Predicting Student Placement Class using Data Mining

Oktariani Nurul Pratiwi
Teknik Informatika,
Universitas Widyatama,
Jl. Cikutra 204A, Bandung
oktarianinurul@gmail.com

***Abstract*—All students in first grade of senior high school in Indonesia have to pass the step called placement class. It divide into Science, Social or Literature class. In traditional method, the placement class process conducted by teachers. But, it needed much time to decide the right class for students. The proposed is using the Knowledge Discovery and Data Mining (KDD). Which is the placement class process using the classification method. In the first experiment classified instances 84.2%. The second experiment use the same data and attributes, give the best percentage of accuracy as 92,1%. The best result are using Naive Bayes and SMO. Hope in the future, it can be the solution to help teacher decide the placement class.**

Keywords—data mining, education, placement class, classification

I. INTRODUCTION

In Indonesia, based on Indonesian government regulations no. 17 in 2010 [3], there is a process to divide students in Senior High School into some majors, namely Science, Social, and Literature class or other courses that required. In traditional method, the placement process conducted by teacher.

The problem of traditional placement class process are teachers have to identify and find useful information in large databases manually and it is a difficult task [1]. A very promising solution to facilitate the placement class process is using knowledge discovery in databases techniques or data mining in education, called educational data mining, EDM [2].

But, educational data and problems have some special characteristics that require the issue of mining to be treated in a different way [4]. Hence, it need to determine the best method for the placement process of students.

There are several important differences and/or advantages between applying data mining with respect to only using statistical models [5]:

1) Data mining is a broad process that consists of several stages and includes many techniques, including statistics. The knowledge discovery process includes the steps of pre-processing, the application of data mining techniques, evaluation and interpretation of results.

2) Statistical techniques (data analysis) is often used as a quality criterion verisimilitude of the data given the model. DM uses a direct approach.

3) In statistics, the search is usually done by modeling based on a hill climbing algorithm in combination with a verisimilitude ratio test-based hypothesis. DM is often used a meta-heuristics search.

4) DM is aimed at working with very large amounts of data (millions and billions). The statistics does not usually work well in large databases with high dimensionality.

In this research, the experiment use the classification method to predict the placement class. The classification method include in supervised method. The supervised methods are able to perform with relatively high accuracy when making coarse grained distinctions in topics [6].

The classification have some algorithms. In this experiment use 6 algorithms to determine the appropriate algorithm method for predicting placement class of students.

II. METHOD

The method proposed in this paper for placement belongs to the process of Knowledge Discovery and Data Mining (KDD) [5].

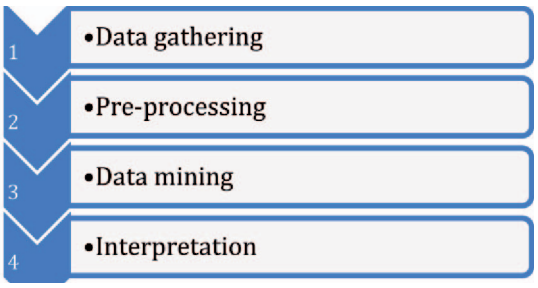

FIGURE 1. METHOD

The main stages of method are:

1. Data gathering, this stage consists of collecting data. For sample data, we collected data from 23 senior high school by accessing

978-1-4673-6355-6/13/$31.00 ©2013 IEEE

http://pdss.snmptn.ac.id/ page. Total data of students is 314 students in 2 semesters.

2. Pre-processing, at this stage we prepared data so it can be applied to data mining techniques. Preprocess data are data cleaning, data integration and transformation.
3. Data mining, at this stage we used weka tools (http://www.cs.waikato.ac.nz/ml/weka/). In order to get the appropriate algorithm to predict the placement class of students, this experiment use some classification algorithms namely J48, SimpleCart, Kstar, Naive Bayes, SMO, oneR.
4. Interpretation, at this stage we analysis of the result. Comparing algorithm in order to provide the best prediction accuracy.

In this experiment, we used 6 algorithms, there are:

1. J48, The algorithms J48 is based on the ID3 algorithm developed by Ross Quinlan, with additional features to address problems that ID3 find difficult to deal [9]. It is one of the decision tree induction has been studied in the areas of pattern recognition and machine learning
2. Naive Bayes, survey on classification algorithms [7] found that simple Bayesian known as Naive Bayesian classifier can be compared in performance with decision trees and other classifiers and exhibit high accuracy.
3. SimpleCART (Classification and Regression Trees) is classification method which uses historical data to construct decision trees [11].
4. Kstar is an instance-based classifier, that is the class of a test instance is based upon the class of those training instances similar to it, as determined by some similarity function. It differs from other instance-based learners in that it uses an entropy-based distance function [12].
5. The new SVM learning algorithm is called Sequential Minimal Optimization (or SMO). SMO is a simple algorithm that can quickly solve the SVM Quadratic Programming (QP) problem without any extra matrix storage and without using numerical QP optimization steps at all [13].
6. OneR learns a one-level decision tree. The algorithm finds wights of discrete attributes basing on very simple association rules involving only one attribute in condition part [14].

