Prediction of the Students’ Behavior in Real-time Education Management System

Kristina Khudaverdyan
NPUA
Yerevan, Armenia
e-mail: kkhudaverdyan@polytechnic.am

***Abstract*—Currently, educational institutions face the problem of continuous growth of educational data and application of this data in the process of upgrading the quality of making decisions. Besides these decisions, predicting students´ behavior and performance is critical, and all the data related to education and performance is based on different factors such as personal information, psychological and social details. Hence, educational databases include such kind of helpful information for prediction purposes. In this context, machine learning is becoming more useful as it is related to enhancing methods for knowledge exploration from data coming from educational establishments. In addition, it enables more production decision making. The data mining techniques are more helpful in classifying educational data. The explanatory variables need to be filtered when logistic regression, decision trees, and other techniques for models are used. However, choosing a variable in a model is a really challenging process, therefore, predictive power, the correlation between variables, simplicity, and reliability need to be used. Nevertheless, predictive power of the variables is the most vital measure to take into account.**

***Keywords*— Prediction, students’ behavior, classification, K-means algorithm, Fuzzy c-means algorithm, making decisions, predictive power.**

I. Introduction

The use of decision making assistance methods in the education management system will allow, based on the analysis of students' involvement, perception and response, quickly make a decision on how to continue further education. It is assumed that in order to solve decision making problem, it is necessary to classify the incoming data, then to use appropriate measures and it will be possible to make a forecast. If operational decisions are made, these forecasts should be as accurate and fast as possible.

Data collected from educational systems can be aggregated from a large number of students and can contain many variables that can be used to build a predictive model using data processing algorithms:

Data processing is the process of obtaining the necessary knowledge from huge data. Data processing has three main components: Clustering, Classification, Association (Correlation) [3].

Clustering refers to the collection and division of data into certain categories. A simple example of clustering is grouping students according to their abilities and knowledge.

Correlation refers to the detection and maintenance of data variables for future use. For example, the correlation between the taught subjects or their sequence may indicate problems with the deterioration of the student's academic performance.

Suppose we need to build a predictive model, this model is designed to predict whether during the lesson each student in the group can correctly answer/react to the regular test, which is designed to check the level of mastering the topics of the lesson or the level of acquiring knowledge during the lesson. Thus, we want to predict the student's reaction, behavior to our actions, to answer correctly or not. The prediction is based on mathematical calculations and statistics.

Prediction accuracy is important since it can be very useful in planning educational interventions aimed at improving the results of the teaching-learning process, saving resources and educators’ time and effort. Moreover, the additional use of pre-processing techniques along with classification algorithms have improved performance prediction accuracy.

In order to make the teaching effective, it is necessary to collect and analyze the data generated during the teaching, the survey conducted during the course, on the basis of which the lecturer will be suggested a way to make an optimal decision to conduct the lesson.

First of all, it is necessary to evaluate the existing knowledge of the students, group them, and then the result collected from each group will form a prediction model, on the basis of which the lecturer will make an effective decision. As we can see in the Figure 1.

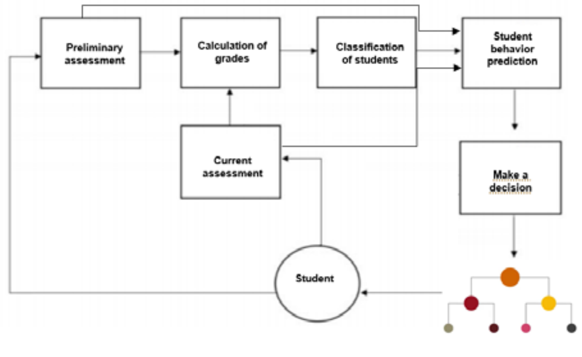

Figure 1.

A preliminary assessment can be the student's final grade, number of attendances, current grades, etc.

All this represents external data for the organization of the given course. In addition to external data, it is also very important to have current or internal data about the student, to make possible the monitoring of the student's progress throughout the course, since the student's progress may be changed during the course. Current internal data or quizzes may be the source of such internal data. The source of obtaining such internal data can be tasks or quizzes related to the course.

II. Student Classification

Cluster analysis was used to classify students [3]. Clustering means grouping individual parts of the integrity according to certain characteristics. There are several clustering criteria .

