some statistical datasets about the admission rate for different regions/cities represented the King Abdulaziz University (KAU) admission system. The present dataset contains about **125** records for all the regions/cities that students apply from. Each record contains 6 attributes, which are MaleRate and FemaleRate for different three academic years 2010, 2011 and 2012. The rate attributes (Rate_{ij}) for the region (i) and academic year (j) is calculated using the formula:

$$Rate_{i\,j} = \frac{A_{ij}}{\sum_k A_{kj}} \qquad (6)$$

Where A_{ij} is the number of accepted students in the region (i) and academic year (j).

4. EXPERIMENTAL RESULT

4.1 Clustering Results

Table 1 shows the DB Index for each of the clustering algorithms along with different numbers of clusters (from 3 to 7 clusters).

Table 1. DB Index for Clustering Quality

N	*3*	*4*	*5*	*6*	*7*
K-Means	**0.3845**	0.6121	0.5117	0.5212	0.6681
SOM	6.1259	3.2186	1.4740	1.8459	1.1413
Fuzzy	0.4258	0.5648	0.5211	0.5315	0.6718

As shown in table 1, K-Means overcomes other algorithms and achieves best clustering results. Because of k-means to achieve the minimum DB value comparing with Fuzzy C-Means and the SOM algorithms. The DB validity index showed that the best number of clusters is 3.

To visualize K-Means clustering results more clearly, we calculate the average male and female rate from input datasets to show output clusters in 2D space. As shown in fig. 1, the data are more compact and form some density and logical meaning. The center of each cluster is donated by black circles.

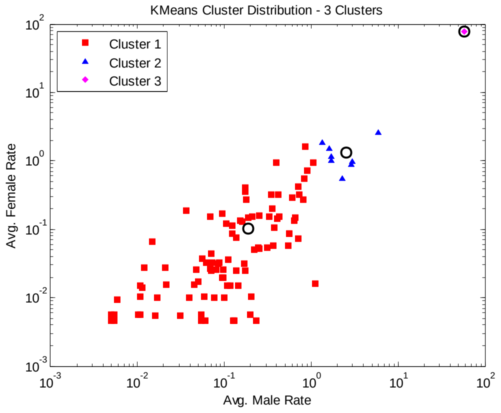

Fig. 1: KMeans Cluster Distribution (3 Clusters)

4.2 Analysis and Discussion

From a university admission system perspective, we can conclude some features for clustering results shown in fig. 1 as following:

- *Cluster 1*: **low** acceptance rate for both males and females (*low acceptance rate*). The regions of this cluster are considered below average source of students which means that their students are not fulfill the requirements of KAU because their average rates are **0.12%** males and **0.10%** females. (**116 regions)**
- *Cluster 2*: **average** male rate and **average** female rate (*average acceptance rate*). The regions here can be average source for female and male students. This group has about **8 regions** that supply the university with average **2.5%** and **1.3%** for male and female rates respectively.
- *Cluster 3*: **high** male rate and **high** female rate (*high acceptance rate*). The regions of this cluster are considered a good student source for both genders. This cluster contains only **1 region** (which is Jeddah) that has about **58%** male rate and **78%** female rate. This means that Jeddah city/region is the source of the most suitable among the many thousands of students that apply to KAU every year.

Figure 2 shows the population number distributions of cities/regions which consider the main source of students apply in KAU. It may be one of the reasons that explain why Jeddah city represented as one cluster with high rate for both Male and Female.

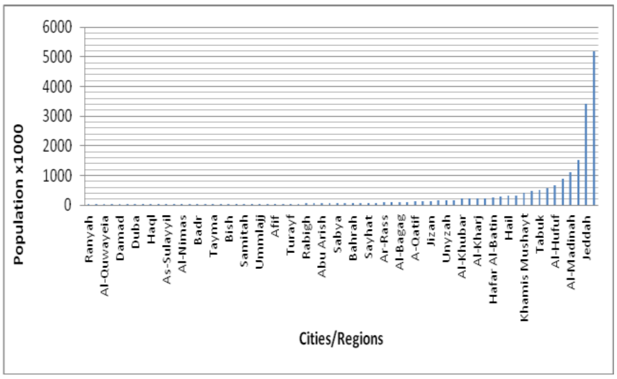

Fig. 2: The Cities/Regions Population Distribution

Table2 shows that there are about **67%** of the applicants are males with acceptance rate **13.5%** while they represent about **49%** of the accepted applicants. On the other side, percentages of accepted females are about **51%** with only **33%** from the initial applications with acceptance rate **27.7%**. This means that female students in the high schools fulfill the university admission requirements. Also, we conclude from table 2 that the university admission needs more studies based on the perspective of secondary schools that sending their students to pursue higher education not only based on the perspective of university that receive the new incoming students.

