ISBN 978-981-11-7861-0
Proceedings of 2018 the 8th International Workshop on Computer Science and Engineering (WCSE 2018)
Bangkok, 28-30 June, 2018, pp. 298-302
doi: 10.18178/wcse.2018.06.053

Embedding Machine Learning Algorithm Models in Decision Support System in Predicting Student Academic Performance Using Enrollment and Admission Data

Ace C. Lagman [1+] and Rossana T. Adao [2]

[1, 2] FEU Institute of Technology

Abstract. Academic Analytics is extracting hidden patterns from educational databases. The main goal of this area is to extract hidden patterns from student academic performances and behaviors. One of the main topics in academic analytics is to study the academic performance of freshman students. Students enrolled in first year are the most vulnerable to low student retention in higher education institution. Research studies from different Higher Educational Institutions already indicated that early identification of students with academic difficulty is very crucial in the development of intervention programs. As such, early identification of potential leavers and successful intervention program(s) are the keys for improving student retention. The study will utilize the available enrollment and admission data. Feature selection technique will be used to determine significant attributes. The study aims to produce predictive and cluster model in which can early identify students who are in need of academic help and program interventions. The extracted predictive and cluster models will be evaluated using confusion matrix and be integrated in the decision support application.

Keywords: information system, decision tree algorithm, education data mining, decision support system.

1. Background of the Study

Academic Analytics is an emerging area of data mining focusing on extracting useful patterns from student databases. One of the major problems in academic analytics is student retention. Addressing this problem is very crucial as all academic institutions aim to increase student retention rate as a long term goal. The consequences of student attrition are significant for students, academic and administrative staff (Ishitani, 2006)[1].

Students enrolled in first year are the most vulnerable to low student retention. Freshman students have the greatest risk of dropping out and not to continue schooling. The research gap motivates the researcher to focus on this area of research.

The study aims to utilize enrollment and admission data in making predictive and cluster model to establish a university admission decision making models in early identification of possible students who are in need of academic and counselling intervention from school administration.

The study aims to develop decision support systems (DSS) that embeds the predictive and cluster models that are extracted using machine learning algorithms. A DSS is a specific class of computerized information system that supports business and organizational decision-making activities. It is an interactive software based system that early identifies students who are prone of having an academic difficulty.

1.1. Research questions

1. How feature selection technique can identify significant attributes affecting student academic performance in their early year in college?

+ Corresponding author.
E-mail address: aclagman@feuteche.du.ph.

2. What data model can be developed using classification algorithms?
3. What cluster model can be designed using clustering technique that can profile students' academic standing?
4. How effective the derived model using Confusion Matrix Test?
5. How effective the developed system decision support system in terms of User Acceptability Test?

1.2. Scope and limitation

The study will focus on the students' enrollment and admission data of Far Eastern University, Institute of Technology. Feature selection technique will be used to determine which among sets of predictors are highly significant in the development of the predictive model.

The study is only delimited to freshman students because according to literature it is considered as the most vulnerable to low student retention. The retention strategy aims to help improve and increase retention rate. This is very crucial to early identified students who are vulnerable to drop their courses. The researchers will use decision tree algorithm to generate data models applicable for prediction. The study will focus only with students with similar first year subjects. The confidentiality of student's data will be protected by deleting personal information of the data from data sources. The gathered data will be processed using Data Mining tool such as WEKA and SPSS which consist of data mining techniques and statistical packages to prevent the researcher for making mistake or even forgetting something on computations.

The historical data will be divided into two parts; training and test data. The training data will be used to generate data and cluster model. The model will be evaluated using the testing data by calculating the accuracy results using confusion matrix test. The models are derived rules sets from decision tree algorithms. The model will be embedded to software that predicts student academic standing.

The data will be prepared to a format that a data mining tool can be recognized. Table below show the sample attributes to be used for the analysis. The first year standing variable will be coded as the dependent variable of the study.

Table 1: Sample Attributes of Admission and Enrollment Data

Attributes	Values	Description
Sex	{male, female}	Gender of the student
Location	{city, municipality}	location of residence of the student
HS Grade	0 – 100	Average grade in Senior High School
Parents Occupation	{ local, international }	Parents work
Scholarship	{ Yes, No }	Scholarship Grant given
HS Grade	0 – 100	Average grade in Senior High School
First Term Grade and Second Term Grade	0-100}	General Weighted Average of the Grades of Students
First Year Standing	{Irregular, Not Regular}	General Weighted Average of the Grades of Students

2. Literature Review

A. Educational data mining

The educational data mining focused on the student understanding aspect. The different computer learning tools which include tutoring and simulations gave opportunities to find patterns and trends of students behaviour. One main possible data source are students' online system in which columns or attributes can be studies and researched to build data models for prediction (Juan Francisco; Vandamme, 2006) [2].