- Clustering algorithms based on fragmentation divide data into separate parts so that each part contains at least one object,
- Hierarchy-based clustering algorithms sort data according to the mean correspondence hierarchy. This means, it starts with a cluster and gradually divides into sub-clusters,
- Density-based clustering algorithms group data by the density of their alignment,
- Network clustering algorithms group the data into separate networks, after which only clustering is carried out, based on the indicators of individual networks,
- Model-based clustering algorithms perform clustering by assessing the relevance of existing data and pre-defined mathematical models.

The centroid model was used in this work. One of the earliest clustering algorithms used in educational programs is the "Fuzzy C-Mean" clustering algorithm. In particular, studies [1, 3] have shown that it is effective for grouping students according to certain characteristics, as it is a prerequisite for course organization. The "FCM" algorithm first finds the unit representing each cluster, called the central axis. The algorithm then calculates the membership unit of an individual unit on each cluster. The “FCM” algorithm, by performing iterations, finds the central axes of the clusters and updates the membership unit of the objects. The membership unit ranges from 0 to 1 and shows how much the object corresponds to the given cluster. The higher the membership score, the higher the level of compliance with the cluster.

The main purpose of using the "FCM" clustering algorithm, as already mentioned, is to understand the learner level. However, it has certain disadvantages.

In particular, if the student's score corresponds to two clusters at the same time, that is, the point representing the data of this student is located at the same distance from the two centroids, will result not in an entirely accurate picture of the situation. For example, if the X student membership rate for cluster 1 is 0.0513, 0.4134 for cluster 2, 0.1429 for cluster 3, 0.3863 for cluster 4, and the efficiency indicator is 0.05%, then both cluster 2 and cluster 4 will be considered for this student, as the difference between them is less than 0.05 (0.4134 - 0.3863 = 0.0271 <0.05).

To solve this problem, you can use the "K-Means" algorithm, which provides accurate data. In this case, the student can belong to only one group.

III. Data Cluster Analysis Techniques

FCM algorithm: The FCM clustering algorithm is represented by the following formula [1]:

$$\boldsymbol{J = \sum_{i=1}^{n}\sum_{j=1}^{c} U_{ij}^{m} D_{ij}^{2}}$$

where **n** is the number of objects, **c** is the number of predefined clusters, **Uij** is the degree of membership of object i for cluster j, m is the approximation index (m> 1),

$$\boldsymbol{D_{ij} = \sqrt{\sum_{i=1}^{n}(p_i - v_j)}}$$

Dij is the Euclidean distance between the ith object p_i and the jth cluster center. v_j represents the jth cluster center calculate by following formula:

$$v_j = \frac{\sum_{i=1}^{n} U_{ij}^{m} p_i}{\sum_{i=1}^{n} U_{ij}^{m}}$$

The membership unit is calculated using the following formula:

$$U_{ij} = \frac{1}{\sum_{k=1}^{n}(\frac{|p_i - v_j|}{|p_i - v_k|})}$$

The first input parameter of the “FCM” algorithm is the collected data. It is presented in the form of objects that reflect students' grades. The initial U membership unit is randomly generated.

The output of the ‘FCM” algorithm is an array which number of rows is equal to the number of students and the number of columns is equal to the number of clusters. Each column of the row includes the student membership for that cluster.

«K-Means» algorithm: The essence of the "K-Means" algorithm is that the entered data are grouped into a number of K clusters. First, a random number of K-centers are selected, and the distances of the points from them are calculated, after which the center closest to each point is selected and the point is assigned to it. This process is repeated until the values of the cluster centers are the same after iteration [2, 3].

The algorithm is represented by the following formula:

$$J = \sum_{j=1}^{k}\sum_{i=1}^{n} ||x_i^j - c_j||^2$$

$||x_i^j - c_j||^2$ is the distance between x_i^j and c_j. x_i^j is the entered data, and c_j is the current coordinates of the center.

The formula for calculating the coordinates of the center is as follows:

$$v_i = (1/c_i)\sum_{j=1}^{c_i} x_i$$

where c_i is the amount of data in cluster i, and x_i is the amount of data entered.

