Journal of Telematics and Informatics (JTI)
Vol.5, No.1, March 2017, pp. 22~28
ISSN: 2303-3703

Student Academic Performance Prediction on Problem Based Learning Using Support Vector Machine and K-Nearest Neighbor

Badieah Assegaf
Departement of Informatics Engineering, Sultan Agung Islamic University, Indonesia
e-mail: badieah.assegaf@unissula.ac.id

Abstract

Academic evaluation is an important process to know how well the learning process was conducted and also one of the decisive factors that can determine the quality of the higher education institution. Though it usually curative, the preventive effort is needed by predicting the performance of the student before the semester begin. This effort aimed to reduce the failure rate of the students in certain subjects and make it easier for the PBL tutor to create appropiate learning strategies before the tutorial class begin. The purpose of this work is to find the best data mining technique to predict student academic performance on PBL system between two data mining classification algorithms. This work applied and compared the performance of the classifier models built from Support Vector Machine (SVM) and K-Nearest Neighbor (KNN). After preprocessed the dataset, the classifier models were developed and validated. The result shows that both algorithms were giving good accuracy by 97% and 95,52% respectively though SVM showing the best performance compared to KNN in F-Measure with 80%. The further deployment is needed to integrate the model with academic information system, so that academic evaluation can be easily done.

Keywords*: Student Academic Performance, Problem Based Learning, Support Vector Machine, K-Nearest Neighbour, Prediction*

1. Introduction

Problem Based Learning (PBL) is a learning methodology that encourages students to better understand the subject being studied. This learning system combines the basic knowledge and skill by positioning the student as a problem solver to the problems that will be faced by the student in the future [1]. In PBL system, students are directed to think of solving the problems based on a real case problem that is then discussed together with the peers in a small group discussion. The effectiveness of the PBL implementation depends on three main areas: clinical cases discussed, the performance of the tutor and the performance of the students [2]. So with this learning model, if the student performance is the output indicating the effectiveness of the learning process, then the collaboration of these three factors will be the main determinant of the success of PBL learning process. The academic evaluation process should also be focused on these factors.

Academic evaluation is one of the essential things applied by universities to know the quality of the learning process. Academic evaluation is also one of the decisive factors assessed in the accreditation process. Implementation of academic evaluation is usually conducted at the end of each semester which results are then used to improve academic quality in the following semesters. So that the evaluation is usually curative. Preventive efforts should then be undertaken to predict student performance early so that appropriate learning strategies can be formulated before the semester begins. Preventive efforts also aim to reduce the failure rate of students in certain subjects.

Currently, many techniques have been proposed to predict student academic performance, including data mining techniques. Data mining is now a popular technique used in education which is then referred to as Educational Data Mining (EDM). EDM is an area of research that integrates several fields of science conducted to develop a method for analyzing a large amount of data. The main purpose of EDM is to find hidden knowledge and pattern to improve student performance in learning [3]. Data mining approach to predict student academic

performance has been proposed in several studies [3]–[7]. However, with the various characteristics of each method, it is necessary to find the best method to predict student performance, especially on PBL system.

Preliminary research was proposed to predict student performance using Artificial Neural Networks (ANN) [4]. In the study, the predictive value were numeric values that caused difficulty for the predictive model to generate an output value to really close to its target value. A large number of epochs were required for generalizing data during the training phase, though theoretically high epoch in ANN is most likely to cause overfitting. Moreover, from the user perspective, predictive value with class label is more meaningful than numerical value. By using numerical value as an output, user still need to determine which student will pass or fail in certain subject. Classification will make it easier for users to immediately identify which students will fail or pass in the subject.

This work focuses on predicting academic performance of dentistry students in Sultan Agung Islamic University (Unissula) using Support Vector Machine(SVM) and K-Nearest Neighbor (KNN) algorithms. Both of the algorithms are widely used in various fields of study because of their excellent ability for classifying data. The dataset used in this experiment contained academic history stored in academic information system database from year 2009 to 2013. Result of this work is performance comparison from both classification algorithms. The result is aimed to find the best algorithm to predict academic performance in PBL system.