Table 2. Admission Statistics data

Year	*Male Students*			*Female Students*		
	#Apps	*#Acc*	*%*	*#Apps*	*#Acc*	*%*
2010	44301	6268	**14**	17755	5907	**33**
2011	46718	5637	**12**	23751	6038	**25**
2012	47474	6702	**14**	27221	7141	**26**
Totals	*138493*	*18607*	***13***	*68727*	*19086*	***27***
	67%	*49%*		*33%*	*51%*	

5. CONCLUSION

In this paper we presented an efficient clustering model for KAU university admission data. The model used DB internal clustering validity index to measure the performance of three different clustering algorithms. Also, we noticed that K-Means performance overcomes both Fuzzy C-Means and SOM algorithms. The clustering results also provide some information that will be helpful in the KAU admission office based on the data attributes. This information shows the accepted applicants' rate (for both males and females) with respect to the region or city they come from. KAU admission office can use this information in adopting some advertisement strategies in the regions with very low rate. Other regions are considered a good student source in which the university can offer some scholarships to its students.

6. REFERENCES

[1] S. Feng, S. Zhou and Y. Liu, (2011) "Research on Data Mining in University Admissions Decision-making", International Journal of Advancements Advancements in Computing Technology, vol. 3, no. 7, pp. 176-186.

[2] J. Beck, (2007) "Difficulties in inferring student knowledge from observations (and why you should care)", Educational Data Mining workshop in conjunction with 13th International Conference of Artificial Intelligence in Education, Marina del Rey, CA. USA, pp.21-30.

[3] N.R. Pal, J.C. Bezdek and E.C.-K. Tsao, (1993) "Generalized clustering networks and Kohonen's self-organizing scheme", IEEE Trans. Neural Networks, vol. 4, no. 4, pp. 549-557.

[4] J. Han and M. Kamber, (2000), Data mining:concepts and techniques, San Francisco, Morgan-Kaufma.

[5] M. Verma, M. Srivastava, N. Chack, A. K. Diswar, N. Gupta, (2012) "A Comparative Study of Various Clustering Algorithms in Data Mining", International Journal of Engineering Research and Applications (IJERA), vol. 2, no. 3, pp. 1379-1384.

[6] J. Hartigan and M.A. Wong, (1979) "A k-means clustering algorithm", Applied Statistics, vol. 28, pp. 100-108.

[7] S. Wu and T. Chow, (2004) "Clustering of the self-organizing map using a clustering cvalidity index based on inter-cluster and intra-cluster density", Pattern Recognition, vol. 37, pp. 175-188.

[8] J.C. Bezdec, (1971), Pattern Recognition with Fuzzy Objective Function Algorithms, New York, Plenum Press.

[9] L.J. Deborah, R. Baskaran and A. Kannan, (2010) "A survey on Internal Validity Measure for Cluster Validation", International Journal of Computer Science & Engineering Surveys (IJCSES), vol. 1, no. 2, pp. 85-102.

[10] M.J.A. Berry and G. Linoff, (1997), Data Mining Techniques: For Marketing, Sales, and Customer Support, Berlin, John Wiley & Sons.

[11] D. Davies and D. Bouldin, (1979) "A Cluster Separation Measure", IEEE transactions on Pattern Analysis and Machine Intelligence, vol. 1, no. 2, pp. 2244-227.

[12] M. Kim and R.S. Ramakrishna, (2005) "New indices for cluster validity assessment", Pattern Recogntion Letters, vol. 26, pp. 2353-2363.

[13] K.R. Zalik and B, Zalik, (2011) "Validity index for clusters of different sizes and densities", Pattern Recognition Letters, 32, pp. 221-234.