An empirical study of the performance of two stage optimal ensemble classification using genetic algorithm

R D R Raharjo1, A M Soleh1 and B Sartono1*

1Department of Statistics, IPB University, Bogor 16680, Indonesia

E-mail: bagusco@apps.ipb.ac.id

Abstract. Zelenkov et al. [1] proposed a two-step classification method (TSCM) based on genetic algorithm to predict the bankruptcy of Russian companies. This current study was conducted to do more comprehensive evaluation than Zelenkov et al. [1] did. We involved more datasets, compared to a greater number of competitive methods, and developed the ensemble using more base classifiers. The datasets consisted of nine datasets and then the result of prediction compared to previous studies from 13 papers that were published between 1996 and 2009. In addition, the previous method using five base classifiers, this study involved seven base classifiers. The purpose of this study is to compare the accuracy of prediction from previous studies with this method. In this study, we used k-Nearest Neighbour (k-NN), Logistic Regression (LR), Naïve Bayes (NB), Decision Tree (DT), support vector machine (SVM), random forest (RF), and boosting as the base classifiers of ensemble model. Genetic algorithm is used to find the best predictor variables and the best weight for each base classifier. The result of TSCM shows that the accuracy of prediction can increase about 0-34% for some datasets compared to the previous studies.

1. Introduction

Data mining is a process of exploring and analyzing large amounts of data to find patterns and rules that have meaning. Data mining methods can be divided into two types, descriptive methods and predictive methods. Descriptive method is a method used to describe or provide data as they are without any analysis and general conclusions, while predictive methods are methods designed to predict new information based on the information we have today, the information produced can later both qualitative and quantitative [2]. Qualitative information can be in the form of a classification method that is widely used in banking and telemarketing-based companies.

Classification method can be divided into two types, supervised classification, and unsupervised classification. Supervised classification is a classification method that is carried out to determine the differentiation between groups, the differentiating factors that are produced can later be used to determine the membership of a new observation, while in the unsupervised classification, it is only to classify observations into a group.

There are many classification methods that can be used to predict the occurrence of an event in the future, such as logistic regression, k-nearest neighbor, support vector machine, and neural network. Along with the times, there is an idea to combine several or more classification methods with ensemble learning [3]. Random forest and boosting are examples of ensemble learning that can provide predictive results with high accuracy [4]. According to Dietterich [5] ensemble learning is a method that forms a collection of classification models which are then used to predict new data based on majority voting from each model. In general, the results of predictions from ensemble learning can provide higher
predictive accuracy results compared to making predictions using a single model. One thing that also plays an important role in obtaining predictive results with high accuracy is the selection of the right predictor variables. The selection of predictor variables will take a lot of time if the data consists of many predictor variables. One method that can be used to choose predictor variables more efficiently is the genetic algorithm method [1]. Genetic algorithm is an algorithm that is built by following the pattern of evolution in living things in order to obtain the best solution. In relation to the weighting of ensemble learning that uses majority vote, genetic algorithms can also be used to determine the best weight so that better predictions can be obtained.

This study will compare the goodness of other models with a two-stage ensemble classification based on genetic algorithms. To see the goodness of these models will be seen from the accuracy of the predictions of some simulation data.

2. Material and Methods

The two-stage ensemble method is a method used to select predictor variables that are able to provide a higher level of accuracy in a prediction and to determine the best weight in the ensemble process [1]. The first step in the method is the selection of the best predictors to be included in the ensemble model. After being selected the best predictors will enter the second stage, namely determining the best weight for the ensemble process. The measure of goodness used in both stages is the accuracy produced in the prediction process.

Genetic algorithm is used to select the best predictors and weights. The use of this genetic algorithm will be very useful when the data that will be predicted consists of many explanatory variables. Genetic algorithms are algorithms that mimic the patterns of evolution of living things in order to obtain the best individuals. This algorithm describes a natural selection process where only the best individuals can survive. In this algorithm, there is also a process of mutation and cross-breeding in each generation so that the best individuals will emerge. The six stages in the genetic algorithm are:

1. Initial population generation
2. Calculation of fitness function value
3. Selection of the best individuals
4. Crossover
5. Mutation
6. Substitution of generations, and repetitive processes for steps 2 to 5 to several predetermined generations.

