Research on Recommendation of Insurance Products Based on Random Forest

Yan Guo
College of Information Engineering, Sichuan Agricultural University, Yaan, Sichuan, China
Key Laboratory of Agricultural information engineering of Sichuan Province, Sichuan Agricultural University, Yaan, Sichuan, China
guoyan@sicau.edu.cn

Xiaonan Hu
College of Information Engineering, Sichuan Agricultural University, Yaan, Sichuan, China
hu15515894667@126.com

Yu Zhou
College of Information Engineering, Sichuan Agricultural University, Yaan, Sichuan, China
1074741743@qq.com

Wenchuan Cheng*
College of Information Engineering, Sichuan Agricultural University, Yaan, Sichuan, China
3154905097@qq.com

***Abstract*—With the rapid development of recommendation system, how to predict user's behavior accurately become more and more important. In this paper, random forest is applied to recommend insurance products and compared with ID3, C4.5, Nave-Bayes and Nearest-neighbor. Experiment results show that the prediction error of random forest is 2.02% lower than ID3, 1.09% lower than C4.5, 1.67% lower than Nave-Bayes and 5.97% lower than Nearest-neighbor. Therefore, it is highly feasible to recommend insurance products with random forests.**

Keywords- random forest; insurance products; recommendation

I. INTRODUCTION

There are many existing recommendation methods. Traditional recommendation methods include collaborative filtering algorithm, content-based recommendation algorithm, bipartite graph-based recommendation algorithm, Association rule-based recommendation algorithm, matrix decomposition-based recommendation algorithm and hybrid recommendation method. Research on traditional recommendation methods has been very in-depth, and traditional recommendation methods have also been widely used. However, these methods more or less encounter some difficult problems, such as data sparseness and cold start, recommendation time based on association rules consumes a lot and the degree of personalization is low.

In recent years, machine learning has achieved great success in image processing, natural language understanding and speech recognition. At the same time, machine learning technology is more and more widely used in recommendation systems. Random forest is an important method in machine learning. It has the advantages of high accuracy, strong robustness and wide application range. In this paper, random forest is applied to the prediction of insurance products and recommended according to the prediction results.

II. RESEARCH SUMMARY

A. Bootstrap

Bootstrap was proposed by Efron B [1], the method was widely used in the 1980s. Over the years, many scholars have made many improvements to the Bootstrap method. Diciccio T J discussed the confidence interval of Bootstrap. Based on the techniques of transformation and deviation correction, a more accurate confidence interval was obtained with only increasing the amount of calculation. In addition, bootstrap interval was developed for non-parametric conditions [2]. Andrews D W K proposed a three-step method to select an appropriate number of bootstraps for calculating bootstrap statistics and error correction [3]. Based on Bootstrap, Hao W proposed a method for estimating and predicting the confidence intervals of Web services. Experiments show that the confidence intervals follow the heavy-tailed distribution [4].

B. CART

CART (Classification and Regression Tree) was proposed by Breiman L in the early 1980s [5]. CART algorithm is an epoch-making progress in the field of tree model. It is also the most widely used tree model algorithm at present. It can not only create classification trees to deal with classification problems, but also create regression trees to deal with regression problems. Rutkowski L proposed a new algorithm based on the CART. This algorithm can determine the best attributes to split at a node, thus reducing the time required to establish the CART [6]. Pakgohar A applied CART to the analysis of human factors in traffic accidents. It was found that human factors such as driver's license and seat belt played a significant role in the severity of serious traffic accidents in Iran [7].

978-1-7281-5094-9/19/$31.00 ©2019 IEEE
DOI 10.1109/MLBDBI48998.2019.00069

C. Bagging

Bagging is a combined classification method of base classifiers constructed by bootstrap method proposed by Breiman L [8]. When using subsets of classification, regression tree and linear regression to process data sets, Bagging can greatly improve the accuracy. Wu W proposed a data mining method consisting of K-means clustering and bag-filling neural network for short-term wind power forecasting, which can process the dynamics of training samples and improve the prediction accuracy [9]. Tien Bui D used bagging to simulate rainfall-induced landslides. It was found that bagging had better prediction performance than AdaBoost and MultiBoost when they were used to simulate rainfall-induced landslides [10].

