Using data mining techniques in higher education

Elena Şuşnea

(1) National Defence University "Carol I", Bucharest, 68-72 Panduri St. Bucharest 5, ROMANIA
E-mail: esusnea@yahoo.com

Abstract

Data mining (DM) is useful for collecting and interpreting significant data from huge database. The education field offers several potential data sources for data mining applications. These applications can help both instructors and students in improving the learning process.

Keywords: Data Mining, Education, k-Means Algorithm

1 Introduction

Development of educational means has become a priority for most member states and the rate regarding higher education presents a tendency increased globally.

Also, universities need to develop a special interest in using ICTs in education: “the application of ICTs to teaching and learning has great potential to increase access, quality and success” (UNESCO, 2009).

Following this direction, the use of e-learning technologies has grown to be an alternative solution to improve traditional education. E-learning systems allow collecting huge quantities of data that can be used both by universities and other institutions. The educational database can provide personal information regarding the users’ profiles, their academical grades, and also data regarding interaction among different users.

The KDD process (knowledge discovery in database) can be very useful in the student-centered educational system due to the fact that the information from educational database allow improvement of the teaching and learning level, also improvement of the students’ grades, better understanding of students’ behavior, adjusting the curriculum to the students’ needs, improvement of the quality of educational management etc. This information can be presented as rules, graphics, decisional trees and networks.

Within the KDD process, there can be used different means of DM analysis, that allow getting important information from the database such as: Bayes classifiers, association rules, tree decision, neural networks, genetic algorithms, support vector machines, clustering etc.

Next, we are presenting some aspects regarding the existing differences between models and patterns, and after that a detailed description of k-means method (frequently used in DM), and an application of this method in the educational field.

2 General aspects regarding discovery of models and patterns within the database

The DM process consists mainly in discovering some “valid, new, possibly useful and comprehensible” structures from the dates (Fayyad, U.M., Pitatesky-Sapiro et al, 1996).
The structures discovered during the DM process can describe the entire (the most of the) set of data and they are called models. There are also cases when the structures discovered get some local properties of the data, and in this case the term of "pattern" is used.

From the geometrical point of view, we can represent the set of data by using an $n \times p$ matrix, where $n$ is the number of samples (logins) and $p$ is the number of characteristics (variable or attribute). Thus, for each sample, a number of $p$ measurements, represented by a $p$-dimensional vector. The set of data $X$ can be used under the form $X = \{x_1, x_2, \ldots, x_n\}$.

If the structure discovered in the $p$-dimensional space is a model, then we can get information for each point in this space and thus for each sample. In case some values of the characteristics are missing (the vector is incomplete), these can be determined by customizing the model. In linear regression, a simple model could have the form $Z = aY + b$, where $Y$ and $Z$ are variable, and $a$ and $b$ are the model parameters (constants determined during the DM process) known as regression points.

Unlike the models, the pattern structures provide information regarding some areas of the $p$-dimensional space. For example, a statement as the following if $Y_i > a$ then $\text{prob}(Y_j > b) = p_i$ has certain constraints imposed to the values of the variables $Y_i$ and $Y_j$. From a semantical point of view, the relationship is equivalent to $\text{prob}(Y_j > b | Y_i > a) = p_i$. For example, a database research can show that the family income is a decisive factor for education. Such a research [Kane, J. 1970], that aims to identify the major risk factors in predicting the students’ grades (passes/failed), shows a strong positive connection between the family income and other attributes that characterize the family educational accomplishments: admission, perseverance and graduation. For example, the students with poor financial status have a much reduced class attendance as compared to the students whose families provided a medium income. Thus, we can identify a group of logins (pattern) different from the others (which can be considered a central cloud in the $p$-dimensional space).

The distinction between models and patterns is useful in many cases. Although, sometimes it is not clear whether a certain structure should be considered a model or a pattern.

3 DM techniques used in e-learning

The problem regarding the educational curriculum is present and complex, leading to opinions, theoretical solutions and practical ways of applying, sometimes different, and even contradictory. In a restricted way, a curriculum represents all the official school documents that contain the main information regarding guiding of the educational process. The core of the curriculum consists of the teaching objectives. They are "are specific statements about exactly what a student should know, be able to do, or value as a result of accomplishing a learning goal" (Reed, 2005).
By using DM techniques, we can get new valid reports, even comprehensible models from the databases available in higher education. The discovery of hidden patterns allows the development of some good decisions and also has the advantage of being student-centered. In this way, the educational process can be improved according to the discovered models and patterns.

