Academic Mining to Assist the Guardian Mechanism Using Naive Bayes Classifier

Veronikha Effendy*, Thee Houw Liong**
* Faculty of Information Technology, Institute of Technology Telkom
** Faculty of Postgraduate, Institute of Technology Telkom

Keywords:

Academic Mining
GPA Prediction
Naive Bayes Classifier
Simple linear regression
Guardianship
Faculty trustee

ABSTRACT

Until now, academic data only used for the purposes of calculating GPA, graduation requirements, transcripts, and others processes related to reports and regular information. Academic data might be a tremendous source of information on educational improvement. One of the information which can be exploited is information on the predictive value of student in the future period (e.g. the next semester).

This study attempted to use a data mining approach to help the guardianship process. The guardianship only needed the estimated range of GPA value to be able to determine whether the student is in the safe position or not based on academic rules and regulations.

Data pre-processing implemented feature selection, raw selection, discretization and data sampling to get the best result. This study implemented Naive Bayes Classifier (NBC) and simple linear regression as the prediction method. The result from both methods will be compared at last.

Comparison results from NBC and simple linear regression in several scenarios shown that NBC with data sampling got the best accuracy among other scenarios. The result means that data mining approaches is worth enough to be considered as a solution for predicting student's academic success.

Corresponding Author:

Veronikha Effendy,
Faculty of Information Technology,
Institute of Technology Telkom,
Telekomunikasi Street No 1, Bandung, Indonesia.
Email: veffendy@gmail.com

1. INTRODUCTION

Student academic success is the desire of all people. Industry also desires to get high quality graduates. However, even a higher education institution has difficulties to control the success of the students. Each person is unique, and has unstable emotions. Therefore, the treatment of each student must also differ from one anothoer according to their characteristics. This reason makes the faculty trustee’s task pretty heavy.

Prediction of student academic success ideally requires different supporting data, such as the results of psychological tests, college entrance test results, previous GPA, family profiles, regional profiles, economic status, and the activities carried out by students during college and so on. Unfortunately, it is not an easy thing to obtain all of the data and merge into a single set of data. This study attempts to utilize data consisting of GPA per semester from each student. Other data still cannot be used because there are quite a lot of missing data from the institution’s information system.

Prediction of the next GPA based on historical data values is feasible. The students and faculty trustee need to know the prediction’s result to be able to estimate their difficulties in the future. Instituition’s rules have an important role here, such as a minimum GPA policy of each level. Furthermore, the academic success of the students could be detected earlier if the data used as input in the system can meet the learning process of the system. Faculty trustee can optimize their duty to be able to guide and assist students in overcoming their academic difficulties by utilizing the results of this prediction system.

There are many researches on student performance prediction. One of those research used several methods, such as Naive Bayes Classifier (NBC), J48, and Multi Layer Perceptron (MLP). The research result stated that NBC had the best perform among the three methods using historical GPA and student's profile data [1].

This study uses data that similar to the data used in previous research. As we know previous research shown that the best method is NBC, so this study tries to apply NBC. NBC can handle both numerical and categorical data with the class label is in the nominal data type. Moreover, we plan to complete data with various nominal and numerical data such as gender, birth location, economic status, student's activities, etc. Unfortunately, this study still uses only the numerical data due to the lack of other supporting data. Since numerical data is used, the result of the experiment using NBC will be compared by the result of the simple regression. This study predicts student success in four category (poor, critical, good, and very good). Oversampling techniques overcomes the imbalanced data problem which occured in the data set. Prediction results will be one important consideration in the mechanism of guardianship. The critical student will need more attention from the faculty trustee.

Bayesian classifier is a statistical-based classifier. The classifier is able to predict the likelihood of membership of a class, for example the possibility of a tuple of data is a member of a particular class. NBC is one of the simplest Bayesian classifier. NBC works as follows [2]:

a. D is a training set that contains the data tuple. Tuple data is represented by n-dimensional attribute vector, $X = (x_1, x_2 \ldots\ x_n)$, where n indicates the number of attributes.

b. Suppose there are m classes, $C_1, C_2 \ldots C_m$.