All of the results of the classification algorithms compare each other to get the appropriate with this problems.

III. DATA GATHERING

Gathered data is derived from scores of students in 23 senior high school for 2 semesters in the academic year 2010/2011. Subjects are Religious Education, Citizenship Education, Indonesian, English, Mathematics, Physics, Chemistry, Biology, History, Geography, Economics, Sociology, Cultural Arts, Physical Education, Sport and Health, Information and Communication Technology, Skills / Foreign Languages. The data classified into 2 classes, Science (IPA) and Social (IPS). It because there are only 2 classes in that school.

Data divided into 2 parts, training dataset and test dataset. Data training is the result of the classification data in the process of placement class of students by teacher manually. The training dataset implemented to build up a model. Data testing implemented to validate the model that have been built.

The sample of data can be see in Figure 2. There are the sample of 3 students, 2 of them classified into class Social (IPS) and 1 of them to class Science (IPA).

```
'65','70','68','64','60','65','65','65','75','75','69','73','75','83','81','70','75','75','
72','75','65','70','67','68','82','72','86','85','86','70','77','70','IPS'

'65','70','68','64','64','66','65','71','71','75','67','68','73','93','75','67','70','72','
70','70','75','75','70','80','70','72','78','80','75','70','77','75','IPA'

'65','70','68','64','60','65','71','65','69','75','67','76','72','98','79','68','76','73','
70','65','65','70','77','70','84','73','85','70','84','86','77','71','IPS'
```

FIGURE 2. EXAMPLES OF DATA

IV. DATA PRE-PROCESSING

Data pre-processing is one of the most important steps in KDD. It is aim is to make the chosen dataset as 'clean' as possible for the later mining step [7, 8]. In this experiment, the data preprocess are data cleaning, data integration and transformation.

Data cleaning tasks are fill in missing value, identify outliers and smooth out noisy data, correct inconsistent data. The missing data caused by the student is a new student, so there are no data value.

The next step is integrating data. Combined data from multiple source files into a coherent database.

After that, transformating data. Every missing value replaced by question mark '?'.

V. DATA MINING AND EXPERIMENTATION

The experiments used 6 classification algorithms in order to try obtain the highest classification accuracy. The algorithms are J48, SimpleCart, Kstar, SMO, Naive Bayes, OneR.

Calculation accuracy used 10 cross-validation method. Cross-validation method is a statistical algorithms by dividing data into two segments: one used to learn or train a model and the the other used to validate the model [15].

The first experiment use all subjects. This is the result of the experiment in Table I.

From the result, using J48 and OneR give the best percentage of correctly classified instances 79.61% and 78.66%. Kstar and NaiveBayes give the worst percentage of correctly classified instances 69.74% and 76.75%.

TABLE I. FIRST EXPERIMENT CLASSIFICATION ACCURACY

	Correctly Classified Instances	Incorrectly Classified Instances
J48	79.61%	20.38%
SimpleCart	78.34%	21.65%
Kstar	69.74 %	30.25 %
SMO	77.38%	22.61%
NaiveBayes	76.75%	23.24 %
OneR	78.66%	21.33 %

The second experiment scenario by remove some of the subject, such as Religious Education, Skills / Foreign Languages, Physical Education. It turns out there was an increase in terms of the accuracy of the data. It can be seen from Table II.

TABLE II. SECOND EXPERIMENT CLASSIFICATION ACCURACY

	Correctly Classified Instances	Incorrectly Classified Instances
J48	79.61%	20.38%
SimpleCart	79.61%	20.38%
Kstar	74.52%	25.47%
SMO	75.15%	24.84%
NaiveBayes	78.34%	21.65%
OneR	78.66%	21.33%

From this experiment, we can see that the best classification accuracy used J48 and SimpleCart. It give same percentage, 79.61% correctly classified instances. Some algorithms give some improvements. It prove that filtered the subject can increase to the accuracy value. This improvement could be by removing some subjects. For the result of comparing, we can see in the Figure 3.

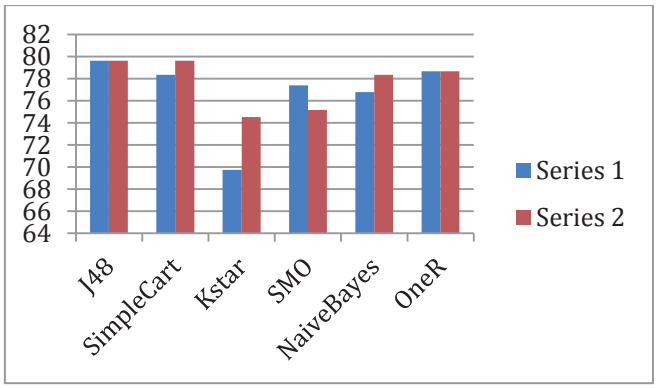

FIGURE 3. RESULT OF 6 ALGORITHMS

Precision is the fraction of retrieved documents precision that are relevant. Recall is the fraction of relevant documents that are retrieved [16]. In Table III and Table IV, we can see the value of Precision and Recall of the data.

