Performance Comparison and Optimized Algorithm Classification

Dedi Saputra¹ *, Weishky Steven Dharmawan², Mochamad Wahyudi³, Windi Irmayani⁴, Juniato Sidauruk⁵, Martias⁶

¹,³,⁶Sistem Informasi, Universitas Bina Sarana Informatika, Indonesia
²,⁴Sistem Informasi Akuntansi, Universitas Bina Sarana Informatika, Indonesia
⁵Bahasa Inggris, Universitas Bina Sarana Informatika, Indonesia
E-mail: dedi.dst@bsi.ac.id

Abstract. The current development of technology is quite rapidly not disengaged in a large data processor covering of all areas such as information technology, computer science, medicine, finance and other. This brings a large computing effect in identifying the processing of data. In data analysis for very large data, data processing is very much needed, in this study the authors propose data mining method as a solution to a large data processing problem, data mining is divided into several techniques including classification method techniques that aims to classify large amounts of data to be relevant data information. In this study the authors compared 5 algorithms in the classification method to get better performance in classification problems. Researchers analyze and test 5 Algorithm classifications with 4 different datasets as a tool in the problem of large data classification. The results of this research show the method SVM is much better to be used 4 comparison methods in calculating the value of AUC by using 4 datasets of UCI Repository. The LSVT Dataset shows the highest AUC value with 0.973, Ionsphere 0.887, Sonar 0.897, Heartstatlog 0.868.

1. Introduction

Information technology provides the convenience of people in obtaining large amounts of data even tends to be very excessive which results in the data information that is obtained invalid, consequently the utilization of the accumulated data becomes not optimal in processing. This carries a huge computing influence in identifying the processing of a data. Data mining is currently widely used in a research related to the processing of large data to obtain accurate information. The literature on Data Mining can be seen for example in [1], [2], [3]. Several studies have been conducted with a comparison of 5 methods of classification algorithms for 1 dataset in the marketing of banking services products [4].

In this study we focus on classification techniques in comparing the performance of 5 classification algorithms to determine which algorithm is better in classifying large data processing. In our opinion the use of data mining is very effective, so that it can predict the level of success in marketing the service product. The evaluation process uses Cross Validation, Confusion Matrix, Roc Curve and T-test, to find out the most accurate data mining classification algorithm in predicting the success of telemarketing in marketing bank service products from trials, the Neural Network algorithm is more accurate with 89.71% accuracy, with an AUC value of 0.872. Our research will apply optimization and comparison of classification algorithms in
data mining. Comparison of classification algorithm testing consists of C45, Navie Bayes (NB), Support Vector Machine (SVM), k-nearest neighbour (k-NN), Classification And Regression Trees (CART), which later will be improved which classification performance is better in data mining processing. Using 4 datasets as a test tool in determining data mining performance, especially in classification algorithms. From the results of experiments that have been done, a new knowledge can be generated in the problem of classifying a large enough dataset, by comparing 5 data mining classification algorithms with 5 different datasets.

2. Methods
In this study, literature study was used as the cornerstone of research in order to know the state of the art about research on the performance ratio in applying classification algorithm. Next, try to analyze and first compare the 5 classification algorithms with 4 different datasets as a test tool to determine which is better in the data classification problem.

2.1. Definition and Function of Data Mining
Definition of Data Mining or Knowledge Discovery in Database (KKD) See [5]. Data Mining analyses data using tools to find patterns and rules in the dataset. The software is tasked to find patterns by identifying rules and features on the data. Tool Data Mining is expected to be able to recognize this pattern in data with minimal input from the use.

2.2. Classification Method
Classification is how to place specific objects into a group based on their nature.[6]. This method aims to study the different functions that describe each of the data selected into one of the predefined groups of classes.

2.3. C.45
C.45 is a method and prediction that is very strong and very widely used. C4.5 used information gain to select the attributes that will be used for object separation [7]. The attribute that has the highest information gain compared to other attributes relative to the set y in data, is chosen to do the solving. In this algorithm, the selection of which attributes will occupy a node is done by calculating the information entropy and looking for the minimum value. The choice of attributes in this algorithm is based on the assumption that the complexity of the decision tree is very closely related to the amount of information provided by the attribute values.

2.4. Naïve Bayes
The technique of Naïve Bayes (NB) is one of the simplest forms of Bayesian tissue for classification. A Bayes network can be viewed partly directed as a table with a combined probability distribution of more than one discrete set and a stochastic variable [8].

2.5. Support Vector Machine
Support Vector Machine (SVM) is a classification method for finding the best hyper-plane values that are capable of finding optimal global solutions. So the accuracy value is not easy to change [9]. In other studies, Support Vector Machine (SVM) is learning that leads to quadratic programming with linear constraints [10].

2.6. K-Nearest Neighbor
The k-nearest neighbor algorithm (k-NN or KNN) is a method for classifying objects based on learning data that is the closest distance to the object. Research on this algorithm mostly discusses how to select and give weight to features, so that the classification performance becomes
better, according to [7], KNN is also an example of lazy learning technique, which is a technique that waits until the query (query) comes to be the same as training data.