Given a tuple X, the classifier will predict that X is a member of a class that has the highest posterior probability on condition X. Therefore, it can be formulated that X is a member of the class C_i, if and only if it meets the following requirements: $P(X \mid C_i)\, P(C_i) > P(X \mid C_j)\, P(C_j)$, where $1 \leq j \leq m$ and $i \neq j$.

This study also uses regression analysis because all predictor and class labels are numerical data. Straight line regression analysis or better known as simple linear regression will select predictor (attribute) that produces the lowest squared error. It involves a response variable y, and a single predictor variable x. It is the simplest form of regression, and model y as a linear function of x. That is $y = w_0 + w_1 x$. The coefficients, w_0 and w_1 can be solved by the method of least squares, which estimates the best-fitting straight line as the one that minimizes the error between the actual data and the estimate of the line. Let D be a training set consisting of predictor variable, x, for some population and their associated values for response variable, y. The training set contains data points $(x_1,y_1),(x_2,y_2), \ldots, (x_i, y_i)$. The coeeficients can be estimated using the equations in figure 1 [2].

$$w_1 = \frac{\sum_{i=1}^{|D|}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{|D|}(x_i - \bar{x})^2}$$

$$w_0 = \bar{y} - w_1 \bar{x}$$

Figure 1. The regression coeficients equations

2. RESEARCH METHOD

Figure 2 represents the method applied in this study:

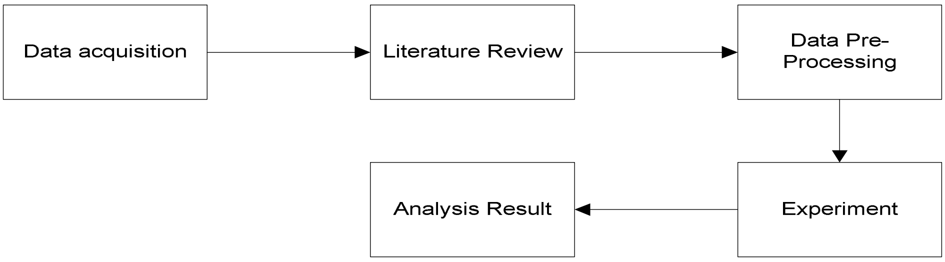

Figure 2. Research procedure

This study has data from SISFO division of Institute of Technology Telkom. The data consists of GPA records per semester in one program study. The original data consists of attributes: ID, high school, hometown, GPA, semester period, academic year and current student status. The data contains many missing value in some of the attributes. We removed the attributes wich has more than 80% missing value.

We process the raw dataset into the data set that ready to use in the classification process. The number of rows of the data set represents the number of students and number of columns represents the attributes. We apply some treatment to the data in order to solve the problem occured in the data, such as the incomplete data entry and also empty value caused by the leave policy applied in the system. Numerical data input is rounded up to one decimal point to get better result. We uses attributes data in both numerical and nominal type for NBC implementation, and numerical type for the simple regression. The study will compare the accuracy results from those scenarios.

The class label that will be predict (poor, critical, good, very good) is formed from the value of GPA in the next semester. The class label will be represented by numbers, which are can be treated as numerical or categorical data type. Table 1 show the labeling criteria equivalent with the class which want to predict in this study. The label data was formed from the GPA in 5th semester.

Table 1. Labeling Criteria

Range Values	Label	Class
0.0 -1.9	0	Poor
2.0 - 2.4	1	Critical
2.5 - 2.9	2	Good
3.0 – 4.0	3	Very Good

Based on the class category, the data distribution in each class is not balanced. We use SMOTE for balancing the data set to apply over sample on the minority data, since SMOTE is good for solving the imbalance data set. SMOTE create a synthetic data rather then duplicate the same data from the minority data [3]. We set 100% oversampling on the category 0 and category 1 with 3 nearest neighbor. We will use the data with and without sampling and compare the result at last.

We use WEKA 3.7 [4] to execute the experiments. This study use 10-fold cross validation to get a fairer accuracy result. The research analyze the experiment result i terms of its performance and its relation to the purpose of this study.

