Classification Problem Using Imbalanced Ratio Based Random Forest Method

Wan Wang¹, Junlin Guo² and Wai Kin Victor Chan¹*

¹Intelligent Transportation and Logistics Systems Lab, Tsinghua-Berkeley Shenzhen Institute, ShenZhen, China. wang-w18@mails.tsinghua.edu.cn
²Intelligent Transportation and Logistics Systems Lab, Tsinghua-Berkeley Shenzhen Institute, ShenZhen, China. guojl19@mails.tsinghua.edu.cn
¹*Intelligent Transportation and Logistics Systems Lab, Tsinghua-Berkeley Shenzhen Institute, ShenZhen, China. chanw@sz.tsinghua.edu.cn

Abstract. In recent decades, as machine learning is getting more and more popular in many fields, multiple classification methods have been proposed and applied to various applications for intelligent decision making. Most of the classification methods are just work on a shallow machine learning classifier. And many of these algorithms are already existed in the machine learning package. In this study, a new Imbalanced Ratio Based Random Forest(IR-RF) method was proposed to make classification and prediction. We focused on dealing with the imbalanced data. So we use several imbalanced datasets and put processed features into our IR-RF model. Three datasets we used are from the UCI machine learning database where RF and IR-RF are compared in the final results. The former RF method indicates the traditional random forest algorithm and the latter IR-RF method is our newly written one. Comparing to other classification model, IR-RF is an algorithm-level method, which is not limited by existing parameters. We can generate more new parameters within the algorithm and train our own classifier. Results show that proposed method has higher accuracy than traditional random forest method after our ratio compared and it can take better account of imbalanced condition. So proposed method not only performs well in high accuracy but also can be applied to especially imbalanced condition which is a new window for practice.

1. Introduction

Reasonable application of prediction algorithms and accurate classification of data are very important in many fields. A number of methods and experiments have been proposed to solve this problem and improvement has been done by many researchers. However, different classes of dataset are always biased in real scenarios which will cause serious impact on the performance of the classifier[1]. Development of open source implementations of Machine Learning algorithms (scikit-learn[2] and Weka[3]) has skyrocketed, leading to the less understanding of the principle and innovation. These observations motivated our work. In this study, we focused on dealing with the imbalanced data and utilize a new Imbalanced Ratio Based Random Forest(IR-RF) algorithm rather than using the open source package to deal with imbalanced classification problem. In our experiment, we found that the bootstrap ratio of different class could highly influence the accuracy and performance of the prediction
results and we can choose the best parameter on different imbalanced problems to achieve a better performance on the algorithm level.

Most of the study usually use a data-level method like SMOTE [4] and sampling method which include the under-sampling and over-sampling method[5] to deal with imbalanced problem. Other methods also include single machine learning model like Logistic Regression, Support Vector Machine (SVM), Decision Tree (DT) or some Bagging and Boosting method like Ada-boost, etc[6]. All these methods have been proposed in many open source packages that we can directly applied to our needs.

In addition, there are some studies focusing on dealing with problems using Random Forest. However, the main idea is to work before or after the main body part(random forest itself)[7]. Some researchers also proposed the weighted RF method, while there are also weight parameters that can be set in the existing package of the Random Forest.

In this study, we examine only binary classification tasks by using IR-RF model and adjust the bootstrap ratio of different class within the algorithm. We integrated a new parameter called bootstrap ratio in the process of forest construction. Then we train the IR-RF classifier and get different results of the ratio respectively. Also like other parameters in the grid search method, we can seek the suitable bootstrap ratio as we need.

In general, we can train classification model with many existing open source models. However, the performance of the proposed model can not consider the special condition of the real data and the parameters contained in the model can not be changed flexibly. In other words, the existing parameters may not meet our need in imbalanced conditions. To dig deeper for finding new classification strategies, algorithm-level method could be a good choice. Comparing to other classification model IR-RF is an algorithm-level method, which is not limited by existing parameters. We can generate more new parameters within the algorithm and train our own classifier.

The structure of this paper is arranged as follows. In Section 2, we will describe our proposed model framework. After that, we discuss three open-sourced dataset adopted in this work. In Section 3, we conducted the experiment and listed the results, then we analyze the results of different dataset and the performance of the prediction accuracy. As for Section 4 and Section 5, we conclude this paper and express our gratitude to our professor and our department.

2. Model Framework

Owing to the fast development of the Machine Learning open source packages, researchers can easily choose one to implement and build model to get the results. Our model aims to build our own IR-RF model which use the imbalanced bootstrap ratio embedded in random forest method to make prediction of the imbalanced datasets.