2.7. Classification and Regression Trees (CART)
Classification and Regression Trees (CART) is one of the methods or algorithms of the decision tree technique. Response variables in this study are categorized on a scale, so the method to be used is the tree classification method [11].

2.8. Confusion Matrix
Confusion matrix provides decisions obtained in training and testing. Confusion matrix provides an assessment of performance classification based on the object correctly or incorrectly [6]. Confusion matrix contains actual and predicted information on the classification system. Table 1 is an explanation of the configuration matrix.

| Classification Predicted Class |
|--------------------------------|
| Observed Class=Yes  | Class=Yes | Class=No |
| Class=Yes          | A         | B        |
|                    | (True Positive=TP) | (False Negative=NP) |
| Class=No           | C         | D        |
|                    | (False Positive=FP) | (True Negative=TN) |

Notes:
True Positive (TP) = positive proportion in data sets that are classified positively.
True Negative (TN) = negative proportion in the data set that is classified as negative.
False Positive (FP) = negative proportion in data sets that are classified as potential.
False Negative (FN) = negative proportion in the data set that is classified as negative.

2.9. K-Fold Cross Validation
One alternative approach to "train and test" which is often adopted in some cases (and several others regardless of size) is called k-fold cross validation, by testing the amount of error in the data test [12]. In this study, the value of k used amounted to 10 or 10-fold Cross validation. Each trial will use one testing data and the k-1 part will become training data, then the testing data will be exchanged for one training data so that for each experiment different testing data will be obtained. Training data is data that will be used in learning while testing data is data that has never been used as learning and will function as data testing the truth or accuracy of learning outcomes [13]. Figure 1 is an illustration of the stages of using cross validation.

2.10. Receiver Operating Characteristic (ROC)
The Receiver Operating Characteristic or commonly called AUC value is a visual tool that is useful for comparing two classification models. Here’s the look of two types of ROC curves (discrete and continuous).

Points above the diagonal line are good classification results, while points below the diagonal line are bad classification results. It can be concluded that, one point on the ROC curve is better than the other if the direction of the transverse line from lower left to upper right in the
graph. For the level of insurance the AUC value in the classification of data mining is divided into five groups [6], namely: (a) 0.90 - 1.00 = excellent classification; (b) 0.80 - 0.90 = good classification; (c) 0.70 - 0.80 = fair classification; (d) 0.60 - 0.70 = poor classification; (e) 0.50 - 0.60 = failure.

3. Results And Discussion

This study applies 5 classification algorithms in determining the performance of the classification performance with 4 datasets used as measurement tools. The results obtained from the calculations of this study by showing better performance in classification problems. Figure 3 is a pattern of the results of testing using the Rapid Miner Tools.

4. After the classification process of all algorithms and datasets that were tested the results of analysts testing the calculation of the value of the AUC (Area Under Curve) of the 5 classification algorithm methods are summarized in the table below:

For data mining classification, AUC values can be divided into several groups (Gorunescu Florin, 2011): (a) 0.90-1.00 = excellent classification; (b) 0.80-0.90 = good classification; (c) 0.70-0.80 = fair classification; (d) 0.60-0.70 = poor classification; (e) 0.50-0.60 = failure. The results of testing the five methods of the classification algorithm are described in Figure 4.
Table 2. Results of AUC (Area Under Curve)

| Dataset       | C 45 | NB | SVM | K-NN | CART |
|---------------|------|----|-----|------|------|
| LSVT          | 0.500| 0.325|0.973| 0.500| 0.500|
| Ionosphere    | 0.858| 0.878|0.887| 0.500| 0.842|
| Sonar         | 0.552| 0.883|0.897| 0.500| 0.782|
| Heartstatlog  | 0.706| 0.878|0.868| 0.500| 0.680|

Figure 4. Graph of Test Results 5 Classification Algorithm methods

The results of the test can be seen from the tables and graphs above with 5 classification algorithms by testing 4 datasets showing the performance of the Support Vector Machine (SVM) algorithm is superior to the other 4 classification algorithms in testing the value of AUC (Area Under Curve). In this study also displays the value of Accuracy, Sensitivity, and Specificity for the prediction of each class in the dataset being tested. The results of the calculation completion and summarized in the table below:

Based on the results of experiments and data analysis in this study, it can be obtained the highest Area Under Curve (AUC) value comparison in the classification algorithm test, the superiority of the SVM algorithm after parameter optimization with the testing phase. The trial result shows SVM on average has a better performance compared to other comparative methods such as C4.5, NB, k-NN and CART in the aspect of Area Under Curve (AUC) values and some Precision, Recall and F-Measure values. In this study Support Vector Machine (SVM) is used because it is known from the results of previous studies that in SVM as a single method has a very good generalization ability to solve problems even with limited samples. From these results it is known that the success of the Support Vector Machine (SVM) is greatly influenced by the selection of appropriate attributes. The more attributes and information used will result in the amount of time and cost being sacrificed, even reducing the higher level of accuracy and complexity. The results of 4 UCI Repository dataset experiments show good performance of the SVM method in the calculation of the AUC value which is quite high from the 4 comparison methods. The LSVT dataset shows the highest AUC value with 0.973, Ionosphere 0.887, Sonar 0.897, Heartstatlog 0.868.

