Improving prediction accuracy of classification model using cascading ensemble classifiers

A A Aziz¹, Indahwati¹ and B Sartono¹*

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

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

Abstract. One way to improve the accuracy of predictive modeling is by combining the models. This research tries to study local cascade. It combined one or more base classifier sequentially. In each stage, the probability prediction of the base classifier was inserted to the data. The data then modeled using a decision tree algorithm. This process continued until the data is homogenous. In the original method, the base classifier used was non-ensemble classifier. Our study included bagging, boosting, and random forest as base classifiers. 11 dataset with binary response was used to assess the accuracy of this method. We also compared the accuracy of our method with others that were published between 1996 and 2009. We found that cascading ensemble classifier slightly improve accuracy and performed better for a dataset with numerical predictors.

1. Introduction

Classification is part of predictive modeling that have categorical target variable. Many classification algorithms were built to handle various data. No algorithms is always better than the others. One approach to increase accuracy in classification modeling is combining prediction from several algorithms. Some research showed that this technique tends to have better accuracy compare to a prediction from single algorithms [1].

Some popular techniques to combine prediction are voting, model averaging, stacking, and cascading. Voting, model averaging, and stacking train the base classifiers at the same time. In voting final prediction was determined from the majority of class prediction between base classifiers [2]. Model averaging use the average probability prediction of the class [3]. Stacking train a new model by using the probability prediction of the class from some base classifiers. Cascade has a different nature compared previous method. This method combines the prediction of the base classifiers in a sequence [2].

Cascading was done by adding a new column in the data which is the probability prediction of the base classifier. The new data then trained again with different classification algorithms. Study of cascade generalization by Gama and Brazdil [2] only used a decision tree, discriminant, naïve Bayes, and C4.5 algorithms. This method improves prediction accuracy compared to the single classifier, stacking, and comparable to boosting algorithms.

Application for cascade generalization in Indonesia also showed the improvement of prediction accuracy. Research by Nugroho [4] applied cascade generalization to improve the prediction of breast cancer using mammography data. The best result using global cascade generalization had a prediction accuracy of 83.689% [4]. Unfortunately, we haven’t found other published research that use cascade generalization in Indonesia.
Therefore, we aim to study cascade generalization with other base classifier algorithms that haven’t been studied by Gama and Brazdil [2]. The other classification algorithms that will be studied include logistic regression, neural networks, linear support vector machine, bagged tree, boosted tree, and random forest. Our method was compared to other research that used binary target variables. The criterion for model performance used was prediction accuracy.

2. Materials and Methods

Cascade generalization is one approach to combine some prediction from the different base classifier. In general, this method adds new variables in a dataset which are the probability predictions of the target variable. New data then trained with different classification algorithm to get the final prediction. Our study will focus on local cascade generalization that had better performance compared to global cascade generalization [2].

The first step in our method is data preparation. We check for any missing data. There were three methods used in our study. The first one imputation of missing data using mean for numerical variable and mode for a categorical variable. The second one is deleting variable with much values ing value. The last one is deleting observation with a missing value. The method applied depends on the condition of the data.

Table 1. Dataset characteristics used in the study.

| No | Dataset               | Number of Observation | Number of Numerical Predictor | Number of Categorical Predictors |
|----|-----------------------|-----------------------|-------------------------------|----------------------------------|
| 1  | Adult                 | 48842                 | 6                             | 8                                |
| 2  | Australian            | 690                   | 6                             | 8                                |
| 3  | Breast Cancer Wisconsin | 699                  | 9                             | -                                |
| 4  | Cleveland Heart Disease | 303                  | 6                             | 7                                |
| 5  | Diabetes              | 768                   | 8                             | -                                |
| 6  | German                | 1000                  | 9                             | 11                               |
| 7  | Hepatitis             | 155                   | 6                             | 13                               |
| 8  | Ionosphere            | 351                   | 34                            | -                                |
| 9  | Mushroom              | 8124                  | -                             | 22                               |
| 10 | Sonar                 | 208                   | 60                            | -                                |
| 11 | Votes                 | 435                   | -                             | 16                               |

The second step was modeling. Dataset will be split into training and testing data with a percentage of 70% and 30% respectively. The modeling process will follow the methodology that was proposed By Gama and Brazdil [2]. For example, we used decision tree as the final classifier, M₁ and M₂ as base classifiers. If there are p predictor variables, Y binary target variable, and data D. The modeling process are:

