JOURNAL OF AERONAUTICS AND SPACE TECHNOLOGIES
JULY 2005 VOLUME 2 NUMBER 2 (53-57)

A DATA MINING APPLICATION IN A STUDENT DATABASE

Şenol Zafer ERDOĞAN
Maltepe Üniversity
Faculty of Engineering
Büyükbakkalköy-Istanbul
senole@maltepe.edu.tr

Mehpare TİMOR
İstanbul Üniversity,
Faculty of Business Administration
Avcılar-Istanbul
timorm@istanbul.edu.tr

ABSTRACT
Data mining is a technology used in different disciplines to search for significant relationships among variables in large data sets. Data mining is mainly used in commercial applications. In this study, we concentrated on the application of data mining in an education environment. The relationship between students university entrance examination results and their success was studied using cluster analysis and k-means algorithm techniques.

Keywords: *Data mining, Cluster Analysis, K-Means Algorithm.*

1. INTRODUCTION

The amount of data maintained in an electronic format has seen a dramatic increase in recent times. The amount of information doubles every 20 months, and the number of databases is increasing at an even greater rate [1,2]. The search to determine significant relationships among variables in the data has become a slow and subjective process. As a possible solution to this problem, the concept of *Knowledge Discovery in Databases – KDD* has emerged [3]. The process of the formation of significant models and assessment within KDD is referred to as data mining [2,4]. Data mining is used to uncover hidden or unknown information that is not apparent, but potentially useful [2,5].

Cluster analysis is a technique used in data mining. Cluster analysis involves the process of grouping objects with similar characteristics [6], and each group is referred to as a cluster. Cluster analysis is used in various fields, such as marketing, image processing, geographical information systems, biology, and genetics.

In this study, university students were grouped according to their characteristics, forming clusters. The clustering process was carried out using a K-means algorithm.

2. CLUSTER ANALYSIS

Cluster analysis is a multivariate analysis technique where individuals with similar characteristics are determined and classified (grouped) accordingly [2,7]. Through cluster analysis, dense and sparse region can be determined in the distribution, and different distribution patterns may be achieved.

The concepts of similarities and differences are used in cluster analysis. Different measures may be used in determining similarities and differences. This study utilises the Euclidian distance measure.

2.1. Euclidian Distance Measure

The Euclidian distance measure is frequently used as a distance measure, and is easy to use in two dimensional planes. As the number of dimensions increases, the calculability time also increases [2].

$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \ldots + |x_{i_p} - x_{j_p}|^2)} \qquad (2.1.a)$$

The formula defines data objects i and j with a number of dimension equal to p. The distance between the two data objects d(i,j) is expressed as given in formula (2.1.a). x_{ip}: is the measurement of object i in dimension p.

2.2. Algorithm

The K-means algorithm is a cluster analysis algorithm used as a partitioning method, and was developed by MacQueen in 1967 [8]. K-means is the most widely used used and studied clustering algorithm. Given a set of n data points in real d-dimensional space, $\mathbf{R}^d$, and an integer k, the problem is to determine a set of k points in $\mathbf{R}^d$, called centers, so as to minimize the mean squared distance from each data point to its nearest center[9].

The K-means algorithm defines a random cluster centroid according to the initial parameters [8]. Each consecutive case is added to the cluster according to the proximity between the mean value of the case and the cluster centroid. The clusters are then re-analysed to determine the new centroid point. This procedure is repeated for each data object.

The algorithm is composed of the following steps:[10]

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

3. APPLICATION

In this study, data gathered from university students was analysed using a k-means algorithm cluster analysis technique.

3.1. Data Set

The data gathered from the students of the Maltepe University was used in this study. The data was gathered in 2003, and included records of 722 students.

3.2. Database

The database management system used in the study was the Microsoft SQL Server 2000. This system was used for two reasons; the software used in analysis was compatible and efficient to use with the database management system, and the data to be analysed was maintained in the database prior to the study.

3.3. Application Software

The programming environment for the application was Matlab. The Matlab software application was suitable for the development of the application, and compatible with the SQL Server 2000 in which the data was maintained. The K-means algorithm used in the application was defined in the Matlab software as a function.

3.4. The Data Mining Process

The data exploration and presentation process consisted of various steps. These steps were data preparation, data selection and transformation, data mining and presentation.

3.4.1. Data Preparation

In these steps, the data that was maintained in different tables was joined in a single table. The ‘students’ and ‘students_log’ tables were joined using the StudentsID field as the key field. After the joining process errors in the data were corrected.