2. Research Method

This research is an experimental research that aims to find the best data mining algorithm to predict student performance on Adult and Elderly Diseases subjects using PBL learning method at Faculty of Dentistry Unissula, Indonesia. Using the classification method, students will be classified into "Passed" and "Failed" classes. The result of the research is a comparison of classification performance that calculated based on Confusion Matrix. The steps and analysis of the study were conducted based on procedures of Cross-Industry Standard Process Data Mining (CRISP-DM) [8].

2.1. Business Understanding

This stage is the initial stage of the research where the objectives and direction of the research identified. At this stage, further identification of the factors that affect student's learning performance and several data mining methods from previous studies were also studied.

2.2. Data Understanding

At this stage, data exploration was conducted for determining predictor variables. In this work, the datasets were derived from the academic information system owned by Unissula Dentistry Faculty. The datasets contain student academic history collected from 2009 to 2013 consisting of 7 predictor variables and 1 class variable. Datasets were taken from Adult and Elderly Diseases courses with a total of 303 datasets. The selected variables are presented and defined in Table 1.

Table 1. Research Variables

Variable Name	Description
Gender	Students gender
Age	Age of student when taking courses
Knowledge Score	The average score of knowledge of all subjects attempted
Skill Score	The average score of skill of all subjects attempted
CGPA	CGPA calculated through dividing the total amount of grade points earned by using the total quantity of credit hours attempted.
Group Heterogeneity	Heterogeneity of study groups
Label Class	Classes that identify students passing or failing in a course

2.3. Data Preparation

In the data preparation phase, there were two steps that were done: data cleaning and data preprocessing. Data cleaning was done to avoid the outlier, missing value, data redundancy, and misclassification. After going through the process of data cleaning, data were collected and then transformed into a form that could be processed by machine learning. For numerical

variables, the min-max normalization formula [8] was used. While the categorical variable, as in the gender variable, uses a binary value.

$$X^* = \frac{X - \min(X)}{range(X)} = \frac{X - \min(X)}{\max(X) - \min(X)} \quad (1)$$

After going preprocessing the data, the dataset was then separated by percentage: the 80% of the dataset was used for the training dataset, while the remaining 20% was used as the testing dataset.

2.4. Modeling

The classification algorithm experimented in this research were Support Vector Machine (SVM) and K-Nearest Neighbor algorithm. The SVM algorithm was selected because of its excellent characteristics against small datasets. While KNN was selected because of its simple computation though has a very good ability to classify data.

2.4.1. Methodology for Supervised Modeling

Supervised learning is a methodology for building a data mining model and then evaluating the model. If we want to predict the value of the target variable based on the predictor variables, then the data mining algorithm requires a large number of records containing complete information about every field including the target variable. So the first step in this methodology is to provide a set of training data in which the data contains the values of the target variable as well as the predictor variable. The data is then performed by the training process algorithm. The main purpose of this training process is so that the provisional model can generalize the data well to the new or future values that were not previously included in the training process. We do not expect provisional model to "memorize" the pattern of the data because it will be difficult for the model later to recognize the new data [8].

To see how well the provisional model is formed, the next step in this methodology is to test the provisional model using a test set. In the test set, the target variable is temporarily hidden and then the test set is classified according to the pattern learned in the training process. The result of classification is then compared with the target variable that had been hidden and evaluating its performance.

2.4.2. Support Vector Machine (SVM)

Support Vector Machine (SVM) is one of the classification method that is widely developed and applied in various research fields. Rooted in statistical learning methods, SVM is said to provide better results compared with other classification methods and can also work on high-dimensional data sets. In contrast to ANN model that will use the entire training data, SVM only use a small number of training data that will be selected to contribute to the classification model, that is called support vector. This is an advantage of SVM because not all training data will be seen to be involved in every iteration of the training.