D. Random Forest

Random Forest is an integrated learning method based on bagging algorithm, which can be used for classification, regression and so on. Breiman L combines bagging method with Ho T K's random subspace method, and then proposed a new machine learning algorithm, random forest algorithm [11][12]. Random forest overcomes the problem of over-fitting, and has high robustness to noise and outliers, it also has good scalability and parallelism for high-dimensional data classification. Hu X developed a random forest model to estimate PM2.5 concentration. Experiments show that the prediction accuracy of the model is similar to that of the neural network, but simpler [13]. Chen J introduced PRF (parallel random forest) algorithm for large data on Apache Spark platform, and verified that PRF algorithm is superior to Spark MLlib and other research in classification accuracy, performance and scalability [14]. Paul A proposed a new random forest, which uses the least number of trees to classify by eliminating some unimportant attributes. A maximum number of trees will be added to the forest to ensure the accuracy of classification [15].

III. MODEL ESTABLISHMENT

A. Bootstrap

For a population F with unknown distribution, a sample with a capacity of n is extracted according to the sampling method with replacement, which is called Bootstrap sample. Bootstrap method continuously and independently extracts n Bootstrap samples from the population, and uses these samples to infer the population F.

Let $X_i = x_i, i = 1,2,\dots,n,$ Among them, X_i is independent and identically distributed and obeys unknown distribution F, the process of inferring the total parameter θ by Bootstrap method is as follows:

(1) Empirical function F is constructed based on sample observation value X_i.

(2) Extracting sample $X^* = [x_1{}^*, x_2{}^*, \dots, x_n{}^*]$ from F with replacement, called Bootstrap sample.

(3) Using Bootstrap Distribution of $\widehat{\theta^*} = \mathrm{S}(X^*)$ to approximate estimate the sample distribution of $\theta = \mathrm{S}(\mathrm{X})$.

(4) For $x_i{}^*$, the estimated value of θ is $\widehat{\theta_i} = \mathrm{S}(x_i{}^*)$.

B. CART

The basic idea of CART algorithm is to select the attributes with the minimum GINI coefficient value as the splitting attributes. According to the splitting attributes of nodes, the current sample set is divided into two sub-sample sets by using the technology of binary recursive segmentation, and a simple binary tree is formed recursively.

In CART algorithm, the smaller the GINI coefficient is, the more reasonable the division is. For sample set F, suppose it contains n classes and p_i is the probability that the i-th class included in F, then the GINI coefficient of sample set is:

$$\mathrm{GINI(F)} = 1 - \sum_{i=1}^{n} p_i{}^2 \quad (1)$$

If F is divided into subsets F_1 and F_2, the GINI coefficient is:

$$\mathrm{GINI}(F_1, F_2) = |\frac{F_1}{F}|\mathrm{GINI}(F_1) + |\frac{F_2}{F}|\mathrm{GINI}(F_2) \quad (2)$$

The implementation of CART classification can be divided into four steps, the process is as follows:

(1) Computing GINI coefficients of attributes in attribute set, then the attribute with the smallest GINI coefficient is selected as the split attribute of the root node.

(2) If the splitting attributes are continuous attributes, let T be the thresholds of splitting. Sample sets with values greater than T on attributes are classified into one class, and sample sets with values less than T on attributes are classified into one class.

(3) If the value of the sample set on the attribute is included in the true subset with the minimum GINI coefficient of the discrete attribute, it is divided into one part. If the value of the sample set on the attribute is not included in the true subset with the minimum GINI coefficient of the discrete attribute, it is divided into one part.

(4) Pruning the generated decision tree.

C. Bagging

The basic idea of bagging is that given a weak learning algorithm and a training data set D, random sampling with multiple replacement is carried out. Each sampling takes n samples from the original training set D to form a subset of the training data set. Some samples may appear multiple times, and some samples may not appear at one time. Each training subset will get a base classifier. For an unknown data set, the results obtained by each base classifier are counted as a vote, and the final classification results are selected by counting votes. The bagging algorithm flow is as follows:

Let data sets $\mathrm{D} = d_i, i = 1,2,\dots,n$

(1) Extracting m samples D_j from D with replacement.

(2) Get the results of sample D_j on the base classifier.