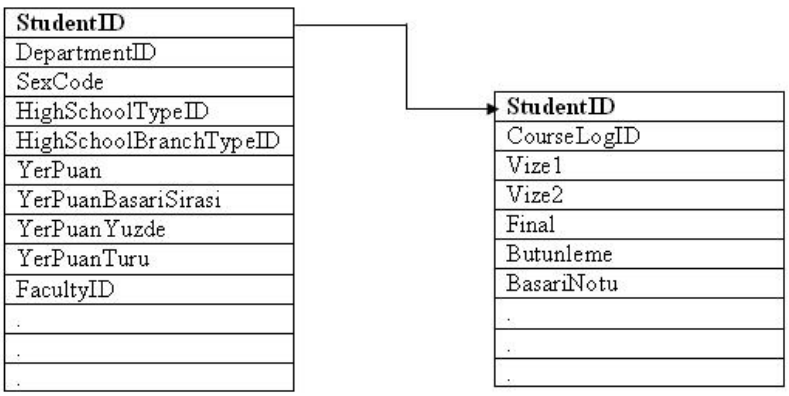

Figure 1. Students and Students Grades Tables

3.4.2. Data Selection and Transformation

After the data preparation, the data selection and transformation process was performed. In this step the fields used in the study were determined and transformed if necessary. For example, the fields in which the responses were yes/no were transformed to 1/0. The AreaPointPercent, SuccessGrade, SexCode, HighSchoolTypeID and FacultyID fields were selected for use in the study, and a new table was created. AreaPointPercent was the percentile the student fell into in the university entrance exam, and SuccessGrade was the grade they obtained. The SexCode variable was coded 5 and 10 for male and female respectively. The HighSchoolTypeID varied between 1 and 10 and the FacultyID varied between 1 and 6.

3.4.3. Data Mining

The prepared data was then put through the data mining process. The K-means algorithm was used in this step. The number of clusters was determined as an external parameter. Different cluster numbers were tried, and a successful partitioning was achieved with 5 clusters. The cluster centroids are given in table 1.

Table 1. Cluster Centroids

Cluster	AreaPointPercent	SuccessGrade	Gender	HighSchoolTypeID	FacultyID
1	12.4774	89.5350	8.1070	6.4650	2.5844
2	16.3113	59.0472	6.9811	5.1981	3.0189
3	46.7851	77.1240	6.9008	5.7190	3.2149
4	80.1095	78.0146	6.6788	3.2774	3.8321
5	79.3565	44.8870	6.3043	3.1391	4.1304

3.4.4. Presentation

The results of the data mining step are presented in this step. For graphs and tables, the MapToolBox plug-in of the Matlab software was used. The resulting clusters are shown in figure 2.

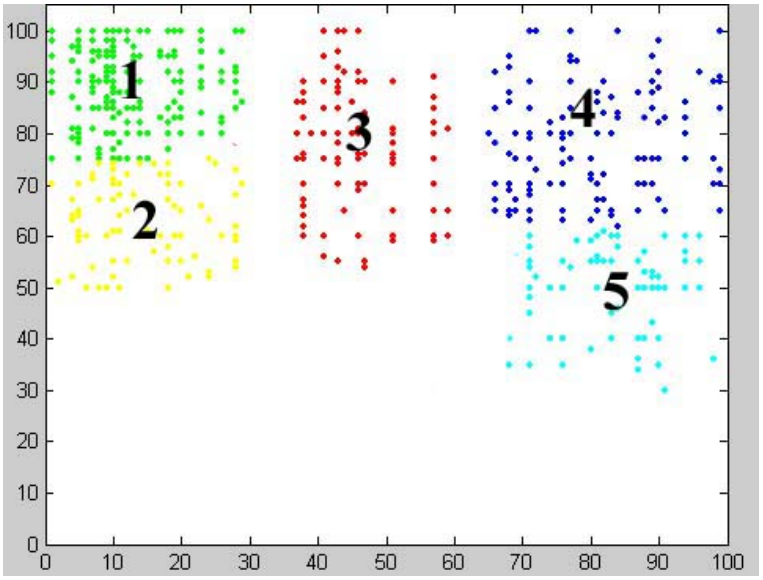

Figure 2. Each number represents different cluster

The figure above shows the University Entrance Exam percentiles of the x axis, and the grades on the y access. The graph shows that the 1st cluster is more successful in regard to grades while the 5th group is the least successful. The distribution of faculties in these two clusters is shown in figures 3 and 4.