The algorithm is implemented in the following sequence of steps:

1. Based on the entered data, the cluster centers are initialized. This process can be performed by various methods, the most common of which is to randomly select one of the entered data and assign it to the center.

2. One center is selected for each point. A special formula is used to select the center, which, calculating the distance of the point from all the centriodes, in turn, selects the shortest distance.

3. The coordinates of the centers are recalculated based on the calculated distances of the received data.

Step 2 and step 3 are repeated until the coordinates of the centers change.

Thus, each of the listed algorithms has its advantages and disadvantages.

After studying, it becomes clear that the "K-Means" algorithm is more appropriate for solving this problem.

IV. Prediction of Students' Behavior

Thus, after solving the problem of students classification, the next step is to predict students' behavior, as it will contribute to effective decision making in the education management system. In many areas where the predictive model is built on raw historical data, training data preparation is the most important step in building a strong model. Thus, proper data preparation will yield stronger results. Applying appropriate data transformations can yield significantly improved results. First of all, it is necessary to find the connection between the target variable "variables".

Evidence Weight (WOE) and Information Value (IV) are simple but powerful methods for variable transformation and selection. These concepts have a lot to do with logistics regression modeling techniques.Each parameter in the model has its own weight: one characteristic may be more important than another.

Compared to the previous approach, which is already outlined, this method does not identify "good" and "bad" students individually, but it estimates the probability that students with a given result will give “right” or “wrong” answers. Prediction results are used as a basis for decision making.

The term "predictive power of variables" is very general, subjective, and non-quantitative. When choosing variables, we can never say, "I think this variable has a strong predictive power, so it should go into the model, right?" We need some specific scores to measure the predictive power of each independent variable and, depending on the size of those scores, to determine which variables are included in the model. IV is such an indicator, it can be used to measure the predictive power of independent variables.

Using IV to measure the predictive power of variables can be roughly understood as follows: we assume that there are two categories of target variables in the classification problem: Y_1, Y_2. To predict individual A, in order to judge whether A belongs to Y_1 or Y_2, we need certain information. Assuming that the total amount of information is I., the required information is contained in all independent variables C_1, C_2, in C_3…, C_n, then for one of the variables C_i the more information it contains, the greater its contribution to determining whether A belongs to Y_1 or Y_2, and the greater the information value of C_i.

Machine learning algorithms mostly take numbers as input, so we have to convert features to numbers before training the model. This requires that all features are numerical. However, we might have categorical features in our datasets that are either nominal or ordinal. There are many methods of assertion. In this case, we can use the concept of WOE (Weight of Evidence) to impute the categorical features. Using WOE - reducing the number of input data columns used to train the model. Imagine you have a categorical variable with 10 different classes, and you do a one-time coding, and you end up with 10 columns with mostly "0" values. Using the WOE technique, the classes are replaced with the corresponding WOE values [6].

As for IV, values provide a basis for further deepening our analysis of the relationship between independent and dependent variables. Also, if the variable is of a quality type, we can use binning, followed by WOE and IV concepts, to design meaningful features.

V. Calculation of Weight of Evidence and Information Value

Information Value (IV) is widely used in financial industry. Information value is a numerical value to measure the predictive power of an independent variable x to describe dependent binary variables y. Mathematically, it is defined as [4]:

$$\mathrm{IV} = \sum_{i=1}^{n}((\frac{g_i}{g} - \frac{b_i}{b})x\ln(\frac{g_i/g}{b_i/b}))$$

where n is the number of bins or groups of variable x, g_i and b_i are the numbers of good and bad answers with bin i, and g and b are the total number of good answers and bad answers. Hence, g_i/g and b_i/b are distributions of good answers and bad answers. Therefore,

$$\sum_{i=1}^{n}\frac{g_i}{g} = \sum_{i=1}^{n}\frac{b_i}{b} = 1$$

Usually, “good” means y=0 and “bad” means y=1. It could be the other way, since IV is symmetric about good and bad. If $gi/g = bi/b$ for all i = 1, . . ., n, then IV = 0; that is, x has no information on y. IV is mainly used to reduce the number of variables as the initial step in the logistic regression, especially in big data with many variables. IV is based on an analysis of each individual predictor in turn without taking into account the other predictors.