1. Train base classifier with M₁ and M₂ algorithms using data D.
2. Make new variable Wᵢ, where Wᵢ is \( \hat{P}(Y=1) \) that were calculated using algorithms Mᵢ; i=1,2
3. Train decision tree with depth = 1, Y as target variables and \{X₁, ..., Xₚ \} \cup \{ W₁, W₂ \} as predictor variables. This step will produce a splitting criterion that will split the data into two, D₁ and D₂ where:
   \[ D₁ \cap D₂ = \emptyset \quad D₁ \cup D₂ = D \]
4. Repeat 1, 2, and 3 for D₁ dan D₂ until the data can’t be split with a decision tree (homogenous).
The base classifier used in our study were logistic regression (LR), Linear Discriminant (Disc), neural networks (Nnet), linear support vector machine (linSVM), bagged tree (Bag), boosted tree (Boost), and random forest (Ranfor). Two algorithms were used for each combination so that in total there were 21 combinations. We chose these algorithms because they were widely used and usually have good performance. This method will be evaluated using accuracy for selected 11 datasets from UCI repository and mlbench library in R that was shown in Table 1.

3. Results and Discussion

We found some data with a missing value. The method used to handle the missing value for each dataset was shown in table 2. 5 out of 11 data that were used had missing value. Missing value for adult and hepatitis data was imputed with mean or mode. For breast cancer and heart disease data, the observation with missing value was relatively small so it was deleted. Mushroom data had many missing values for one variable, so the variable with missing value was deleted.

Table 2. Missing value handling.

| No | Dataset | Number of Missing Value | Handling Method |
|----|---------|-------------------------|-----------------|
| 1  | Adult   | 6465                    | Impute with the mean for numerical predictors and mode for categorical predictors |
| 2  | Australian | -                 | -               |
| 3  | Breast Cancer Wisconsin | 16           | Delete observation with a missing value |
| 4  | Cleveland Heart Disease | 6              | Delete observation with a missing value |
| 5  | Diabetes | -                       | -               |
| 6  | German   | -                       | -               |
| 7  | Hepatitis | 167                  | Impute with a mean for numerical predictors and mode for categorical predictors |
| 8  | Ionosphere | -                    | -               |
| 9  | Mushroom | 2500                    | Delete variable with a missing value |
| 10 | Sonar    | -                       | -               |
| 11 | Votes    | -                       | -               |

The next step is modeling. Table 3 showed the accuracy of our study compares to the previous study. In general, accuracy is close to the previous study. Highest accuracy for each dataset was in bold. The additional symbol “►” was added if the best accuracy from our method. Border with the triple line was used the separate our result and previous study. Mean accuracy for the previous study was mean accuracy 1 and for our study mean accuracy was mean accuracy 2.

Combination with logistic regression classifier was omitted for sonar data because of too many predictor variables for relatively small data. As shown in table 3, the best accuracy with our method found for breast cancer data. Unfortunately for data with relatively low accuracy like diabetes data, our method didn’t improve the accuracy compared to the previous study. We also found that mean accuracy for our method was slightly better to compare to mean accuracy from the previous study for breast cancer data.
### Table 3. Accuracy of dataset with numerical predictor variables.

