Hierarchical ensemble learning method in diversified dataset analysis

Ze yuan Liu1,a*, Xin long Li2,b
1Department of Management Science, Soochow University, Suzhou, Jiangsu, China
2Nanyang Business School, Nanyang technological University, Singapore, Singapore
*a email: 1802409088@stu.suda.edu.cn
b email: xinlong.li@ntu.edu.sg

Abstract. The remarkable advances in ensemble machine learning methods have led to a significant analysis in large data, such as random forest algorithms. However, the algorithms only use the current features during the process of learning, which caused the initial upper accuracy’s limit no matter how well the algorithms are. Moreover, the low classification accuracy happened especially when one type of observation’s proportion is much lower than the other types in training datasets. The aim of the present study is to design a hierarchical classifier which try to extract new features by ensemble machine learning regressors and statistical methods inside the whole machine learning process. In stage 1, all the categorical variables will be characterized by random forest algorithm to create a new variable through regression analysis while the numerical variables left will serve as the sample of factor analysis (FA) process to calculate the factors value of each observation. Then, all the features will be learned by random forest classifier in stage 2. Diversified datasets consist of categorical and numerical variables will be used in the method. The experiment results show that the classification accuracy increased by 8.61%. Meanwhile, it also improves the classification accuracy of observations with low proportion in the training dataset significantly.

1. Introduction

Ensemble learning algorithms are widely used in classification or regression tasks since they were discovered. However, the widespread used ensemble learning algorithms cannot break the upper accuracy limit or lower error limit. Meanwhile, the accuracy to classify observations with low proportion in training dataset is lower than that each category of observation with almost equal proportion in many application scenarios.

Classifiers based on ensemble learning like random forest have reached relative high accuracy in the early year. The accuracy was almost the same as widely-used SVM in some application scenarios (M. Pal 2005) [1]. The ride-hailing industry also started to use ensemble learning algorithms to finish the task of demand forecasting (Y Jin, X Ye, Q Ye, T Wang, J Cheng, X Yan 2020) [2]. However, these applications of ensemble learning algorithms only use the known features without extracting new features, which caused the upper accuracy limit or lower error limit [3]. Recent years a large amount of hierarchical machine learning methods for mining new features have appeared. The accuracy of the method used in Pedestrian Detection has reached 98% and breaks the upper limit significantly (CHONG Yan-Wen1 KUANG Hu-Lin1 LI Qing-Quan 2012) [4]. But most research on classifiers didn’t focus on hierarchical ensemble learning methods and couldn’t be applied in the
car-hailing industry. Meanwhile, though some known ensemble learning algorithms can reach accuracy larger than 90% in some applications scenarios. The machine learning models should ensure that each category of observation has almost equal proportion (Fahd Saleh Alotaibi 2019) [5]. The accuracy of categories which has low proportion in the whole dataset is always sacrificed due to the upper accuracy limit.

This passage attempts to develop a hierarchical method of ensemble learning to break the upper accuracy limit of ensemble learning algorithms and increase the accuracy of categories with low proportion in the whole dataset. Moreover, this method should be applied in car-hailing industry. Datasets with diverse variables are more realistic and easier to extract features. Diversified datasets collected from real world’s car-hailing orders consist of both numerical and categorical variables. Previous research used random forest algorithm as an example of ensemble learning [6]. As the result, random forest is chosen as one of algorithms to extract features. Factor analysis is also a method to extract features in the study. Other ensemble learning algorithms like XGBoost are used for comparison. This method will improve the accuracy of recognizing offline taxis or online car-hailing significantly and contribute to reduce the dispatch cost of car-hailing companies.

2. Theoretical foundation
Machine learning is the scientific field focusing on the ways that machines analyze on historic data and produce meaningful conclusions automatically [7]. Ensemble learning algorithms are one type of machine learning algorithms. Random forest is the typical example of Ensemble learning [8]. Random forest output is the arithmetic mean value of each decision tree regressor or the voting result of each decision tree classifier [9]. XGBoost is also an example of ensemble learning, but the mathematical principles are quite different from random forest. It is a scalable tree boosting system [10].

Factor analysis is a statistical method to get the value of invisible features from known and correlated variables [11]. The factor analysis process can be considered as a multistep process depicted in Figure 1.

![Figure 1. the basic steps of factor analysis](image)

3. Method
The method used in this passage is a two-stage machine learning method to classify observations of ride-hailing. At the beginning we separate variables into categorical variables and numerical variables. Frequency feature (FF) is extracted from the first 7 categorical variables and trained by random forest regressor and Similarity feature (SF) is extracted from all 17 numerical variables left through factor analysis process. Both FF and SF will be added into the whole dataset and contribute to the classification on types of ride-hailing cars trained by Random Forest classifier. The whole flow chart of the proposed ride-hailing classification method depicted in Figure 2.
Figure 2. the whole flow chart of proposed ride-hailing cars trained by Random Forest.

