Jurnal Mantik, 6 (1) (2022) 582-586

Published by:Institute of Computer Science (IOCS)

Jurnal Mantik

Journal homepage: www.iocscience.org/ejournal/index.php/mantik/index

PREDICTION OF STUDENTS DROP OUT WITH SUPPORT VECTOR MACHINE ALGORITHM

Sartika Dewi Purba[1], Leliana Harahap[2], Jonas Franky R Panggabean[3]

AMIK Medicom, Jl.Darat No.74 Medan

E-mail: leliharahap05@gmail.com

ARTICLE INFO

Article history:
Received: Mar 9, 2022
Revised: Aprl 24, 2022
Accepted:May 29, 2022

Keywords:
Drop out prediction, kernel, support vector machine, unified modeling language.

ABSTRACT

The quality of a university can be seen from the high level of student success and the low level of student failure. As for the cause of student failure is the case of drop out. To overcome these problems, predictions are made using the support vector machine method. The Support Vector Machine tries to find the optimal hyperplane where the two pattern classes can be separated maximally, the parameters used in the Support Vector Machine are only kernel parameters in one C parameter which gives a penalty on randomly classified data points. In the Support Vector Machine the weights (w) and biases (b) are global optium solutions from quadratic programming so that just running once will result in a solution that will always be the same for the same kernel and parameter choices. Through the implementation of the support vector machine, it is expected to get the parameters of the Support Vector Machine that are used correctly to obtain the best margin in predicting students dropping out.

1. Introduction

Drop out or dismissal of student status is the process of revocation of student status for students, caused by certain things that have been determined by the university concerned. The high number of students dropping out at universities can be minimized by policies from universities to direct and prevent students from dropping out (Dekker, 2009) that detecting at-risk students in the early stages of education is very important to do to keep students from dropping out. This allows the education department to provide guidance to students in need. Therefore, it is necessary to study or predict student drop outs so that it can be used as useful information to estimate student drop out rates in the coming years and reduce student drop out rates. Prediction of student drop out can help the institution in making decisions. How to analyze the factors causing student drop out and get the best parameters from the method used to predict students who have the opportunity to drop out. Drop out prediction can be done by a series of processes to get knowledge or patterns from a data set called data mining. Data mining solves the problem by analyzing the data already in the database. Several data mining classification algorithms have been used to predict the behavior of students who have the potential to drop out including decision trees, neural networks, nave Bayes, instance-based learning, logistic regression and support vector machines.

Based on previous research conducted by (Hastuti, 2012) at Dian Nuswantoro University by performing a comparative analysis of algorithms for prediction of Non-Active students, it showed that the decision tree algorithm is the most accurate algorithm, however, the decision tree is not dominant over other algorithms. Based on the results of this study, logistic regression, decision tree, nave Bayes and neural networks are included in the category of excellent classification category. According to the research conducted by (Hidayat M. M., 2013) regarding the Prediction Analysis of student DO in educational data mining using artificial neural networks, it shows that the proposed model can be used to predicting potential drop out with an accuracy of possible drop out classification of 98.91% with the highest level of social behavior significant sensitivity of 4.737. The results of the sensitivity analysis show that the input variables for the most influential social parameters are the quality of interaction with friends and family relations variables, while the GPA variables for further study parameters and motivation for the most influential academic parameters are SKS and GPA.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

The method that can be used to predict is the Support Vector Machine (SVM). The Support Vector Machine was developed by Boser, Guyon, Vapnik in 1992. The Support Vector Machine (SVM) is a classification method that works by finding the hyperplane with the largest margin. Hyperplane is the dividing line of data between classes, while the margin is the distance between the hyperplane and the closest data in each class. The data closest to the hyperplane in each class is called the support vector.

Research (Sumarni, 2014) with the title Credit Risk Modeling with Support Vector Machine Approach This research uses the Support Vector Machine approach for classification applied in credit risk management. Classification is done to separate credit applications from two classes, namely good class and bad class. The data used is from the BPR Tasikmalaya bank from September 2005 to March 2010. Of the 602 data used, 402 data are used as training data and 200 testing data with a polynomial kernel function of degree 2. The accuracy obtained when compared with the classification results based on Bank Indonesia is 71.5% .