3. RESULTS AND ANALYSIS

The result of each scenario is shown on Table 2.

Table 2. Comparison Result of Experiments

Scenario	Accuracy on Train Set	Accuracy on CrossVal
NBC + Data Numeric + SMOTE	67.85%	67.54%
NBC + Data Nominal + SMOTE	76.92%	70.92%
NBC + Original Data Numeric	65.85%	64.84%
NBC + Original Data Nominal	72.97%	63.41%
Simple Linear Regression + Original Data	47.36%	47.36%
Simple Linear Regression + SMOTE	58.15%	58%

Based on the results shown in Table 2, applying NBC with balancing and transforming data numeric into nominal get the best accuracy on about 70%. It also shows that NBC and simple linear regression get their best accuracy by applying balancing data process before prediction.

Referring to the original purpose of that prediction is intended to help faculty trustee to perform guardianship mechanism, then the prediction error is very risky when the data are included in the poor and critical category are incorrectly predicted as good and very good category, which means that the students are assumed to be in a safe position, while the truth is they are not. Table 3 shows incorrect data prediction examples. It can be seen from the table that there is movement on the fluctuating value of GPA each semester, even some of the data shows a fairly extreme movements. Such conditions presumably influenced by outside factors other than academic. So it can be concluded that the prediction error are made due to limited data.

The same condition also can be seen in the result of simple linear regression which has prediction data >= 0.5 far from the actual data. The data examples are shown in Table 4.

Table 3. False predicted data examples (NBC)

sm1	sm2	sm3	sm4	predicted idxsm5	idxsm5
3.6	2.5	3.3	2.3	3	0
3.5	2.8	2.6	2.8	3	0
2.8	2.9	3	2.6	3	0
2.6	3.1	3.2	2.2	3	1
2.8	2.5	3.3	2.3	2	0
3.1	1.1	3	1.7	2	0
2.9	2.8	3	2.3	2	1
3.1	2.5	2.7	2.2	2	1

Table 4. False predicted data examples (simple linear regression)

Sm1	Sm2	Sm3	Sm4	Predicted Sm5	Sm5	Predicted class	Actual Class
3.7	3.3	3.6	3.2	3.2	2.1	3	1
2.3	2.0	2.3	3.0	3.1	1.7	3	0
3.3	1.7	1.5	3.0	3.0	1.6	3	0

4. CONCLUSION

With accuracy above 70%, prediction GPA categories using NBC to aid the process of guardianship is worth enough to be considered, although the parameters of the process of balancing the data has not been performed optimization significantly.

Further study is needed to get better accuracy result and reduce false prediction in case of higher data prediction for the low actual data. Additional data such as the entrance examination score, student activities, student family profile, the level of economic can be more valuable to get the better prediction. Exploring on other classification method may produce better results

ACKNOWLEDGEMENTS

The authors would like to express their highest gratitude to Sisfo division on Institute of Technology Telkom for the willingness to help provide data for this study.

REFERENCES

[1] O. Edin and S. Mirza, “Data Mining Approach for Predicting Student Performance,” *Journal of Economic and Business*, vol.X, May 2012.

[2] H. Jiawei and K. Micheline, “Data Mining Concept and Technique,” 2nd edition, San Francisco, Morgan Kaufmann, 2006.

[3] C. Nites V, et al., ,”SMOTE : Synthetic Minority Over-sampling Technique,” *Journal of Artificial Intelligence Research,* vol. 16, pp 321-357, 2002.

[4] Machine Learning Group at University of Waikato, 2013 [cited 2013 Aug 15], available from : http://www.cs.waikato.ac.nz/ml/weka/

BIBLIOGRAPHY OF AUTHORS

	Veronikha Effendy received bachelor's degree in Institute of Technology Telkom and have been worked in informatic technology industry in Indonesia for five years. Now, she is studying in Postgraduate Faculty in Institute of Technology Telkom with concentration on Data Mining to get her Master Degree.
	Thee Houw Liong . Now he is a professor at Institute of Technology Harapan Bangsa, ITB and ST Inten, Bandung.