(3) Repeat steps (1), (2) for K times.

(4) Make statistics and return the final classification results.

D. Random Forest

Random forest is an integrated learning method based on Bagging. Its essence is to apply bootstrap method to CART algorithm. Random forests are sampled by bootstrap method, and then independent decision tree models are constructed by CART algorithm. The trees in the forests are not pruned. Finally, all decision trees are combined to form random forests. After the construction of random forest, the forest is used for analysis. If the dependent variable is the classification variable, the voting method is used to obtain the mode of all CART decision tree model prediction results as the final classification result. If the dependent variable is a continuous variable, the average of each CART decision tree model prediction results is taken as the final classification result.

Let data sets $D = d_i, i = 1,2, \ldots, n$, and there are m decision trees as base classifiers. The generation of random forests is shown in Fig. 1:

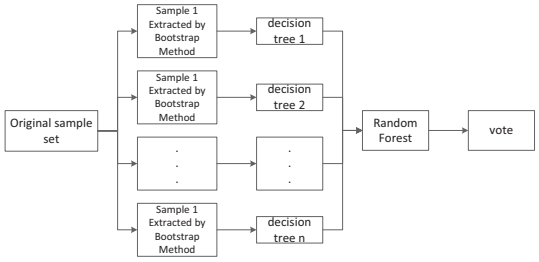

Figure 1. The generation of random forests

IV. EXPERIMENT

A. Data Source and Processing

The data set used in this experiment is from insurance data of insurance companies, including age, education, relationship, occupation, marital status, working hours per week, workclass, nationality, race and sex. Firstly, data cleaning is carried out to remove invalid data, and then the data is digitized for random forest modeling. Some experimental data are shown in Fig. 2:

age	workclass	education	marital_status	occupation	relationship	race	sex	hours_per_week	native_country	Insurance_purchase
39	7	10	1	8	2	1	1	40	1	0
50	5	10	2	4	4	1	1	13	1	0
38	3	14	6	6	2	1	1	40	1	0
53	3	7	2	6	4	4	1	40	1	0
28	3	10	2	5	3	4	0	40	13	0
37	3	13	2	4	3	1	0	40	1	0
49	3	5	4	14	2	4	0	16	19	0
52	5	14	2	4	4	1	1	45	1	1
31	3	13	1	5	2	1	0	50	1	1
42	3	10	2	4	4	1	1	40	1	1
37	3	9	2	4	4	4	1	80	1	1
30	7	10	2	5	4	2	1	40	8	1
23	3	10	1	8	5	1	0	30	1	0
32	3	11	1	3	2	4	1	50	1	0
34	3	4	2	10	4	3	1	45	21	0
25	5	14	1	9	5	1	1	35	1	0
32	3	14	1	7	1	1	1	40	1	0
38	3	7	2	3	4	1	1	50	1	0

Figure 2. Data

B. Analysis of experimental results

Cluster analysis of data by SPSS24 showed that the error rate of out-of-pocket estimation was 15.45%. The importance of random forest variables is shown in Fig. 3:

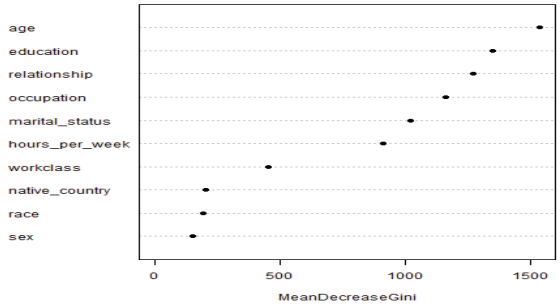

Figure 3. The importance of random forest variables

As can be seen from figure 3, age, education, relationship, occupation, marital status and working hours per week are of high importance, while workclass, nationality, race and sex are of low importance.

Only age, education, relationship, occupation, marital status and working hours per week. The error rate of out-of-pocket estimation was found to be 16%. It shows that the reduction of four predictive variables only reduces the accuracy of prediction by 0.55%. It can be predicted by six variables: age, education, relationship, occupation, marital status and working hours per week.