4. Conclusion
In analyzing data for very large data, data processing is needed, in this study the authors propose a minimum data method as a solution in the problem of large data processing, data mining is divided into several techniques including classification method techniques that aim at classifying large enough data to become a relevant data information. In this study, the authors compared 5 algorithms in the classification method to get better performance in classification problems. The results of this study indicate that the SVM method is far better compared to the 4 comparative methods in calculating the AUC value using 4 UCI Repository datasets as a test.
Table 3. Results for Accuracy, Sensitivity, Specificity Calculation

| Dataset      | C.45 Accuracy | C.45 Sensitivity | C.45 Specificity | NB Accuracy | NB Sensitivity | NB Specificity |
|--------------|---------------|------------------|------------------|-------------|----------------|----------------|
| LSVT         | 0.620         | 0.222            | 0.850            | 0.540       |                | 0.287          |
| Ionosphere   | 0.911         | 0.977            | 0.757            | 0.854       | 0.930          | 0.676          |
| Sonar        | 0.602         | 0.568            | 0.640            | 0.797       | 0.972          | 0.605          |
| Heartstatlog | 0.742         | 0.758            | 0.724            | 0.833       | 0.896          | 0.760          |

| Dataset      | SVM Accuracy | SVM Sensitivity | SVM Specificity | KNN Accuracy | KNN Sensitivity | KNN Specificity |
|--------------|--------------|-----------------|-----------------|--------------|-----------------|-----------------|
| LSVT         | 0.836        | 0.555           | 0.993           | 0.544        | 0.211           | 0.731           |
| Ionosphere   | 0.835        | 0.9              | 0.633           | 0.854        | 0.957           | 0.614           |
| Sonar        | 0.731        | 0.836           | 0.630           | 0.733        | 0.677           | 0.795           |
| Heartstatlog | 0.841        | 0.728           | 0.920           | 0.503        | 0.534           | 0.468           |

| Dataset      | CART Accuracy | CART Sensitivity | Specificity |
|--------------|---------------|------------------|-------------|
| LSVT         | 0.488         | 0.466            | 0.50        |
| Ionosphere   | 0.860         | 0.965            | 0.614       |
| Sonar        | 0.704         | 0.750            | 0.655       |
| Heartstatlog | 0.848         | 0.893            | 0.796       |

The LSVT dataset shows the highest AUC value with 0.973, Ionosphere 0.887, Sonar 0.897, Heartstatlog 0.868. Researchers added descriptions for subsequent research by trying to use a much larger number of datasets and with more attributes, so the latest test results can be seen, and tried to merge using a future selection method such as artificial bee colony and others for the initial stages of the attribute screening.

References

[1] H. Jiawei, M. Kamber, J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. 2012.
[2] E. Turban, J. E. Aronson, and T. Liang, “Decision Support Systems and Intelligent Systems,” 2007.
[3] D. T. Larose, Data Mining Methods and Models. 2006.
[4] S. Dewi, “KOMPARASI 5 METODE ALGORITMA KLASIFIKASI DATA MINING PADA PREDIKSI KEBERHASILAN PEMASARAN PRODUK LAYANAN PERBANKAN,” vol. XIII, no. 1, pp. 60–66, 2016.
[5] W. J. Frawley, G. Piatetsky-shapiro, and C. J. Matheus, “Knowledge Discovery in Databases: An Overview,” vol. 13, no. 3, pp. 57–70, 1992.
[6] F. Gorunescu, Data mining: concepts and techniques. 2011.
[7] J. Wu, Advances in K-means Clustering, vol. 53, no. 9. 2013.
[8] H. B. Zheng, R. J. Liao, S. Grzybowski, and L. J. Yang, “Fault diagnosis of power transformers using multi-class least square support vector machines classifiers with particle swarm optimisation,” IET Electr. Power Appl., vol. 5, no. 9, p. 691, 2011.
[9] N. Deng, Y. Tian, and C. Zhang, Support Vector Machines. 2013.
[10] Y. Liu, X. Yu, J. X. Huang, and A. An, “Combining integrated sampling with SVM ensembles for learning from imbalanced datasets,” Inf. Process. Manag., vol. 47, no. 4, pp. 617–631, 2011.
[11] J. Han, M. Kamber, and J. Pei, Data Mining Concepts and Techniques. 2012.
[12] I. H. Witten, E. Frank, and M. a. Hall, Data Mining Practical Machine Learning Tools and Techniques Third Edition, vol. 277, no. Tentang Data Mining. 2011.
[13] C. J. Mantas and J. Abellán, “Credal-C4.5: Decision tree based on imprecise probabilities to classify noisy data,” Expert Syst. Appl., vol. 41, no. 10, pp. 4625–4637, 2014.