Next, we are presenting the clustering, that is a DM technique often used when creating a classifying model, and then we are showing an example on how this technique can be used for devising the educational database in homogenous groups. For example, by using the clustering we can identify the main elements for creating and producing of the educational curriculum within an e-learning system.

As far as the automatic learning is concerned, clustering represents an unsupervised learning method. Unlike the classification, which involves the existence of some predefined training classes and clusters meant to develop some predictions, clustering has a descriptive target and learning is done by observational learning, instead of exemplifying learning.

The main objective of the clustering process consists of deviding the data set so that the distance among the clusters should be minimal, whereas the inter-cluster distance should be maximal. In order to verify if two objects are similar or not, we use two types of measures: similarity measures and dissimilarity (distance) measures. Often, in order to determine the dissimilarity ratio between certain objects, we use euclidian distance.

The database can be devided through clustering methods either by using partition-based methods, or hierarchical ones. One of the methods frequently used in data partition is the k-means data.

The k-means algorithm is the easiest and the most common algorithm based on squared error criterion. This represents a simple clustering procedure that desires to minimize $J$ criterion function in an iterative way:

$$J = \sum_{j=1}^{k} \sum_{n \in C_j} |x_n - \mu_j|^2$$

where, $k$ is the number of clusters, and $\mu_j$ represents the point average of $C_j$ cluster and it is given by

$$\mu_j = \frac{1}{N_j} \sum_{n \in C_j} x_n$$

This criterion measures how well represented the X data set is by the cluster centers $\mu = \{\mu_1, \mu_2, \ldots, \mu_n \}$. The methods that use such a criterion are called minimum variance methods (Duda et al., 2001).

The algorithm can be summarised as it follows:

**Step 1.** defines $k$, the number of clusters to which the data set should be partitioned

**Step 2.** initializes the clusters, providing a random set of $k$ logins that will initially be considered centers

**Step 3.** finds the closest cluster center for each login. Usually, the ”closeness: of the cluster center is determined by using the euclidian distance.
Step 4. for each of the k clusters, the average is determined [2] and each sample is allotted to the appropriate cluster according to the nearest average obtained. Step 5. repeats step 3 till convergence or finish point.

There are two major steps in the algorithm, identifying the distance among all the points and center re-evaluation. The cost is determined according to the number of iterations as it follows:

\[ T = O(nkdI) \]

where n represents the number of points, k the number of clusters, and I the number of iterations.

Case study

Course Management Systems (CMS), offers a variety of channels and working spaces in order to improve information sharing and communication among a course participants. This system collects huge quantities of information which are very useful for the analysis of the educational process. Traditional analysis of the data within the e-learning system is based on "hypothesis or assumptions" (Gaudioso, E., Talavera, L., 2006) meaning that the analyst starts with data exploring according to personal intuition, whereas data mining allows an inductive approach with automatic discovery of some patterns hidden within the database.

The study has been conducted in 5 institutions of higher education that dispose of the CMS system, such as Moodle, Ilias or Blackboard. 200 students have been questioned.

By using the k-means algorithm we have realised 3 clusters taking into account the study year and the answer to the question "How much do the teaching objectives stated in the educational plan and analytical program correspond to your aspirations?".

In Figure 2 we have presented the dissimilarity matrix. Based on the dissimilarity criterion, we can determine the three clusters (Figure 3 and Figure 4).
Figure 4. Medium distance within the clusters

It can be noticed that there is a bigger dissimilarity between clusters 1 and 3 and a bigger similarity between the elements belonging to cluster 1.

The disadvantage of using k-means algorithm is represented by the fact that "best number of k clusters is not known, as it is chosen relatively as an initial value of the algorithm" (Molder, C., 2004).

4 Conclusions
The development of some adequate and efficient teaching strategies is not a simple operation. It implies a contextual, original and unique combination of the elements of the entire teaching-training process. Lately, most teachers with an open mind regarding the teaching methods, have begun to reorganize their classes, tackle new topics, and present the contents in a dynamic form so that to make the students get better scores.

REFERENCES
http://sacs.utdallas.edu/sacs_glossary

<table>
<thead>
<tr>
<th>Cluster</th>
<th>#Obs</th>
<th>Average distance in cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster-1</td>
<td>72</td>
<td>0.467</td>
</tr>
<tr>
<td>Cluster-2</td>
<td>78</td>
<td>0.524</td>
</tr>
<tr>
<td>Cluster-3</td>
<td>50</td>
<td>0.712</td>
</tr>
<tr>
<td>Overall</td>
<td>200</td>
<td>0.55</td>
</tr>
</tbody>
</table>