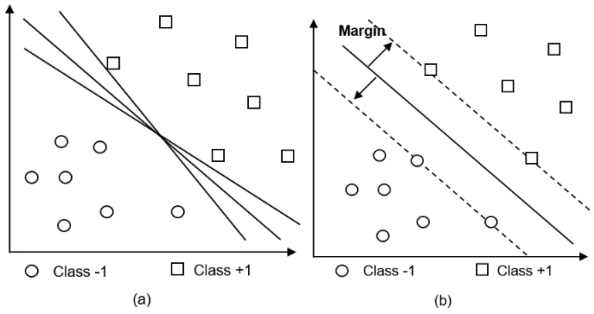

Figure 1. Decision boundary (a) Possible decision boundaries (b) Decision boundary with maximized margin

The basic concept of SVM is to find a hyperplane with a maximum margin value as shown in Figure 1. In Figure 1 it appears that there are many possible hyperplanes on the data set, but only hyperplane with maximum margin that will give better generalization on the classification method. Margin is the distance between the hyperplane and the closest data from each class.

Basically SVM can be used to solve linear and non linear problems. For data which has non linear distribution, kernel approach will be used in the initial features in the dataset. Kernel can be defined as a function that maps the data features from the (low) initial dimension to other features in higher dimension[9].This distinguishes SVM from other classification methods that typically reduce the initial dimension to simplify the computational process and obtain high accuracy values.

The SVM kernel commonly used on various classification engines is defined in Table 2.

Table 2. SVM Kernel Function [9]

Kernel Name	Kernel Function
Linear	$K(x,y) = x.y$
Polinomial	$K(x,y) = (x.y + c)^d$
Gaussian RBF	$K(x,y) = exp\left(\frac{-\|x-y\|^2}{2.\sigma^2}\right)$
Sigmoid (Tangen Hiperbolik)	$K(x,y) = \tanh(\sigma(x,y) + c)$
Invers Multiquadratic	$K(x,y) = \frac{1}{\sqrt{\|x-y\|^2 + c^2}}$

In the kernel formula above (Table 2), x and y are the data pairs from the training set, while $\sigma, c, d > 0$ are constant values.

2.4.3. K-Nearest Neighbor (KNN)

KNN is part of Nearest Neighbor Rule (NNR). Rule on NNR does not indicate that this method is a rule-based classification method, but a rule for determining the nearest neighbor based on statistical rules. NNR is also known as lazy learner because there is no learning process (from data) but learning from neighboring data (nearest) directly by the time of classification [10].

KNN works by searching for a number of k data objects or patterns (of all existing training data) closest to the input pattern, then selecting the class with the largest number of patterns among the k patterns.

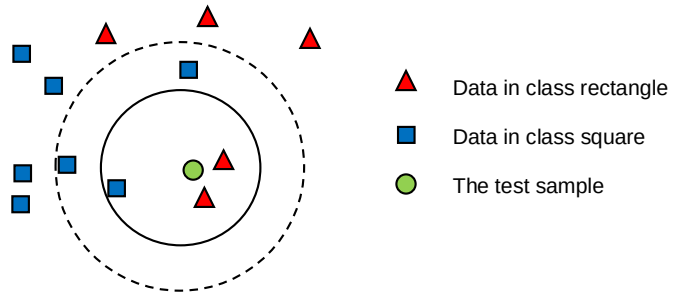

Figure 2. Classification of two classes using KNN

Simply put, KNN classifies the pattern by voting, as illustrated in Figure 2. The determination of the nearest pattern k is based on the size of the distance, similarity or dissimilarity, depending on the type of attribute. The KNN algorithm is described in Table 3.

Table 3. KNN Algorithm [10]

	Steps of KNN Algorithms
1.	For each training pattern $< x, f(x) >$, add the pattern to the list of training patterns
2.	For an input pattern x_q:
	• Suppose $x_1, x_2, \ldots\ldots, x_k$ is k pattern that has the closest distance (neighbor) with x_q.
	• Return the class that has the most number of patterns among the k pattern as the decision class.