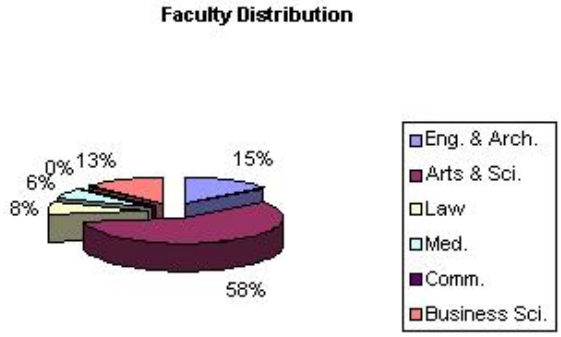

Figure 3. Distribution of Cluster 1

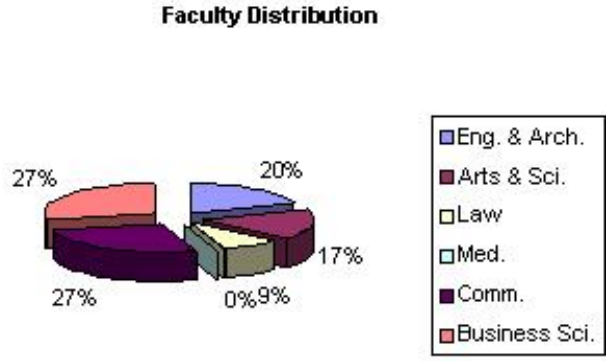

Figure 4. Distribution of Cluster 5

The majority of the students in cluster 1 are from the Faculty of Arts and Sciences. The reason is that most students in the faculty have high success grades and scholarships. They study hard to keep their scholarship, and therefore have good grades.

Cluster 5, however, is mainly made up of Faculty of Communication and Faculty of Business Sciences students. These students have lower grades and lower results in the university entrance exam.

4. CONCLUSION

This study utilises data mining in the field of education. Cluster analysis and K-means analysis were used as data mining techniques. The steps of the data mining process were carried out and explained in detail. The area of application was education, different from the usual data mining studies. The use of the data mining technique in education may provide us with more varied and significant findings, and may lead to the increase in the quality of education.

5. REFERENCES

[1] Vahaplar, A.,İnceoğlu, M., "Veri Madenciliği ve Elektronik Ticaret", Türkiye'de İnternet Konferansları, Harbiye İstanbul, 1-3 Kasım 2001.

[2] Erdoğan, Ş. Z., "Veri Madenciliği ve Veri Madenciliğinde Kullanılan K-Means Algoritmasının Öğrenci Veri Tabanında Uygulanması", Yüksek Lisans Tezi, İstanbul Üniversitesi, 2004.

[3] Akpınar, H., "Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği", İ.Ü. İşletme Fakültesi Dergisi, Sayı:1 (1-22), Nisan 2000.

[4] Thearling, K., "An Introduction to Data Mining",http://thearling.com/text/dmwhite/dmwhite.htm, 01 December 2003.

[5] Fayyad, U.M., Piatesky-Shapiro, G., Smyth, P., Uthurusamy, R., "Advances in data mining and knowledge discovery", MIT Pres, USA, 1994.

[6] Han, J., Kamber, W., "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, USA, 5-10, 2001.

[7] Menteş, G. T., "Faktör ve Kümeleme Analizi Yardımıyla Bankacılık Ürün ve Hizmetlerinin Araştırılması Üzerine Bir Uygulama", Doktora Tezi, İstanbul Üniversitesi, 2000.

[8] Yuqing, P., Xiangdan, H., Shang, L., "The K-Means Clustering Algorithm Based On Density and Ant Colony", IEEE Int. Conf. Neural Networks & Signal Processing Nanjing, China, 457-460, December 14-17, 2003.

[9] Kanungo, T., Mount, D., S. Netanyahu, N., Piatko D. C., Silverman, R., Wu, A.Y., "An Efficient k-Means Clustering Algorithm: Analysis and Implementation", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 7, July 2002.

[10] Luke, B. T., "K-Means Clustering", http://fconyx.ncifcrf.gov/~lukeb/kmeans.html, 20 October 2004.

VITA

Şenol Zafer ERDOĞAN

He graduated from Computer Engineering at Trakya University in 2001.He received his Msc degree in Istanbul University in july 2004. He joined Computer Engineering Department at Maltepe University in 2001. He is now a research assistant at Maltepe University.

Mehpare TİMOR

She graduated from Business Administration at Istanbul University in 1986. She received her Msc degree in 1988 and she received her Phd degree at Istanbul University in 1993. She is now a assistant professor at Istanbul University Faculty of Business Administration.