Baker &Yacef (2009) [3] and further study of Baker (2010) [4], suggests four key areas of application for EDM: improving student models, improving domain models, studying the pedagogical support provided by learning software, scientific research into learning and learners; and five approaches/methods: prediction, clustering, relationship mining, distillation of data for human judgment and discovery with models..

B. Feature selection

Gareth et al (2013) [5] discussed that in machine learning and statistics, feature selection is used to as a variable selection mechanism to determine relevant and significant attributes or predictors that can be used for model construction.

C. Classification algorithm

Decision trees are able to process both numerical and categorical data without requiring any domain knowledge to classify their data. The data is partitioned according to the best split and this in turn creates a new second partition rule. The process goes on until there are no more splits. The resulting tree is known as a maximal tree. (Kesavulu, Reddy, & Rajulu, 2011) [6].

D. Clustering algorithm

The k-means algorithm provides an easy method to implement approximate solution to. The reasons for the popularity of k-means are ease and simplicity of implementation, scalability, speed of convergence and adaptability to sparse data.

```
1     MSE = largenumber;
2      Select initial cluster centroids {m_j}_j
      K = 1;
3     Do
4        OldMSE = MSE;
5        MSE1 = 0;
6        For j = 1 to k
7           m_j = 0; n_j = 0;
8        endfor
9        For i = 1 to n
10          For j = 1 to k
11          Compute squared Euclidean
            distance d^2(x_i, m_j);
12          endfor
13       Find the closest centroid m_j to x_i;
14          m_j = m_j + x_i; n_j = n_j+1;
15          MSE1=MSE1+ d^2(x_i, m_j);
16       endfor
17       For j = 1 to k
18          n_j = max(n_j, 1); m_j = m_j/n_j;
19       endfor
20     MSE=MSE1;
   while (MSE<OldMSE)
```

Fig. 1: K-Means algorithm.

Figure 1 shows the algorithm for K-Means under Clustering Technique. The k-means always converge to a local minimum. The particular local minimum found depends on the starting cluster centroids.

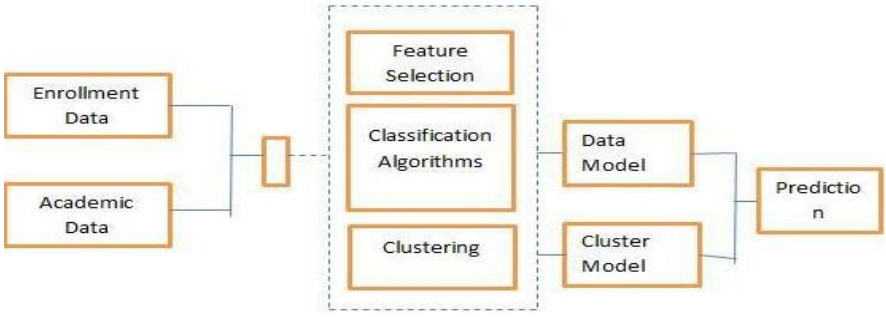

Fig. 2: Conceptual framework

Figure 2 reveals the processes of how the system processed inputs and the different procedures used in extracting data models. The data consist of enrolment and academic data as inputs. These inputs will be processed using feature selection

3. Research Methods or Methodology

The researcher will utilize the steps of Knowledge Discovery in Databases and CRISP-DM methodologies in creating the study. The life cycle of a data mining project as defined by CRISP-DM format consists of six phases and the knowledge discovery in database consists of nine steps.

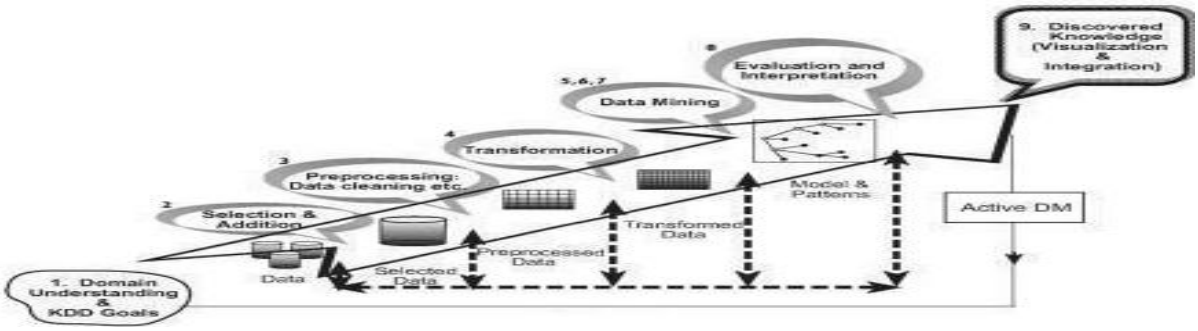

Fig. 3: CRISP – DM methodology

A. Knowledge in the domain

This step involves understanding and defining the goal of the end users, then where knowledge discovery process will take place and other relevant prior knowledge. The researcher will focus on the educational domain particularly on admission and enrollment data.