3.1. dataset used in the study
This dataset is about ride-hailing cars. People may use applications like Uber to call ride-hailing cars or wait for taxis. The dataset records more 5 million cars dispatched by ride-hailing companies to finish orders.

The dataset used in the study has 693071 observations in total. Each observation stands for a ride-hailing record. There are 25 variables in this dataset. We focus on the type (name) of each ride-hailing record, so the variable named 'name' becomes independent variable. The other 24 variables left are dependent variables. The dataset is a diversified data which consist of categorical variables and numerical variables. The specific distribution of categorical variables and numerical variables are depicted in the following table.

Table 1. the proportion of two categories of variables

| type           | number | proportion |
|----------------|--------|------------|
| Categorical    | 7      | 29.1%      |
| Numerical      | 17     | 70.9%      |

3.2. feature extraction

3.2.1. FF Frequency feature (FF) is a feature based on 7 categorical variables. It describes the frequency of categorical variables combination.

**Step 1:** 7 categorical variables are transferred into sets. The elements of each set are the levels of each categorical variable. Then find the Cartesian Product of these sets. The elements in the Cartesian Product set represent all possible combination of these categorical variables.

\[ S = \prod_{i=1}^{7} S_i \]  

In this formula, \( S_i \) represents the \( i^{th} \) set among the 7 sets. \( S \) is the Cartesian Product of these sets.
Table 2. The element number of seven sets

| Categorical variable | Level number |
|----------------------|--------------|
| Hour                 | 24           |
| Day                  | 17           |
| Weekday              | 7            |
| Short_summary        | 9            |
| Cab_Type             | 2            |
| source               | 12           |
| destination          | 12           |
| Cartesian Product of 7 sets | 7402752 |

**Step 2:** Calculate the number of offline taxis dispatched by companies under the circumstance of each element in $S$. The relationship between number of taxis and Cartesian product of 7 categorical variables is depicted in the following formula:

$$\sum n(\prod_{i=1}^{7} S_i) = N$$  \hspace{1cm} (2)

In this formula, $N$ means the whole number of offline taxis. $n(x)$ is the number of taxis under current situation $x$.

**Step 3:** Once all the information known by this system, the Cartesian Product set and number of taxis serve as the training dataset and FF is the regression result of stage1’s regressor.

Table 3. Basic information of stage 1

| Independent variable | Dependent variables |
|----------------------|---------------------|
| Training dataset     | number of offline taxis under the circumstance of each element in $S$ |
| validation dataset   | FF Combination of 7 categorical variables |

FF is extracted by random forest regressor based on 10 decision tree regressors. Each CART regressor’s predicted value is averaged by equal weights.

3.2.2. **SF** Similarity feature (SF) is a feature based on 17 numerical variables. It describes the similarity of numerical variables combinations.

**Step 1:** All the numerical variables function as the sample factor analysis process. All the factors are recorded and sorted by eigenvalue corresponding to it. Then only factors whose accumulated eigen value is less than 80% will be remained for further calculation.

**Step 2:** each factor remained will be weighed. The weight of each factor is its explained variance ratio. The formula for explained variance ratio depicted in the following.

$$E_n = \frac{e_n}{\sum e_i}$$  \hspace{1cm} (3)

In this formula, $E_n$ is explained variance ratio of the $n$th factor and $e_n$ means that the $n$th factor’s eigenvalue corresponding to it.

**Step 3:** The SF value on each observation is the weight average of its factors’ value concluded from the factor analysis process. For example, the value on observation $n$ calculated through the following formula.

$$SF_n = \sum f_i * E_i$$  \hspace{1cm} (4)

In this formula, $f_i$ is the $i$th factor’s value on the $n$th observation and $E_i$ is the eigen value corresponding to the $i$th factor.

3.3. **classifier construction based on SF and FF**

Now the whole dataset used has 26 variables included 2 new features. These 2 new features serve as
the dependent variables. The suitable features which related to target output variables may improve the whole method’s accuracy significantly [12]. The Pearson correlations coefficient between features and independent variables depicted in the following table:

Table 4. the correlations coefficient between features and independent variables

| Features | Pearson’s r with independent variable |
|----------|---------------------------------------|
| FF       | -0.246229                             |
| SF       | -0.468590                             |

The specific distribution of variables after extracting features depicted in the following table.

Table 5. basic information of stage 2

| Variable      | Type                      | Number | Proportion |
|---------------|---------------------------|--------|------------|
| Independent   | Initial independent variables | 7      | 26.92%     |
| Dependent     | Initial dependent variables | 17     | 65.39%     |
| features      |                           | 2      | 7.69%      |

The division of training dataset and validation dataset are depicted in the following table.