![Figure 1. Flowchart of genetic algorithm](image-url)
The first step before carrying out the data analysis stage is to prepare data. At this stage, check on the presence or absence of missing data (missing value) at 9 datasets that will be used in the analysis stage. If there is a missing value in the dataset, it is handled first using one of the following 3 methods:

1. **Imputation**, using average values on numerical data and using mode values on categorical data.
2. **Eliminate incomplete observations.**
3. **Categorize missing value.**

The next step after handling the missing value is to do modeling. Data will be divided into 2 parts, namely 75% used as training data and 25% as testing data. The modeling stage itself consists of 2 parts. The first step in the method is the selection of the best predictors that will be included in the ensemble model. After being selected as the best predictors, it will enter the second stage, namely determining the best weight for the ensemble process. The measure of goodness used in both stages is the accuracy produced in the prediction process. Accuracy is the percentage of the class that is predicted correctly in the testing data.

### 2.1. Material

The data used in this study are data taken from the UCI machine learning repository. There are 9 datasets used to evaluate the model in this study. The characteristics of each dataset can be seen in Table 1.

| Name of dataset            | Number of observations | Number of numerical variables | Number of numerical variables | Total predictor variables |
|----------------------------|------------------------|-------------------------------|-------------------------------|---------------------------|
| Australian                 | 690                    | 6                             | 8                             | 14                        |
| Breast cancer              | 699                    | 9                             | 0                             | 9                         |
| Wisconsin                  | 303                    | 6                             | 7                             | 13                        |
| Clevenland heart disease   |                        |                               |                               |                           |
| Diabetes                   | 768                    | 8                             | 0                             | 8                         |
| German                     | 1000                   | 20                            | 0                             | 20                        |
| Hepatitis                  | 155                    | 6                             | 13                            | 19                        |
| Ionosphere                 | 351                    | 33                            | 0                             | 33                        |
| Sonar                      | 208                    | 60                            | 0                             | 60                        |
| Votes                      | 435                    | 0                             | 16                            | 16                        |

### 2.2. Method

#### 2.2.1. Stage 1

- Generate an initial population consisting of n individuals. Suppose that the problem to be solved is a data consisting of p predictor variables, then each variable will be numbered from 1 to p. An individual in a population can be seen as a p×1 binary vector containing only 0 and 1. The 0 and 1 values indicate whether the predictor variables are in the model or not, if a predictor is not included in the model then given a value of 0 and if a predictor is entered into the model it will be given a value of 1. Different vectors represent different compositions from the predictor variables to be used in the model.
- Calculate the fitness value which is the average value of accuracy obtained from the prediction process using the algorithm. The algorithms that be used include, k-NN, logistic regression, NB, DT, SVM, random forest, boosting.
- Choose the best individuals by looking at their accuracy.
- Conduct interbreeding between the best individuals selected in the process c.
The descendants generated from the process d will be calculated for the accuracy value and will later become elders for the next generation.

- Perform mutation processes for individuals to obtain better derivatives. The mutation process is done by changing several labels on individual chromosomes.
- Repeats the previous steps as many as G generations

2.2.2. Stage 2

- Generate an initial population consisting of n individuals. Each individual in the population is a \( m \times 1 \) vector that contains the weight values of each model.
- Give the weight of \( w_i \) which is a positive real, to the models used in the ensemble process with the provisions \( \sum_i w_i = 1 \).
- The fitness value of the ensemble is determined by calculating accuracy values in training data. Suppose \( C_E \) is a class of objects obtained from \( C_E = \sum_i w_i C_i \) where \( C_i \) is the result of a prediction from a general classifier.
- Choose the best individuals by looking at the fitness value.
- Conduct interbreeding between the best individuals who have been elected.
- Perform mutations by replacing some of the initial weight values on individual chromosomes with values at intervals (-0,1; 0,1), so that \( w_i^* \) is obtained.
- If \( \sum_i w_i^* \) from the result of crossing and mutation is not equal to 1 then it is normalized first, \( w_k = \frac{w_i^*}{\sum_i w_i^*} \) so that \( \sum_k w_k = 1 \).
- Repeat the previous steps as many as G generations.