The process of our experiment are shown in Fig 1. Our model firstly deals with the data based on the given raw datasets. As there are some missing value and wrong data, we need to make some data pre-processing such as data cleaning and value filling for those null items using mean value of the target feature. Then we use these data to build the decision tree classifier and the random forest classifier. In this process, we implement our new building principles according to the newly defined parameters. As we can see that bootstrap parameter can only be set to True or False in the package. Our new parameter added the bootstrap ratio which can be set to 0.1--0.9. In this way, we can describe the model more precisely and considered the imbalanced condition in detail.

After we rewrite the new forest building structure, which is more complicated than only using the package, we can make prediction and see the impact and different results with various ratio. We also make comparison of the results between using RF and IR-RF method.
2.1. Dataset Description
There are numerous open-sourced datasets and we experiment with these three datasets of UCI[8]. We have a detailed look at three open datasets online and analyse their performance across the IR-RF algorithm. The datasets information are shown below in Table 1:

| dataset | min: maj | attributes | Total instances |
|---------|----------|------------|-----------------|
| ma      | 44:516   | 5          | 961             |
| wdbc    | 21:357   | 30         | 569             |
| spect   | 55:212   | 44         | 267             |

These three datasets record the following information respectively:
- spect dataset is divided into: training data (80 instances: 40 class_0 instances and 40 class_1 instances) and testing data (187 instances: 15 class_0 instances and 172 class_1 instances). The predicted attribute is the overall diagnosis which is binary. And each of the patients is classified either normal or abnormal.
- ma dataset can be used to predict the severity (benign or malignant) of cancer. It has 961 instances (445 class_0 instances and 516 class_1 instances) and 6 attributes. We use 80% of the data as train data and the other as test data.
- wdbc dataset which are computed from a digitized image. The predicted attribute is binary, which can be explained as benign or malignant. 80% of the data are train data and the other are test data.

2.2. Classification Algorithm
In our model, we choose the random forest algorithm as our baseline model. According to existing work, decision tree is commonly used to be the classifier to construct a general RF. RF has a better performance compared with other classification algorithms[9] as we can aggregate different trees to form a forest and make the final decision by selecting the majority decision of each trees.
The IR-RF algorithm is shown below:
1. For each iteration in random forest, draw a bootstrap sample from the majority class while randomly choose the majority class according to the certain bootstrap ratio that set.
2. Induce a classification tree base on these records
3. Repeat the steps above as the number desired. Aggregate the predictions and get the final results.

In our IR-RF, the parameters we set are listed in Table 2. The parameters not on the list means we just use the default value. The meaning of each parameter we set are also listed below.

Table 2. Parameters in IR-RF

| Parameter in IR-RF | Value |
|-------------------|-------|
| n_points          | 0.8   |
| n_features        | 0.8   |
| n_estimators      | 50    |
| ratio             | 0.1-0.9 |
| total             | 1000  |
| max_depth         | 5     |
| api               | True  |
| criterion         | gini  |
| min_samples_split | 100   |

explanation of some important parameters:
1. n_points: percentage of the total data we take each time
2. n_features: percentage of the total features we take each time
3. n_estimators: the number of trees in the forest.
4. ratio: bootstrap ratio of different class
   count_maj = total_num*ratio
   count_min = total_num*(1-self.ratio)
5. total: total number of data we choose each time
6. max_depth: the maximum depth of the tree
7. api: as we use our own written algorithm, we need to call the api and set it True
8. criterion: the function to measure the quality of a split with the Gini impurity
9. min_sample_split: the minimum number of samples required to split an internal node

3. Results and discussion
In recent studies, few of the researchers have paid attention to the algorithm itself. In our model, we use our own IR-RF algorithm to predict and compare our results with the traditional RF algorithm. In IR-RF, We've added a new adaptive parameter that can be set to different values which is different from traditional random forest algorithm. Different ratio will affect the final predicting results. There is always an optimal ratio and it can exceed the original accuracy that we call the package. The parameters in our algorithm are shown in our algorithm part and we use the 5-fold cross-validation to get our following results.