Table 6. the division of training dataset and validation dataset

| Dataset        | Division ratio | NO. of subjects |
|----------------|----------------|-----------------|
| Training       | 70%            | 485,150         |
| Testing        | 30%            | 207,921         |

The random forest classifier used in stage 2 is voting result of 10 decision trees. Each tree has the same weight in the whole voting process.

4. Simulation experiment

4.1. experiment environment

This experiment is implemented by python 3.8. The computer configuration is depicted in table 7. Both training dataset and testing dataset come from a public database named kaggle.com.

Table 7. hardware environment to implement the method

| Hardware        | Hardware model         |
|-----------------|------------------------|
| CPU             | Intel core i7 CPU 3.2 GHZ |
| Main memory     | 16 G                   |

4.2. experiment result and results of other one-stage ensemble learning methods

Table 8. the comparison results of different ensemble learning methods

| Method          | NO. of subjects | Accurate NO. of subjects | Best accuracy of both sides | Best accuracy of offline taxis |
|-----------------|-----------------|--------------------------|-----------------------------|--------------------------------|
| Hierarchical    | 207922          | 207905                   | 99.98%                      | 99.98%                         |
| XGBoost         | 207922          | 207905                   | 92.05%                      | 0.00%                          |
| Random forest   | 207922          | 207905                   | 92.07%                      | 0.37%                          |

From the perspective of the experiment result, the method can correctly classify more than 99% observations. The accuracy of Hierarchical method improves more than 8.61% compared to other ensemble learning methods. Moreover, this method still returns quite accuracy more than 99% regardless the proportion of taxis in the whole training dataset. The accuracy of offline taxis breaks 1% up to 99%. The proportion of taxis is only about 7.94% among the whole dataset used.

5. Conclusions

This passage includes a hierarchical machine learning method to solve the classification of online car-hailing or offline taxis. Based on the features extracted separately from categorical and numerical
variables, the method breaks the upper accuracy limit of one-stage ensemble learning algorithms. It improves the accuracy by 8.61% compared to one-stage known ensemble learning classifier (random forest and XGBoost classifier). It can be used as a well-performed method to predict taxi or not taxi under specific circumstance. The method also improves the accuracy to classify offline taxis which is only 7.94% among the dataset used. The accuracy to classify observations with low proportion in training dataset achieves exponential growth from less than 1% to 99%. This method can deal with diversified dataset consisting of both categorical and numerical dataset very well, which has a wide range of application scenarios.

The hierarchical method faces disadvantage over other one-stage ensemble learning methods too. The mean runtime of this method is 184.6s. The mean runtime of Random forest is 10.3s and that of XGBoost is 24.1s.

Future research may be carried out depicted in the following 1) In the process of calculating FF, some dimensionality reduction statistical methods such as PCA will be used to reduce mean time of the method. 2) Try to explain the actual meaning of FF and SF clearly 3) try to extract more features with high quality which may contribute to improve the accuracy and mean time of the method.

Acknowledgments
We would like to thank Mohamad Kazem Shirani Faradonbeh for his valuable tutorial on statistical machine learning. We sincerely thank David Cournapeau and all other contributors in Python Scikit-learn module. Also, we would like to thank our parents for their support in our work.

References
[1] Pal, M. (2007) Random forest classifier for remote sensing classification. International journal of remote sensing, 26:217-222.
[2] Jin, Y., Ye, X., Ye, Q., Wang, T., Cheng, J., Yan, X. (2020) Demand Forecasting of Online Car-Hailing With Stacking Ensemble Learning Approach and Large-Scale Datasets.IEEE Access, 8: 199513-199522.
[3] Kira, K., Rendell, L.A. (1992) Machine learning proceedings 1992.Elsevier, The Netherlands.
[4] Chong, Y., Kuang, H., Li, Q. (2012) Two-stage Pedestrian Detection Based on Multiple Features and Machine Learning.ACTA AUTOMÁTICA SINICA, 38:375-381.
[5] Alotaibi, F.S. (2019) Implementation of Machine Learning Model to predict Heart Failure Disease.IJACSA, 10:261-268.
[6] Zhang, C., Ma, Y. (2012) Ensemble Machine Learning.Springer, New York City.
[7] Jordan, M.I., Mitchell.T.M. (2015) Machine learning: Trends, perspectives, and prospects.Science, 349: 255-260.
[8] Dietterich, T.G. (2002) Ensemble learning.The handbook of brain theory and neural networks, 2:110-125.
[9] Breiman, L. (2001) Random Forests.Machine learning, 45: 5-32.
[10] Chen, T., Guostrin, C. (2016) XGBoost: A Scalable Tree Boosting System .In: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco. pp. 785-794.
[11] Spearman, C. (1904) The Proof and Measurement of Association Between Two Things. American Journal of Psychology, 15:72-101.
[12] Blum, A.L., Langley, P. (1997) Selection of relevant features and examples in machine learning.Artificial Intelligence, 97:245-271.