| Methods                  | Breast Cancer (W) | Diabetes | Ionosphere | Sonar |
|--------------------------|-------------------|----------|------------|-------|
| Discriminant             | 0.945             | 0.773    | 0.895      | 0.747 |
| Neural Network           | 0.981             | 0.715    | **0.962**  | **0.904** |
| SVM                      | 0.961             | 0.765    | 0.870      | 0.759 |
| CART                     | 0.942             | 0.735    | 0.906      | 0.716 |
| Bagged Tree              | 0.963             | 0.761    | 0.931      | 0.822 |
| AdaBoost                 | 0.967             | 0.781    | 0.940      | -     |
| Random Forest            | **0.973**         | 0.757    | 0.925      | 0.820 |
| Stacking                 | 0.973             | 0.758    | 0.930      | -     |
| Cascade Generalization   | 0.974             | 0.776    | **0.934**  | 0.808 |
| Mean Accuracy 1          | 0.963             | 0.756    | 0.920      | 0.806 |
| C1_LR Disc               | 0.971             | 0.748    | 0.867      | -     |
| C2_LR Nnet               | 0.971             | 0.743    | 0.829      | -     |
| C3_LR LinSVM             | 0.976             | 0.757    | 0.886      | -     |
| C4_LR Bag                | 0.971             | 0.735    | 0.924      | -     |
| C5_LR Boost              | 0.971             | 0.748    | 0.933      | -     |
| C6_LR RanFor             | 0.971             | 0.739    | 0.933      | -     |
| C7_LR Disc               | 0.971             | 0.717    | 0.829      | 0.710 |
| C8_LR LinSVM             | 0.971             | 0.761    | 0.838      | 0.581 |
| C9_LR Disc               | 0.971             | 0.761    | 0.895      | 0.694 |
| C10_LR Boost             | 0.971             | 0.748    | 0.933      | 0.839 |
| C11_LR Disc              | 0.971             | 0.709    | 0.914      | 0.758 |
| C12_LR Disc              | 0.961             | 0.770    | 0.914      | 0.742 |
| C13_LR Disc              | 0.971             | 0.735    | 0.895      | 0.710 |
| C14_LR Boost             | 0.976             | 0.734    | 0.942      | 0.603 |
| C15_LR Boost             | 0.971             | 0.722    | 0.914      | 0.839 |
| C16_LR LinSVM            | 0.976             | 0.722    | 0.895      | 0.710 |
| C17_LR Boost             | 0.980             | 0.739    | 0.933      | 0.823 |
| C18_LR LinSVM            | 0.980             | 0.722    | 0.924      | 0.887 |
| C19_LR Boost             | 0.985             | 0.778    | 0.914      | 0.855 |
| C20_LR Disc              | 0.971             | 0.761    | 0.914      | 0.839 |
| C21_LR Boost             | 0.976             | 0.751    | 0.933      | 0.790 |
| Mean Accuracy 2          | 0.973             | 0.743    | 0.903      | 0.759 |

Table 4 showed the accuracy of dataset with categorical and numerical predictor variable. 7 dataset was used to assess the performance of our methods. The highest accuracy of each dataset was indicated in the same manner with table 3. Only 6 combinations were used for the following data. We found that our method had best accuracy compared to previous study for adult data. The other interesting thing that we found was in votes data. One variable seems to have good predictability and always used as splitting criterion in the tree. After the first split, the data was relatively homogenous. The new variable from the probability prediction didn’t help to differentiate the class of target variable. We also found that mean accuracy for our method was slightly better compare to mean accuracy from previous study for adult data.
Table 4. Accuracy of dataset with categorical and numerical predictor variables.

| Methods                | Australian | Adult | Heart Disease(C) | German | Hepatitis | Mushroom | Votes |
|------------------------|------------|-------|------------------|--------|-----------|----------|-------|
| Discriminant           | 0.859[2]   | 0.781[2]   | 0.839[2]       | 0.780[2] | 0.784[2] | 1.000[2] | 0.964[2] |
| Neural Network         | 0.842      | 0.806       | 0.890[13]     | 0.750   | 0.819     | 1.000    | 0.961 |
| SVM                    | 0.869[8]   | -           | 0.825          | 0.750   | 0.854     | 1.000    | 0.968 |
| CART                   | 0.859[2]   | 0.861[2]   | 0.793[2]       | 0.721[2] | 0.805[2] | 1.000[2] | 0.967[2] |
| Bagged Tree            | 0.879[9]   | 0.855       | 0.819[9]       | 0.772[9] | 0.835[9] | 1.000    | 0.964[9] |
| AdaBoost               | 0.874[9]   | -           | 0.826[9]       | 0.771[9] | 0.862[9] | 0.031    | 0.956[9] |
| Random Forest          | 0.872      | 0.859       | 0.738          | 0.738[12] | 0.852    | 1.000    | 0.954[1] |
| Cascade                | 0.868[2]   | 0.863[2]   | 0.849[2]       | 0.770[2] | 0.847[2] | 1.000[2] | 0.967[2] |
| Mean Accuracy 1        | 0.864      | 0.838       | 0.824          | 0.751   | 0.827     | 0.879    | 0.963 |
| C4_LR Bag              | 0.807 ▶ 0.865 | 0.775     | 0.747          | 0.766   | 1.000     | 1.000    | 0.954 |
| C5_LR Boost            | 0.855      | -           | 0.697          | 0.720   | 0.745     | 1.000    | 0.954 |
| C6_LR RanFor           | 0.855      | 0.856       | 0.809          | 0.720   | 0.830     | 1.000    | 0.954 |
| C19_Bag Boost          | 0.826      | -           | 0.742          | 0.737   | 0.851     | 1.000    | 0.947 |
| C20_Bag Ranfor         | 0.855      | 0.863       | 0.775          | 0.707   | 0.766     | 1.000    | 0.954 |
| C21_Boost Ranfor       | 0.850      | -           | 0.753          | 0.693   | 0.851     | 1.000    | 0.947 |
| Mean Accuracy 2        | 0.841      | 0.861       | 0.759          | 0.721   | 0.802     | 1.000    | 0.952 |