Accuracy Evaluation And Comparison
After processing with our data and training our model, prediction accuracy of three dataset are shown in Fig 2, Fig 3 and Fig 4 respectively. We implement proposed model in our own written random forest algorithm in python by 5-fold cross-validation to get the average accuracy to evaluate.
Table 3. Results Summary.

| Dataset | Ratio | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | Package |
|---------|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|---------|
| spect   | 0.694 | 0.759 | 0.759 | 0.792 | 0.789 | 0.837 | 0.863 | 0.856 | 0.830 | 0.754 |
| ma      | 0.714 | 0.767 | 0.771 | 0.791 | 0.791 | 0.803 | 0.803 | 0.752 | 0.742 | 0.772 |
| wdbc    | 0.895 | 0.904 | 0.895 | 0.912 | 0.895 | 0.930 | 0.929 | 0.957 | 0.930 | 0.913 |

Fig 2. accuracy comparison using different bootstrap ratio (wdbc dataset)

Fig 3. accuracy comparison using different bootstrap ratio (spect dataset)

Fig 4. accuracy comparison using different bootstrap ratio (ma dataset)

From Fig 2 to Fig 4, we can see that the accuracy changes with the change of ratio. We first look at the SPECT dataset, accuracy reaches a maximum value of 86.3% at the ratio of 0.8. The algorithm accuracy in the package can only reach 75.4% when all parameters are the same, which increased by 10.9%. And likewise, in ma dataset, the maximum accuracy reaches 80.3% at the ratio of 0.7 while the package only reaches 77.2%. In wdbc dataset, the maximum accuracy achieves at 95.7% contrast to the package accuracy at 91.3%. More intuitively, results summarization is shown in Table 3.

This improvement reveals that our new IR-RF algorithm can adaptively obtain the bootstrap ratio of according to different dataset. And we can search a best ratio like other parameters to help us obtain a more accurate result. Meanwhile, it is also a guidance and a new try for deeper research of ML.

4. Conclusion
In this study, we proposed an IR-RF method, which can set our own bootstrap ratio adaptively and achieve a higher accuracy than traditional random forest method. Also it takes better account of a variety of imbalanced condition by embedding the parameter of imbalanced bootstrap ratio to the
algorithm. We can change the ratio adaptively and obtain a more precise results according to different dataset and different imbalanced condition.

Imbalanced classification is a very common phenomenon in real world scenarios. Existing researches mainly focus on subsampling approach and using the existing algorithm in packages. Due to the popular application and good performance of Random Forest, we introduce the IR-RF method into the classification problem in this paper. The base classifier we used is the tree model. We generate thousands of trees in our new algorithm based on the new bootstrap method, and the final prediction can be computed. By implementing the algorithm from scratch, we will no longer be limited by the parameters and the existing algorithm in the package.

In real world scenarios, datasets are always imbalanced in various degrees, which tends to cause weak performance of the model if we only use the existing method. IR-RF method is a new method based on the tree model and embedded a new imbalanced bootstrap ratio parameter which considered the imbalanced condition of the dataset in real world. According to our experiment, the results have shown that the proposed method can improve the performance of the model and able to apply to imbalanced condition in algorithm level which is a new window for experiment and application.

Acknowledgments
The author would like to acknowledge the funding support of TBSI, Intelligent Transportation and Logistics Systems Lab and the support of Prof. Victor Chan for his guidance and help in the process of research.

References
[1] Chawla N, Japkowic N, Kotcz A, et al. Editorial: Special issues on learning from imbalanced data sets[J]. Annals of Nuclear Energy, 2004, 36(3):255-257.
[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Pret-tenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, Journal of Machine Learning Research 12, 2825 (2011).
[3] Frank E, Hall M, Trigg L, et al. Data mining in bioinformatics using Weka[J]. Bioinformatics, 2004, 20(15):2479-2481.
[4] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic Minority Over-sampling Technique[J]. Journal of Artificial Intelligence Research, 2002, 16(1):321-357.
[5] C. Drummond and R. Holte. C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling. In Proceedings of the ICML’03 Workshop on Learning from Imbalanced Data Sets, 2003.
[6] T. M. Khoshgoftaar, M. Golawala and J. V. Hulse, "An Empirical Study of Learning from Imbalanced Data Using Random Forest," 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007), Patras, 2007, pp. 310-317.
[7] Khalili M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest[J]. BMC Medical Informatics and Decision Making, 2011, 11(1):51-51.
[8] C. Blake, E. Keogh, and C. J. Merz, UCI Repository of Machine Learning Databases. Irvine, CA: Dept. Inf. Comput. Sci., Univ.California. [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html
[9] H. Valecha, A. Varma, I. Khare, A. Sachdeva and M. Goyal, "Prediction of Consumer Behaviour using Random Forest Algorithm," 2018 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), Gorakhpur, 2018, pp. 1-6.