IV and WOE. One advantage of IV is its close tie with weight of evidence (WOE), defined by ln((gi/g)/(bi/b)). WOE measures the strength of each grouped attribute in separating good and bad answers [5]. According to [4], WOE is the log of odds ratio, which measures odds of being good. Moreover, WOE is monotonic and linear. Indeed, g_i/g and b_i/b are from two different probability distributions. They represent the number of good answers in bin i divided by the total number of good answers and the number of bad answers in bin i divided by the total number of bad answers, respectively.

When a continuous variable x has a large IV, we make it a candidate variable for logistic regression, because we know that logistic regression models take as input both categorical and numerical data and output the probability of the occurrence of the event. It makes by giving the student data, what is the probability that the student will give the “right” answers to the evaluation process of student's knowledge.

A Rule of IV. Intuitively, the larger the IV, the more predictive the independent variable. However, if IV is too large, it should be checked for over predicting. For instance, x may be a post knowledge variable. To quantify IV, a rule is proposed in Table 1. [4]:

Information Value	**Predictive power**
<0.02	Useless
0.02 to 0.1	Weak predictors
0.1 to 0.3	Medium Predictors
0.3 to 0.5	Strong predictors
>0.5	Suspicious

Table 1.

In addition, mathematical reasoning of the rule is given in [4].

Suppose we randomly select students from multiple groups to test and collect the results of these students' responses as a simulation dataset, of which some of the students responded. In addition, it is assumed that we also extracted some of the variables of these students, as a set of candidates for our model, these variables include the following (in some situations we may have much more than they are, the variables listed here are only intended to explain our problem): A, B, C, D, E.

Assuming we discretized these variables, the statistical results are shown in Table 2.

Feature (values) X	N of events g_i	N of non-events b_i	Percentage events	Percentage non-events	WoE	IV
A	80	3500	0.16	0.37	0.838	0.176
B	90	2400	0.184	0.25	-0.3065	0.02
C	90	1100	0.184	0.12	0.427	0.026
D	130	1300	0.16	0.14	0.659	0.175
E	100	1210	0.2	0.18	0.105	0.002
Sum	490	9510				0.399

Table 2.

The weight of evidence tells the predictive power of a single feature concerning its independent feature. If any of the categories of a feature has a large proportion of events compared to the proportion of non-events, we will get a high value of WOE which, in turn, says that class of the feature separates the events from non-events. For example, consider category A of the feature X in the above example, the proportion of events (0.16) is very small compared to the proportion of non-events (0.37). This implies that if the value of the feature X is A, it is more likely that the target value will be 0 (non-event). The WOE value only tells us how confident we are that the feature will help us predict the probability of an event correctly.

VI. Conclusion

Data science plays a decisive role in the field of education as well. Data science and working with large amounts of data greatly help in decision making. No less important for making an effective decision is the analysis of data formed at different stages of the educational process. Data analysis is defined as the process of finding useful information for decision-making in different areas. Clustering analysis was used to classify students based on their performance in the education management system. A comparative analysis was performed between the following clustering algorithms: K-means algorithm and Fuzzy c-means algorithm.

As a result of this paper underlined the importance of measurement of predictive power that is related to the weight of evidence and information value, which can be used for prediction of students' behavior in the education management system.

References

[1] R. Suganya, R. Shanthi, *Fuzzy C- Means Algorithm- A Review, International Journal of Scientific and Research Publications,* vol. 2, no. 11, 2012.

[2] Harikumar Rajaguru; Sunil Kumar Prabhakar, *KNN classifier and K-means clustering for robust classification of epilepsy from EEG Signals: a detailed analysis,* pp. 40-47, 2017.

[3] А. А. Барсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод, *Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP* 2-е издание, Санкт-Петербург, pp.59-67, 156-163, 2007.

[4] N. Siddiqi, *Credit Risk Scorecards—Developing and Implementing Intelligent Credit Scoring, John Wiley & Sons,* 2006.

[5] Guoping Zeng, *A Necessary Condition for a Good Binning Algorithm in Credit Scoring,* 2014.

[6] I. J. Good, *Weight of Evidence: A Brief Survey.* BAYESIAN STATISTICS 2, pp. 249-270, 1985.