3. Result

The results of the pre-processing data found that there are missing values in several datasets. The handling process carried out on each dataset in which there is a missing value can be seen in Table 2. Of the ten datasets used, 4 of them contained a missing value. In the Hepatitis data, an imputation process was carried out to deal with the missing value problem. The imputation process is done by entering the average value for the predictor in the form of numerical data and entering the mode value for the predictor in the form of categorical data. Handling the missing value for the Breast Cancer Wisconsin data and Cleveland Heart Disease was carried out by removing observations in which there was a missing value. This is done because the number of missing values in the two datasets is small. The missing value problem in Mushroom data is handled by removing one of the predictor variables. The abolition is done because of the amount missing value very large on the predictor variable.

| Name of dataset       | Number of missing values | Handling method                      |
|-----------------------|--------------------------|--------------------------------------|
| Australian            | 0                        | -                                    |
| Breast cancer Wisconsin| 16                      | Remove observations with a missing value |
| Clevenland heart disease| 6                      | Remove observations with a missing value |
| Diabetes              | 0                        | -                                    |
| German                | 0                        | -                                    |
| Hepatitis             | 167                      | Imputation                           |
| Ionosphere            | 0                        | -                                    |
| Sonar                 | 0                        | -                                    |
| Votes                 | 0                        | -                                    |
The next step is to do the modeling process. The modeling process is carried out in 2 stages. The results obtained from the first stage modeling are predictor variables which will be used in the second stage modeling, the results are shown in Table 3.

Table 3. Number of predictor variables from the first stage modelling.

| Name of dataset | Number of Predictor Variables | Number of Selected Predictors in Stage 1 (min; max) |
|-----------------|------------------------------|-----------------------------------------------|
| Breast Cancer (W) | 9                            | 3;6                                           |
| Diabetes        | 8                            | 3;5                                           |
| Ionosphere      | 34                           | 15;21                                         |
| Sonar           | 60                           | 27;35                                         |
| Australian      | 14                           | 4;10                                          |
| Heart Disease (C) | 13                         | 5;9                                           |
| German          | 20                           | 11;17                                         |
| Hepatitis       | 19                           | 5;11                                          |
| Votes           | 16                           | 6;11                                          |

Table 4. Median accuracy of 10 repetitions.

| Method | Breast Cancer | Votes | Diabetes | Heart Disease | Ionosphere |
|--------|---------------|-------|----------|---------------|------------|
| TSCM   | 0.9591        | 0.9679| 0.7344   | **0.8467**    | 0.9375     |
| RL*    | 0.9737        | 0.9541| 0.7708   | 0.8267        | 0.8693     |
| GNB*   | 0.9678        | 0.9541| 0.7578   | **0.8467**    | 0.8977     |
| KNN*   | **0.9795**    | 0.9358| 0.7083   | 0.6000        | 0.8239     |
| SVM*   | 0.9591        | 0.9541| 0.6693   | 0.5067        | 0.9261     |
| DT*    | 0.9444        | 0.9220| 0.7135   | 0.7000        | 0.9034     |
| AdaBoost* | 0.9620   | 0.9541| 0.7552   | 0.7733        | 0.9148     |
| RF*    | 0.9649        | 0.9679| 0.7448   | 0.7667        | 0.9318     |
| Stacking | 0.9561    | 0.9358| 0.6953   | 0.7133        | 0.8750     |
| NN     | 0.9678        | 0.9495| 0.6693   | 0.5400        | 0.9205     |
| AD     | 0.9678        | 0.9495| 0.7734   | 0.8400        | 0.8693     |
| Bagging | 0.9532   | 0.9495| 0.7578   | 0.7467        | 0.9034     |