Support Vector Machine (SVM) is a semi-eager learner classification technique. Because apart from requiring a training process, the Support Vector Machine also stores a small portion of the training data for reuse during the prediction process. The Support Vector Machine provides a classification model whose solution is globally optimal, i.e. it always gives the same model and the solution with the maximum margin. Does not require selecting parameters, only specifying the kernel function that should be used (for the case of data whose class distributions cannot be separated linearly). The use of a kernel matrix has the advantage that the performance of data sets with large dimensions but the amount of data will be slightly faster because the size of the data in the new dimension is reduced a lot (Nugroho, 2008).

2. Methods

Classification is a technique that can be used to map inputs into discrete outputs called labels and categories. The category for an observation is predicted by using a mapping function. For example, the performance of a group of students may be classified as “passed” or “failed”. The classification technique consists of Support Vector Machine, Nave Bayes, Discriminant Analysis, Nearest Neighbor, Neural Network, and Nave Bayes (N. Mohammad Suhaimi, 2019). The method used for classification is Support Vector Machine (SVM). Several researchers have used SVM to predict student graduation times as done by (Hiryanto, 2017) as a classification method. The Support Vector Machine algorithm is used to determine the predictive model for the study period of students from the faculty of computer science. Based on the results obtained from the test, the Support Vector Machine algorithm provides a high accuracy value of 83.64% in the first experiment and 77% in the second experiment.

In research (Saifudin, 2018) used ten machine learning algorithms in determining the model to predict timely graduation for prospective students, one of which is SVM. Based on testing using data that is implemented in each machine learning algorithm, the SVM algorithm provides the highest accuracy value, which is 65% compared to other algorithms. This shows that the model generated from the SVM algorithm is able to predict timely graduation for prospective students. Research that also uses SVM to determine the best model in predicting student graduation is (Pratama, 2018) by using predetermined parameters. The design used in this research is entering data, normalizing data, conducting SVM training, doing SVM testing and getting classification results. The data used are 188 data using different amounts of training data, testing the effect of parameter values on Sequential Training SVM and the kernel. Based on the test, it was found that the effect of using the Gaussian RBF kernel on SVM gave the best results with an accuracy rate of 80.55%. Research on the application of the SVM algorithm to get a predictive model of student graduation on time was also carried out by (Utari, 2021). In this study, it is also compared with other machine learning algorithms, namely Naïve Bayes. The data used in this study such as GPA, profile data, origin of high school and place of residence. Based on the test results with these data, the SVM algorithm provides a better accuracy value than the Naïve Bayes algorithm, which is 69.15%. The prediction model for on-time student graduation generated by the SVM algorithm is better for use with this data.

Research conducted by (Utami, 2018) discusses the percentage of misclassification resulting from the algorithm in determining the student graduation model on time at FMIPA UNTAD. Using two algorithms to compare the error values generated from each algorithm, namely SVM and binary logistic regression. The test results show that the misclassification value generated by SVM is 16.84% while the misclassification value from binary logistic regression is 19.3%. The SVM algorithm has a smaller error value compared to binary logistic regression so that SVM can be used in the classification process for graduating students on time at FMIPA UNTAD.

Research conducted by (R.A. Permana, 2019) also uses SVM as an algorithm in generating models to predict student graduation by using electronic learning. Based on the results of tests that have been carried out on the use of log data from students on the algorithm SVM is known that the SVM algorithm is able to provide

an accuracy rate of 85.02%. This matter prove that the model generated from the SVM algorithm is able to predict graduation students on electronic learning.

Matlab stands for MATrix Laboratory, is a programming language developed by the Mathwork .Inc (http://www.mathworks.com). Matlab programming language is widely used for technical numerical calculations, computation, symbolic, visualization, graphics, analysis and mathematics, statistics, simulation modeling and GUI design (Prasetyo, 2012)). Guide or GUI builder is a graphical user interface (GUI) built with graphic objects such as buttons, text boxes, sliders, menus and others. Applications that use a GUI are generally easier to learn and use using a GUI are generally easier to learn and use because the person running them doesn't need to know the commands and how they work.

UML (Unified Modeling Language) is a notation used to create a visualization model of a system. The system contains information and functions, but is normally used to model computer systems. In object modeling to present an object-oriented system to others, it will be very difficult to do it in the form of programming language code (Yasin, 2012).