After modelling we tried to visualize the accuracy of our method based on algorithm of base classifier. Figure 1 was used to determine whether any algorithm used for base classifier will make the optimal prediction. In general, the base classifier used in our study had same impact on accuracy. The highest median used linear SVM and the lowest one used neural networks as base classifier.

![Figure 5](image-url)  
**Figure 5.** Accuracy based on algorithms of base classifier.
The last step used in our study was to evaluate the performance of cascade for a dataset with numerical predictor only and others. This step was done by counting how many times the best combination of the base classifier using cascade win or have the same accuracy compared to the previous study. Border with double line separated dataset with numerical predictor only and others.

| Dataset            | Number of wins | Number of losing | Percentage of win |
|--------------------|----------------|------------------|-------------------|
| Breast Cancer (W)  | 11             | 0                | 100.00            |
| Diabetes           | 10             | 1                | 90.91             |
| Ionosphere         | 10             | 1                | 90.91             |
| Sonar              | 7              | 1                | 87.50             |
| Australian         | 2              | 7                | 22.22             |
| Adult              | 6              | 0                | 100.00            |
| Heart Disease(C)   | 2              | 7                | 22.22             |
| German             | 3              | 4                | 42.86             |
| Hepatitis          | 5              | 1                | 83.33             |
| Mushroom           | 8              | 0                | 100.00            |
| Votes              | 1              | 8                | 11.11             |

Table 5 showed the summary of cascading ensemble classifier compares to the previous study. Cascade performed better for a dataset with numerical predictors only. Cascade win with a percentage of 87.5 – 100.0. For a dataset with categorical and mixed of numerical-categorical predictors, cascade win with a percentage of 11.11 – 100.0. We believe cascade algorithms performed better for data with numerical predictor only because of its nature. The algorithms split the data and base model were trained using chosen algorithms on smaller data. Numerical predictor could have any values in real number for each observation on small data. In the case of categorical predictor, the values is limited to the number of its category. Because of that, it is possible to produce a better result for data with numerical predictors only.

4. Conclusion
The aim of this research was to study local cascade using ensemble model as the base classifier. We found that cascading ensemble classifier slightly improve accuracy for breast cancer and adult data. Based on base classifier used in our study, cascade had similar prediction accuracy. We also found that cascade performed better for a dataset with numerical predictors only. This study recommends cascading ensemble classifier for a dataset with numerical predictors only.

References
[1] James G, Witten D, Hastie T and Tibshirani R 2013 An Introduction to Statistical Learning Vol. 112 New York(US), Springer.
[2] Gama J and Brazdil P 2000 Machine learning 41(3) 315.
[3] Ting KM and Witten IH 1999 Journal of Artificial Intelligence Research 10 271.
[4] NUGROHO KA, Setiawan NA and Adji TB 2013 Proceedings of the Information Technology and Electrical Engineering 57.
[5] Abbass HA 2002 Artificial Intelligence in Medicine 25(3) 265.
[6] Jiang Y, Zhou ZH and Chen ZQ 2002 Proceedings of the 2002 International Joint Conference on Neural Networks 2 1416.
[7] Gorman RP and Sejnowski TJ 1988 Analysis of Hidden Units in a Layered Network.
[8] Baesens B, Viaene S, Van Gestel T, Suykens JA, Dedene G, De Moor B and Vanthienen J 2000
[12] Proceedings of the Fourth International Conference on Knowledge-Based Intelligent Information
[13] Engineering Systems & Allied Technologies 1 313.
[14] Maclin R and Opitz D 1997 An Empirical Evaluation of Bagging and Boosting AAAI/IAAI p 546.
[15] Gorman RP and Sejnowski TJ 1988 Neural Networks 1(1) 75.
[16] Breiman L 1996 Machine Learning 24(2) 123.
[17] Breiman L 2001 Machine Learning 45(1) 5.
[18] Das R, Turkoglu I and Sengur A 2009 36(4) 7675.