| Method | Sonar | Hepatitis | Australian | German |
|--------|-------|-----------|------------|--------|
| TSCM   | **0.9038** | 0.7179   | 0.8613     | 0.7500 |
| RL*    | 0.7788 | 0.7179   | 0.8526     | **0.7600** |
| GNB*   | 0.6346 | 0.7179   | 0.8035     | 0.7280 |
| KNN*   | 0.8173 | 0.6282   | 0.6763     | 0.6920 |
| SVM*   | 0.5769 | 0.5769   | 0.5665     | 0.7280 |
| DT*    | 0.7308 | 0.5256   | 0.8208     | 0.6860 |
| AdaBoost* | 0.7981 | 0.6795   | 0.8382     | 0.7460 |
| RF*    | 0.7692 | 0.6282   | 0.8642     | 0.7460 |
| Stacking | 0.7212 | 0.5256   | 0.8237     | 0.6900 |
| NN     | 0.8173 | 0.5256   | 0.6705     | 0.7480 |
| AD     | 0.7115 | 0.6923   | 0.8613     | 0.7540 |
| Bagging | 0.7692 | 0.6410   | **0.8642** | 0.7260 |
Predictor variables used in the first stage modeling are then used to do the second stage modeling. In the second stage modeling, the best weights were used in the ensemble process. The median value of a prediction curation from nine datasets with 10 repetitions using TSCM and using the method from previous studies is shown in Table 4. The highest accuracy in each data is marked with a value written in bold. The method given the symbol "*" is the base classifier used in the ensemble process in this study.

In general, the method used in the study was able to improve accuracy compared to using Method in previous studies. Judging from the median value, the method used in this study provides the highest accuracy value in five datasets, namely Votes, Cleveland Heart Disease, Ionosphere, Sonar, and Hepatitis data. The last step is to compare the goodness of the model in this study with the methods used in previous research. The goodness of the model is seen by calculating how much the prediction accuracy using the model in this study is higher than the models in previous studies.

### Table 5. TSCM comparison with other methods.

| Comparison method | Percentage |
|-------------------|------------|
| RL*               | 55.56%     |
| GNB*              | 55.56%     |
| KNN*              | 88.89%     |
| AD                | 66.67%     |
| SVM*              | 100%       |
| DT*               | 100%       |
| Bagging           | 77.78%     |
| AdaBoost*         | 77.78%     |
| RF*               | 66.67%     |
| Stacking          | 100%       |
| NN                | 88.89%     |

The method used in this study provides better predictive results compared to 11 other methods, although when compared with the RL and GNB methods the difference in victory is relatively small. Based on Table 5, it is also seen that the ensemble method that uses seven base classifiers (k-Nearest Neighbor, Logistic Regression, Naïve Bayes Gaussian, Decision Tree, Support Vector Machine, Random Forest, AdaBoost) provides better predictive results than if the methods in base classifier work alone.

### 4. Conclusion

The ensemble method with the selection of variables and weights using Genetic Algorithms (TSCM) has been able to provide better predictive results compared to the methods used in previous studies. Future research may be able to explore data further so that it is expected to provide better predictive results. In addition, adding a base classifier such as XGBoost can also be done to get a higher accuracy value.

### References

[1] Zelenkov Y, Fedorova E and Chekrizov D 2017 *Expert Systems with Applications* 88 393.
[2] Tuffery S 2011 Data Mining and Statistics for Decision Making. West Sussex (UK): John Wiley & Sons, Inc.
[3] Kuhn M and Johnson K 2013 *Applied Predictive Modeling* New York (US): Springer.
[4] Breiman L 2001 *Machine Learning* 45 5.
[5] Dietterich TG 2000 *Multiple Classifier Systems* 1.