Empirical Study on the Predictive Power of Rotation Forest

Bagus Sartono¹*, Mulianto Raharjo¹, Cici Suhaeni¹

¹Department of Statistics – Bogor Agricultural University
E-mail: bagusco@gmail.com, bagusco@ipb.ac.id

Abstract. This paper presents the comparison of the predictive power of rotation forest and other classification techniques in the case of classification problem. Rotation forest is an ensemble of many decision trees that utilizing principal component analysis to do some rotation to the original predictor variables before constructing the trees. An empirical study using fourteen different datasets was held to observe how good the prediction resulted by the rotation forest. The authors found that in most of the cases rotation forest perform better compared to logistic regression, tree, and discriminant analysis. We also revealed that the rotation forest fails to have excellent prediction when the predictors are dominated by categorical variables. In general, rotation forest could be a competitive approach to handle classification tasks.

Keyword: classification, ensemble method, rotation forest, tree

1. Introduction
In many applied researches, as well as practical analytics, people likes to do some prediction concerning the class of new observations based on their characteristics. The predictions were made by using models or rules that use the characteristics as inputs and result in a kind of probability of the membership of those observations to the unknown classes. For example, a bank officer should predict whether a loan applicant would be able to pay or not after the loan is given. A model such as a credit scoring model is helpful to guide the officer to make a decision. Another example is that a marketing executive officer should decide whether or not he calls a customer and offers a certain product. A statistical model might be helpful to identify the level of the propensity to buy of persons. In general we called the model as a classification model.

There are several basic popular classification techniques, e.g. discriminant analysis, logistic regression, decision tree and support vector machine. Some of them have evolved into more complicated algorithms and it happens especially on the decision tree. One of recent development of the decision tree is such an ensemble technique called rotation forest by [1] which was inspired by two previous algorithms: Bagged Tree [2] and Random Forest [3]. An ensemble tree does the prediction by relying not only to a single tree
but combining the prediction from many trees. The authors of [4] described that the ensemble approach is a good strategy to have better prediction performance.

Suppose that we have a set of characteristic variables $X = \{X_1, X_2, \ldots, X_p\}$ that would be used as inputs to have a rule to classify observations into class $Y$. The dataset should be consisted of $n$ rows representing $n$ observation, and $p$ columns of input characteristics and a column of class. The rotation forest just like random forest works by producing $L$ decision trees and at the end the prediction will be produced by the aggregation of individual predictions that resulted by the trees. Therefore, it is an ensemble method. Let’s have a look of the general algorithm as described by [1].

First, a bootstrap sample was withdrawn from the training dataset. The set of characteristics of the sample was then rotated using the following procedure. The characteristic set is randomly divided into $K$ subsets. In each set, separately, a rotation matrix was produced using a principal component analysis approach and then collected into a single big rotation matrix. The last matrix is then use to rotate the whole data consisting all characteristic variables $X$, after some column reposition to ensure that the column of the rotation matrix rotate the correct corresponding column. The rotated $X$ matrix is then used as an input predictor matrix to build a decision tree. This step is then repeated $L$ times, so that at the end we will have $L$ trees which differ one to others. The variation among trees is contributed by the bootstrap process and by the random division of the characteristic set during the rotation. Those $L$ trees are then resulting $L$ predictions that ensembled into one single final prediction. A majority vote is the common combining process that is used to obtain the final prediction.

As mentioned in [5] who did experimentation by applying the rotation forest algorithm for classification to some well known datasets that are available in UCI Machine Learning Repository. There are thirty two datasets involved and the classification accuracy of the algorithm was compared to the accuracies of three other ensemble methods: Bagging, Boosting, and Random Forest. They noted that the rotation forest was working well and mostly better than the others, but comparable to boosting algorithm.

This recent paper did similar idea to [5] in term of the aim that it is evaluating the prediction power of the rotation forest. However, unlike [5] that ran several algorithms to each dataset, here the authors implemented only the rotation forest to several datasets and compare the accuracy to the one of the original techniques reported by the theses authors. There are fourteen datasets that are in interest, and all datasets are the ones that used by students in the Department of Statistics – Bogor Agricultural University, Indonesia. The datasets were analyzed and the results were reported in the theses that were the final projects of them to get degree of Bachelor and Magister from the university during 2011 – 2015.

2. Data and Method

As mentioned previously, this study works with fourteen datasets to evaluate the performance of rotation forest in classifying observations. We coded the datasets into A1, A2, …, A14 and the brief description of them are shown in Table 1. In the original thesis papers, two of authors were using discriminant analysis, five with decision tree using various splitting algorithms, and seven others are using logistic regression. The data covered a range of topic such as educational matter, medical problem, up to social and business.

At the very first stage, we did a careful discussion with the theses authors to get deep information on the variables and the treatment that they did during the analysis. Fortunately, we could reach all the theses author so we could confirm all the things they did during the theses writing. We did that to guarantee that we did the same treatment in the data preparation so that the result can be compared fairly. The failure of this step may mislead the conclusion. It is why we did not involve older theses to avoid the case that the author could not remember exactly what they performed when they wrote the theses papers.
Next, the dataset were then submitted in the algorithm of rotation forest. The author implemented this using the available package in R called rotationForest prepared by Michel Ballings and Dirk Van den Poel. The implementation in each of the datasets was repeated 100 times, and the prediction performance of the results was recorded. In every repetition, the performance was slightly different due to the randomness in the algorithm. The best performance among those is selected to be compared with the original performance reported by the theses authors.