2.5. Evaluation

To evaluate the performance of the classification model, several attempts were made to measure accuracy, recall, precision, Error Rate and F-Measure.The measurement is calculated based on the value obtained from the confusion matrix. Confusion matrix is very useful for analyzing the classifier's quality in recognizing the tuples of the existing class.

In confusion matrix, four values are obtained from the classification result, ie True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) values.TP and FN state that the classifier recognizes the tuple correctly, meaning that the positive tuple is recognized as the positive tuple and the negative tuple is recognized as the negative tuple. Instead, FP and FN state that the classifier is wrong in recognizing the tuples. These values will then be used to calculate accuracy, recall, precision, Error Rate and F-Measure.

$$Accuracy = \frac{TP+TN}{P+N}\ x\ 100\% \quad (2)$$

$$Recall = \frac{TP}{P}\ x\ 100\% \quad (3)$$

$$Precision = \frac{TP}{P}\ x\ 100\% \quad (4)$$

$$F - Measure = \frac{2\ x\ precision\ x\ recall}{precision+recall}\ x\ 100\% \quad (5)$$

$$Error\ Rate = \frac{FP+FN}{P+N}\ x\ 100\% \quad (6)$$

Where :

$$P = TP + FN \quad (7)$$

$$N = FP + TN \quad (8)$$

2.6. Deployment

This step is meant to explain the finding from this research and to plan the further deployment from the model in the future. However, the detail result in this paper will be explained entirely in the next section (Result and Analysis) and the future deployment will be mentioned in Conclusion section.

3. Results and Analysis

In this section, the results obtained from this experiment will be explained which then conducted further analysis. In this work, comparison of two classifier including SVM and KNN will be conduct. After the dataset was split into two parts (80% and 20% respectively), data with the largest percentage was used as a training set (236 data) and the remaining 20% (67 data) was used for testing set. The training process was then performed using the training data. This process is aimed to train the classifier model to recognize the data patterns that exist in the training set. Training results were then validated with new data that has not been included in the training process, using testing set. This validation process was performed to test the classifier model whether it can recognize the data pattern on the testing set and classify it accurately.

After the classifier model was formed, the important question that arises is how well the classifier model has been built. To do so, we can evaluate it by forming a confusion matrix.

Confusion matrix is a table that is often used to describe the performance of the classification model against a number of test data.The results of confusion matrix are four values, namely True Positive (TP) value (number of positive value which is classified as positive value), True Negative (TN) value (number of negative value which is classified as negative value), False positive (FP) value (number of positive value which is incorrectly classified), False Negative (FN) value (number of negative value which incorrectly classified).

The four values generated by the confusion matrix were then used to calculate the accuracy, precision, recall, F-Measure and Error Rate values of each model. Formula (2) - (6) were used to calculate the performance of the models. Related result of these measurements are demonstrated in Table 3.

Table 3. Comparison of model accuracy in SVM and KNN using testing set

Algorithm	Accuracy	Precision	Recall	F-Measure	Error Rate
Support Vector Machine (SVM)	97%	100%	66.67%	80%	3%
K-Nearest Neighbor (KNN)	95.52%	80%	66.67%	72.73%	4.48%

The experiment using SVM, the training and validation process were done using polynomial kernel whereas in KNN the K value used was 11. It can be seen from Table 3 that the accuracy results obtained by SVM is greater than to KNN with accuracy value of both models are 97% and 95.52% respectively. These results indicate that SVM is able to classify tuples in test data with 2% better than KNN. In addition, the positive test data is perfectly classified by SVM with 100% precision whereas KNN is only able to classify a positive dataset with 80% precision. However, the recall of both methods get the same value of 66.67% which is a good value because it is more than 0.5. To see the performance of the two methods in one value, we use the F-Measure calculation which is the harmonic average between precision and recall. The F-Measure value of SVM and KNN 80% and 72.73% respectively, indicating that, overall, SVM method is the best model compared to KNN to predict the academic value of PBL students using existing dataset.