B. Selection and addition

This step involves selection of attributes needed to be performed based on goals. This section answers what data is needed and are these data are available for consumption.

C. Pre-processing and cleaning

Data reliability is enhanced in this stage. It includes data clearing, such as handling missing values, and removing of outliers. The researcher will use SQL script commands as a data processing technique to query relevant and potential data from databases.

D. Data transformation

In this stage, the generation of better data, for the data mining is being prepared and developed. The data will be transformed in to a proper format using data pre-processing technique using a data mining tool.

E. Modeling

This study will use statistical predictive algorithms and clustering techniques. The classification algorithms will be then tested to determine the best model. The clustering technique will be used to profile the students to extract cluster models.

3.1. Logistic regression

In feature selection, logistic regression will be used since the target variable is dichotomous or binary in essence. Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). A logistic regression produces a logistic curve, which is limited to values between 0 and 1. Logistic regression is similar to a linear regression, but the curve is constructed using the natural logarithm of the "odds" of the target variable, rather than the probability. Moreover, the predictors do not have to be normally distributed or have equal variance in each group

To determine the statistical significance of a predictor the p value will be used. The predictor is statistically significant when a p value is less than the significance level. The p-value is the probability of observing an effect given that the null hypothesis is true whereas the significance or alpha (α) level is the probability of rejecting the null hypothesis given that it is true. In practice significance level is chosen before data collection and is usually set to 0.05(Agresti, 1990) [7]

3.2. Cluster analysis

To determine the similarity profile of the students cluster analysis using K-Means will be used. The clustering technique identifies similar patterns present in the cluster models extracted from the data sets. The cluster models will group students based on similarity using Euclidean distance. This will help the Administrators to identify what recommendation can be derived or designed based on the cluster model.

3.3. Classification analysis

To determine the predictive model, decision tree algorithm will be used. This entails to produce an if then else rule statements from the datasets. The target variable will be coded as regular as 0 and 1 as irregular. The purpose of the decision tree model is to build decision rule sets or patterns of 0's and 1's.

4. Evaluation

This stage involves evaluating the models built in the model building stage. The most common way to evaluate models is to verify their performances on the test datasets. Evaluation of the models can be easily determined by observing the number of correct predictions to the total number of predictions.

The receiver operating characteristic (ROC), illustrates the performance of a binary classifier as its discrimination threshold is varied. The curve is created by plotting the true positive rate against the false positive rate at various threshold settings.

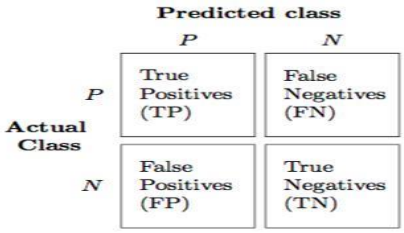

Fig. 4: The Receiver Operating Characteristic (ROC)

The true positives (TP) and true negatives (TN) are correct classifications. The true positive rate is TP divided by the total number of positives, which is TP + FN; the false positive rate is FP divided by the total number of negatives, FP + TN. The overall success rate is the number of correct classifications divided by the total number of classifications.

$$\textbf{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$
$$\textbf{Error Estimation} = 100 - \textbf{Accuracy}$$

Eq. 1: Error estimation

This evaluation scheme will help the researcher determine the best suited classification algorithm for this study.

5. Future Works

The researchers will develop a decision support system. The decision support system is an interactive software that early identifies students who are prone of having an academic difficulty. The systems' main goal is to predict student who will have an academic difficulty based on enrollment and admissions data so proper remediation and policy can be formulated to help students to improve their academic standing. The system will be evaluated using ISO 9126 metrics to determine the acceptability of the application in terms of its operational use.

6. References

[1] Ishitani, T. (2006). Studying attrition and degree completion behaviour among first-generation

[2] Juan Francisco; Vandamme, Jean-Philippe; Meskens, Nadine. (2006) "Determination of factors influencing the achievement of the first-year university students using data mining methods".

[3] Baker, R., Yacef, K. (2009) The State of Educational Data Mining in 2009: A Review and Future Visions. Journal of Educational Data Mining, 1, 1, 3-17.

[4] Baker, R. S. J. D. (2010). Data mining for education. International encyclopedia of education, 7, 112-118.

[5] Gareth James; Daniela Witten; Trevor Hastie; Robert Tibshirani (2013). An Introduction to Statistical Learning. Springer. p. 204.Ishitani, T., & Snider, K. Longitudinal effects of college preparation programs on college

[6] Kesavulu, E., Reddy, V., & Rajulu, P. (2011). A Study of Intrusion Detection in Data Mining. *World Congress on Engineering 2011. III.* London, UK: WCE.

[7] Agresti, A. (1990). Categorical Data Analysis. Wiley, New York.