Table 1. Brief description of the datasets used

| Data sets | Thesis Topics                              | Theses Authors | Year of Theses | Classification Method Used | Method |
|-----------|--------------------------------------------|----------------|----------------|-----------------------------|--------|
| A1        | Subclinical status of mastitis in cattle   | I Dewa G Amory | 2012           | Stepwise discriminant analysis | discriminant |
| A2        | Credit status of retail loan               | Resty Indah Sari | 2011           | Stepwise discriminant analysis | discriminant |
| A3        | Micro credit status                        | Gitania Rahisti | 2015           | CART                        |        |
| A4        | Ketepatan waktu lulus mahasiswa FMIPA      | Rindy Pertiwi   | 2013           | CHAID                       |        |
| A5        | Ketepatan waktu lulus mahasiswa IPB        | Rindy Pertiwi   | 2013           | CHAID                       |        |
| A6        | Child labor status in Jakarta              | Dimas Adiangga  | 2015           | C5.0                        |        |
| A7        | Infection of Toxicara cati on cats         | Nur Fitriani    | 2015           | QUEST                       |        |
| A8        | Student graduation status in Bogor Agr Univ. | Nita Nurgenita | 2015           | Logistic regression         |        |
| A9        | Student preference on entrepreunership     | Adi Nugraha     | 2015           | Logistic regression         |        |
| A10       | Jamu (herbal medicine) composition         | Rossi Barro     | 2013           | Logistic regression         |        |
| A11       | Village classification in Indonesia        | Shafa Surbakti R | 2015           | Logistic regression         |        |
| A12       | Hypertency trigger factor                  | Meita Rubiati A | 2014           | Logistic regression         |        |
| A13       | Fisherman preference to adopt new technology | Alfin Khairi    | 2014           | Logistic regression         |        |
| A14       | Credit risk modeling                       | Rakhmawati      | 2011           | Ridge logistic regression   |        |
3. Result and Discussion

Figure 1 depicted the comparison of the prediction accuracy between the original algorithm and rotation forest in all fourteen datasets. The accuracy is the percentage of correct classification that calculated using testing set. The bars with pattern int the figure represent the accuracy of the rotation forest, while the clear bars represents the accuracy of the original method used in the theses.

![Figure 1. The accuracy of prediction of the original method and the rotation forest in each datasets](image)

We could learn from Figure 1 that in most cases the rotation forest worked better than the original algorithm used by the previous researchers. There are three datasets where it is not true, that are A5, A10, and A11. However if we look at carefully, the difference is very small. In those three datasets, the ratios between the accuracy of the rotation forest and the competitor are around 99.5%.

Our investigation about this weakness revealed that those three datasets have characteristic variables that are dominated by categorical variables, and it is not the case in other datasets. The possible explanation about this is as follow. The rotation forest may work ideally with continuous (numerical) predictor variables. When we have a categorical variable, it is then turned into binary dummy variables. A variable with $k$ categories will be replaced by $(k - 1)$ dummy variables. Using this process the dummy variables must be orthogonal each other. We should aware that the principal component analysis does not result significant transformation or rotation to the original variables if the variables are already orthogonal or have very small correlation. By this reason, rotation forest fails to handle the data to obtain better classification model.

We explored further the result of one hundred replications. As mentioned in the previous section that we do likely get different results for a different run of rotation forest. There are at least two random sources of the difference. First it is caused by the random of bootstraping in selecting observations. Second, it is caused by the random split of the variable sets during the determination of the rotation matrix.

Figure 2 summarizes the variation of the performance of the random forest compared to the original algorithm of the theses. We plot a box plot for each dataset to show the distribution of the ratio between accuracy of both algorithm. A ratio that is greater than 100% means that the rotation forest is better than the competitor, and a ratio less than 100% means the other way around.
In most of the cases the “left” whiskers of the boxplot lie above the line of 100%. It means that the worst result of a hundred trials to have rotation forest model is still better than the original algorithm. Therefore, even if the result of the rotation forest is not a constant, but it has high likelihood that the prediction power is higher than the other classification algorithm.

![Boxplot of prediction power ratio between rotation forest and previous methods](image)

Figure 2. Distribution of prediction power ratio between rotation forest and previously used methods

4. Conclusion

Rotation forest works well in producing classifier with high predictive power. In most of the datasets we studied, rotation forest provides larger value of the accuracy which is equivalent to smaller number of false prediction. We should also aware that because of random process inside the algorithm, we may have different result of prediction from the rotation forest approach. However, the small simulation we performed shows that in general the power of prediction do not differ a lot from one run to another.

In conclusion, the author would say that rotation forest could be a good alternative in handling some classification task in business sector or researches. It is potential to replace some more popular approach such as logistic regression which is already familiar to most people.

5. References

[1] Rodriguez JJ, Kuncheva LI, and Alonso CJ 2006 Rotation Forest: A New Classifier Ensemble Method in *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28** (10) 1619-1630.

[2] Breiman L 1996 Bagging predictors *Machine Learning* **24** (2) 123–140

[3] Breiman L 2001 Random Forests *Machine Learning* **45** (1) 5–32

[4] Sartono B and Syafitri UD 2010 Metode Pohon Gabungan: Solusi Pilihan untuk Mengatasi Kelemahan Pohon Regresi dan Klasifikasi Tunggal *Forum Statistika dan Komputasi* **15** (1).

[5] Kuncheva LI and Rodriguez JJ 2007 An Experimental Study on Rotation Forest Ensembles *Multiple Classifier Systems* **4472** of the series Lecture Notes in Computer Science 459-468.