The reason why the recall value of both methods could not achieve a high value is the number of datasets available to form the classifier is not so large and the data variations of the two classes are not balanced. In the existing dataset, the sample data classified as "failed" in the course is not much so it makes it difficult for the classifier to generalize the data patterns, especially data with negative value. This is why the TN values obtained in confusion matrix weren't much.

In machine learning, the dataset size has a large contribution to the needs of the learning process. However, it is quite difficult to say a dataset is large or not. A study by Ajiboye et al was conducted to determine the effect of dataset size on the performance of the predictive model [11] . In that study, it was found that the quantity of data partitioned for the learning process must have good representation of all the available dataset and sufficient enough to span through the input space.It also concluded that, from various experiments using different number of datasets, the learning process using a large number of datasets can result in more accurate predictions and more stable results [11].

So from this work, it can be concluded that, although the resulting recall value is not quite high, but this experiment proved that SVM is the best method to be able to generalize the exsisting dataset well compared to KNN.

4. Conclusion

In this work, two widely used classification algorithm were implemented on dentistry student academic dataset. The two algorithm both shows good performance to predict student academic performance using the exsisting dataset. The main reason for both algorithms weren't achieved higher value in recall is because the quality and the size of the dataset. Despite of that condition, both algorithms still give good performance in accuracy and precision. Based on the F-Measure achieved in the study, its concluded that SVM has the best performance to predict academic performance of dentistry students on PBL system compared to KNN. This experiment needs to be expanded to deploy further prediction model that can be integrated in academic information system, so that it could easily implemented for academic evaluation purpose.

References

[1] M. H. Bidokht and A. Assareh, “Life-long learners through problem-based and self directed learning,” *Procedia Computer Science*, vol. 3, pp. 1446–1453, 2011.
[2] N. Alajmi, “Factors That Influence Performance in a Problem-Based Learning Tutorial,” Bond University, 2014.
[3] H. Hamsa, S. Indiradevi, and J. J. Kizhakkethottam, “Student Academic Performance Prediction Model Using Decision Tree and Fuzzy Genetic Algorithm,” *Procedia Technology*, vol. 25, pp. 326–332, 2016.
[4] B. Badieah, R. Gernowo, and B. Surarso, “Metode Jaringan Syaraf Tiruan Untuk Prediksi Performa Mahasiswa Pada Pembelajaran Berbasis Problem Based Learning (PBL),” *Jurnal Sistem Informasi Bisnis*, vol. 6, no. 1, pp. 46–58, 2016.
[5] Z. Kovačić, “Early Prediction of Student Success: Mining Students Enrolment Data,” in *Proceedings of Informing Science & IT Education Conference*, 2010, pp. 647–665.
[6] V. O. Oladokun, A. T. Adebanjo, and O. E. Charles-Owaba, “Predicting students’ academic performance using artificial neural network: A case study of an engineering course,” *The Pacific Journal of Science and Technology*, vol. 9, no. 1, pp. 72–79, 2008.
[7] J. Laokietkul, N. Utakrit, and P. Meesad, “A Forecasting model to evaluate a freshman’s ability to succeed by using particular full-scaled class association rules (PFSCARs),” in *2009 International Association of Computer Science and Information Technology - Spring Conference, IACSIT-SC 2009*, 2009, pp. 40–44.
[8] D. T, Larose, *Discovering Knowledge in Data : An Introduction to Data Mining*. Canada: Wiley-Interscience, 2005.
[9] E. Prasetyo, *Data Mining Konsep dan Aplikasi Menggunakan MATLAB*. Yogyakarta? Andi Publisher, 2013.
[10] Suyanto, *Data Mining untuk Klasifikasi dan Klasterisasi Data*. Bandung: Penerbit Informatika, 2017.
[11] A. R. Ajiboye, R. Abdullah-Arshah, H. Qin, and H. Isah-Kebbe, “Evaluating The Effect of Dataset Size on Predictive Model Using Supervised Learning Technique,” *International Journal of Computer Systems & Software Engineering*, vol. 1, no. 1, pp. 75–84, Feb. 2015.