TABLE III. PRECISION

	First Experiment		Second Experiment	
	IPA	IPS	IPA	IPS
J48	0,773	0.881	0.773	0.881
SimpleCart	0.823	0.712	0.826	0.738
Kstar	0.81	0.568	0.815	0.64
SMO	0.799	0.729	0.795	0.67
NaiveBayes	0.821	0.678	0.832	0.701
OneR	0.76	0.9	0.76	0.9

TABLE IV. RECALL

	First Experiment		Second Experiment	
	IPA	IPS	IPA	IPS
J48	0.96	0.513	0.96	0.513
SimpleCart	0.839	0.687	0.859	0.687
Kstar	0.683	0.722	0.774	0.696
SMO	0.859	0.626	0.819	0.635
NaiveBayes	0.809	0.696	0.824	0.713
OneR	0.97	0.47	0.97	0.47

VI. CONCLUSION

Predicting student placement class manually by teachers is a difficult tasks. To resolve this problems, we can use data mining to help predict the classification. Before implement, we need to know the best algorithm that appropriate with this data. Because of that, we did this experiment.

This paper propose 6 algorithms that can use to classified the student's data. From the first experiment, that use all of attributes, show that the best classification methods are J48, with the percentage accuracy of both experiment are 79.61%.

This experiment shown that the data balancing can also be very useful for improving accuracy. Hence, it necessary process of finding the best attributes in classification process.

Furthermore the algorithm will be into the system so it can be use to assist teacher predicting the placement class of students. The system also can help students to give early information about their major.

REFERENCES

[1] M. N. Quadril and N. V. Kalyankar, "Drop out feature of student data for academic performance using decision tree techniques," *Global J. Comput. Sci. Technol.,* vol. 10, pp. 2-5, Feb. 2010.

[2] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Expert Syst. Appl.,* vol. 33, no.1, pp. 135-146, 2007.

[3] Peraturan Pemerintah Republik Indonesia nomor 17 tahun 2010 Tentang Pengelolaan dan Penyelenggaraan Pendidikan.

[4] C. Romero and S. Ventura, "Educational Data Mining: A Review of the State of the Art," *IEEE Transactions on Systems, Man, and Cybernetics-PART C: Applications and Reviews*, Vol. 40, No. 6, November 2010.

[5] C. M. Vera, C. R. Morales and S. V. Soto, "Predicting School Failure and Dropout by Using Data Mining Techniques," *IEEE Journal of Latin-American Learning Technologie,* Vol. 8, No. 1, February 2013.

[6] A. Padhye. "Comparing Supervised and Unsupervised Classification of Messages in the Enron Email Corpus," *A Thesis Submitted to The Faculty of The Graduate School of The University of Minnesota.* 2006.

[7] J. Han and M. Kamber, *Data mining: concepts and techniques.* Morgan Kaufmann, 2000.

[8] S. F. Crone, S. Lessmann, and R. Stahlbock, "The impact of preprocessing on data mining: an evaluation of classifier sensitivity in direct marketing," *European Journal of Operational Research,* vol. 173, no. 3, pp. 781-800, 2006.

[9] Quinlan, J. R., "C4.5: Programs for Machine Learning", San Mateo, CA: Morgan Kaufmann Publishers, 1993.

[10] N. Gayatri, S Nickolas, A. V. Reddy, R. Chitra, "Performance Analysis of Data Mining Algorithms for Software Quality Prediction," *2009 International Conference on Advances in Recent Technologies in Communication and Computing.*

[11] R. Timofeev, "Classification and Regression Trees (CART) Theory and Applications", *A Master Thesis,* Humboldt University, Berlin, 2004.

[12] Pentaho, Kstar, http://wiki.pentaho.com/display/DATAMINING/KStar. Accessed: 24 April 2013.

[13] J. C. Platt, "Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines,"

[14] -, "OneR algorithm," http://artax.karlin.mff.cuni.cz/r-help/library/FSelector/html/oneR.html. Accessed: 24 April 2013.

[15] -, "Cross-Validation," http://www.cse.iitb.ac.in/~tarung/smt/papers_ppt/ency-cross-validation.pdf. Accessed: 24 April 2013.

[16] C. D. Manning, P. Raghavan, H. Schutze, "An Introduction to Information Retrieval," Cambridge UP, 2009.