UML is referred to as a modeling language not a method. Modeling language is a model notation that is used to design quickly. UML is a standard language for writing software blueprints used for visualization, specification, creation and documentation of tools from software systems. UML is usually presented in diagrams or drawings that include classes and their attributes and operations, as well as relationships between classes. UML consists of many diagrams including use case diagrams, activity diagrams, class diagrams and sequence diagrams.

Action research methods, namely methods that aim to develop new information to address the world of work or other practical human needs. To find the basics and appropriate steps to take practical corrective action (Darmawan, 2013).

2.1 Data Collection

The data collection methods used in this study are:

a. Interview
b. Observation

2.2 Data Analysis

In this study using the Support Vector Machine algorithm by changing the data representation, data normalization, and transformation as well as the Support Vector Machine learning process by conducting a training process and data testing to find the hyperplane with the largest margin. The data consists of three study programs, namely management of informatics, computer engineering and computerized accounting.

a. Processing data
b. Data transformation and data selection
c. Variable selection
d. Training and testing

2.3 Prediction Process

To predict students dropping out is the Support Vector Machine (SUPPORT VECTOR MACHINE). The stages of the Support Vector Machine algorithm are as follows:

a. Carry out the transformation process
b. Define kernel function
c. Define kernel parameters and cost parameters
d. Selecting parameters
e. Calculating the accuracy of prediction

3. Results And Discussion

Support Vector Machine testing to predict student drop out is carried out in several stages, namely the process of sharing data, optimizing parameters and learning Support Vector Machine.

a. Data sharing. The student data set is divided into two parts, namely training data and testing data
b. Parameter optimization of Support Vector Machine, Kernel Support Vector Machine used in this research are Polynomial and RBF kernels. Kernel Polynomial has one parameter degree (p), degree is 1, 2, 3, 4, ..., n.
c. Learning Support Vector Machine, the performance evaluation method used is 10-flod cross validation. K-fold cross validation was chosen because it is more accurate in estimating performance. Cross-validation is an evaluation method that divides data into two segments, one is used for learning or training models and the other is used as a validation model.

The following is an interface implementation of the prototype that was built:

a. Training data menu The training data menu is used to carry out the training process.

b. Data Testing Menu The data testing menu is used to carry out the data testing process.
c. Prediction Data Menu Prediction data input is done by the user. The data used as input is new data that has never been used before.

Support Vector Machine (SVM) Parameters

The following are the results of the experiments in table 1 and table 2 which have been carried out with several kernel functions and include the cost (C) value and the specified range (k-fold) value to test each study program.

a. Informatics Management Study Program

TABLE 1
SUPPORT VECTOR MACHINE PREDICTION TEST 2008-2010

K-fold = 10					
Year	type kernel	parameter			accuracy
		C	P	ɣ	
2008	1	0.1	3	0.001	87.27%
2009	1	0.1	3	0.001	80.35%
2010	2	0.7	6	0.1	93.23%

b. Computer Engineering Study Program

TABLE 2
SUPPORT VECTOR MACHINE PREDICTION TEST 2008-2010

K-fold = 10					
Year	type kernel	Parameter			accuracy
		C	P	ɣ	
2008	2	0.1	2	0.001	97.27%
2009	1	0.3	3	0.1	89.45%
2010	2	0.5	4	0.1	93.43%

c. Computerized Accounting study program]

TABLE 3
SUPPORT VECTOR MACHINE PREDICTION TEST 2008-2010

K-fold = 10					
Year	type kernel	parameter			accuracy
		C	P	ɣ	
2008	2	0.1	2	0.001	97.27%
2009	1	0.2	2	0.1	91.05%
2010	1	0.7	5	0.1	96.32%

The results of the evaluation with the SVM algorithm need to measure the level of performance accuracy and error using student data in 2011 with a total of 200 data records with a polynomial kernel type, k-fold = 10 cost value 0.5 gamma 0.001 and degree = 8.

For measurement of accuracy can use the equation:

$$accuracy = \frac{number\ of\ correct\ predictions}{total\ predictions}$$

$$accuracy = \frac{182}{200} = 92$$

While the measurement of the error rate (error rate) using the equation:

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

$$error = \frac{number\ of\ error\ predictions}{total\ predictions}$$

$$error = \frac{12}{200} = 0.09$$

4. Conclusion

Based on the results of the research that has been done, it can be concluded that:
Using the Support Vector Machine (SVM) Algorithm can predict students who will be dropped out with variables that influence each other academically and non-academicly. The results of the test with data from students of the informatics management study program, with the best parameter with a value of k-fold = 10, polynomial kernel type, error value = 0.080, cost value = 0.1, degree = 5 and gamma = 0.1. The results of the test using computer engineering student data, with the best parameter with a value of k-fold = 10, polynomial kernel type, error value = 0.093, cost value = 0.2, degree = 7 and gamma = 0.1. The results of the test with student data from the computerized accounting study program, with the best parameter with a value of k-fold = 10, polynomial kernel type, error value = 0.091, cost value = 0.1, degree = 7 and gamma = 0.1.

References

[1] Darmawan, D. (2013). *Metode Penelitian Kuantitatif.* Bandung: Remaja Rosdakarya.

[2] Dekker, G. (2009). Prediction student drop out: A case study, USA, Academic Press. *2nd international Conference On Educational Data Mining.* Spain: Cordoba.

[3] Hastuti, K. (2012). Analisis Komparasi Algoritma Klasifikasi Data Mining untuk Prediksi Mahasiswa Non Aktif. *Seminar Nasional Teknologi Informasi & Komunikasi Terapan* .

[4] Hidayat M. M., P. D. (2013). Analisis Prdiksi DO Mahasiswa dalam Educational Data Mining menggunakan JaringanSyaraf Tiruan. *Jurnal IPTEK* .

[5] Hiryanto, T. H. (2017). Predicting And Analyzing The Student's Length of Studi-Time Using Support Vector Machine. *ComTech Comput. Math. Eng. Appl*, 107–114.

[6] N. Mohammad Suhaimi, S. A. (2019). Review on Predicting Students' Graduation Time Using Machine Learning Algorithms. *Int. J. Mod. Educ. Comput*, 1-13.

[7] Nugroho. (2008). Support Vector Machine: Paradigma Baru dalam softcomputing dan Aplikasinya. *Konferensi Nasional Sistem & Informatika.* Bali.

[8] Prasetyo, T. W. (2012). *Analisis dan Desain Sistem Kontrol dengan MATLAB.* Yogyakarta: Andi.

[9] Pratama, R. C. (2018). Implementasi Algoritme Support Vector Machine (SVM) untuk Prediksi Ketepatan Waktu Kelulusan Mahasiswa. *J. Pengemb. Teknol. Inf. dan Ilmu Komputer*, 1704–1708.

[10] R.A. Permana, S. S. (2019). Metode Support Vector Machine Sebagai Penentu Kelulusan Mahasiswa. *Jurnal Khatulistiwa Informatika*, 9.

[11] Saifudin, A. (2018). Metode Data Mining Untuk Seleksi Calon Mahasiswa Pada Penerimaan Mahasiswa Baru di Universitas Pamulang. *J. Teknologi*, 25-36.

[12] Sumarni, A. (2014). *Kalsifikasi Data Nap (Nota Analisis Pembiayaan) untuk Prediksi Tingkat Keamanan Pemberian Kredikt (Studi Kasus : Bank Syariah Mandiri Cabang Luwuk Sulawesi Tengah).* Yogyakarta: Ilmu Komputer, Universitas Gajah Mada.

[13] Utami, I. T. (2018). Perbandingan Kinerja Klasifikasi Support Vector Machine (SVM) Dan Regresi Logistik Biner Dalam Mengklasifikasikan Ketepatan Waktu Kelulusan Mahasiswa Fmipa Untad. *J. Ilm. Mat. Dan Terap*, 256-267.

[14] Utari, A. K. (2021). Predicting Patterns of Student Graduation Rates Using Naïve Bayes Classifier and Support Vector Machine. *AIP Conf. Proc.*

[15] Yasin, V. (2012). *Rekayasa Perangkat Lunak Berorientasi Objek Pemodelan, Arsitektur, dan Perancangan (Modeling, Architecture and Desaign).* Mitra Wacana Media.