C. comparative analysis

In this paper, ID3, C4.5, Nave-Bayes and Nearest-neighbor are used for classification and recommendation. The error rates of each model are shown in Table 1:

Table 1. ERROR RATES

model	Random-forest	ID3	C4.5	Naïve-Bayes	Nearest-neighbor
Error rate	16%	17.47%	16.54%	17.12%	21.42%

As can be seen from table 1, the prediction errors of random forests are lower than those of ID3, C4.5, Nave-Bayes and Nearest-neighbor. The prediction error of random forest is 2.02% lower than ID3, 1.09% lower than C4.5, 1.67% lower than Nave-Bayes and 5.97% lower than Nearest-neighbor. Therefore, the recommendation based on random forest is superior to ID3, C4.5, Nave-Bayes and Nearest-neighbor in accuracy.

V. CONCLUSION

Traditional recommendation algorithms encounter problems such as cold start and sparse data when recommending goods. Machine learning has been applied to recommendation system by more and more researchers, and has shown good recommendation effect. In this paper, random forest is applied to recommend insurance products. Experiments show that the prediction error of random forest is lower than ID3, C4.5, Nave-Bayes and Nearest-neighbor. The prediction error rate of random forest is 1.47% lower than ID3, 0.54% lower than C4.5, 1.12% lower than Nave-Bayes and 5.42% lower than Nearest-neighbor. This shows that random forest has high feasibility in insurance product prediction.

ACKNOWLEDGMENT

The authors are extremely grateful to the journal editorial team and reviewers who provided valuable comments for improving the quality of this article. This work was supported by Key Laboratory of Agricultural information engineering of Sichuan Province, Research Topic of Government Affairs in Sichuan Province in 2019 (Research on the Development of Intelligent Agriculture in Sichuan Province) and Social Science Foundation of Sichuan Province in 2019(19GL030).

REFERENCES

[1] Efron B. Bootstrap methods: Another look at the jackknife[J]. Ann. Stat. 1979, 7(1): 1-26.

[2] Diciccio T J, Efron B. Better Bootstrap Confidence Intervals[J]. Journal of the American Statistical Association, 1996, 11(3): 189-228.

[3] Andrews D W K, Buchinsky M. A Three-Step Method for Choosing the Number of Bootstrap Repetitions[J]. Econometrica, 2000, 68(1): 23-51.

[4] Hao W, Jian-Mao X, Hao L, Le-Yue W, Software S O, University J N. An Approach to Estimate and Predict the Confidence Interval of Web Service QoS Based on Bootstrap[J]. Acta Electronica Sinica, 2018, 46(3): 665-671.

[5] Breiman L, Friedman J, Olshen R, Stone C J. Classification and Regression Trees[M]. NewYork: Chapman&Hall, 1984.

[6] Rutkowski L, Jaworski M, Pietruczuk L, Duda P. The CART decision tree for mining data streams[J]. Information Sciences, 2014, 266: 1-15.

[7] Pakgohar A, Tabrizi R S, Khalili M, Esmaeili A. The role of human factor in incidence and severity of road crashes based on the CART and LR regression: a data mining approach[J]. Procedia Computer Science, 2011, 3(none): 764-769.

[8] Breiman L. Bagging Predictors[J]. Machine Learning, 1996, 24(2): 123-140.

[9] Wu W, Peng M. A Data Mining Approach Combining K-Means Clustering with Bagging Neural Network for Short-term Wind Power Forecasting[J]. IEEE Internet of Things Journal, 2017: 1-1.

[10] Tien Bui D, Ho T C, Pradhan B, Pham B T, Nhu V H, Revhaug I. GIS-based modeling of rainfall-induced landslides using data mining-based functional trees classifier with AdaBoost, Bagging, and MultiBoost ensemble frameworks[J]. Environmental Earth Sciences, 2016, 75(14): 1101.

[11] Breiman L. Random Forests[J]. Machine Learning, 2001, 45(1): 5-32.

[12] Ho T K. The random subspace method for constructing decision forests[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(8): 0-844.

[13] Hu X, Belle J H, Meng X, Wildani A, Waller L, Strickland M, et al. Estimating PM2.5 Concentrations in the Conterminous United States Using the Random Forest Approach[J]. Environmental Science & Technology, 2017: acs.est.7b01210.

[14] Chen J, Li K, Member S, Tang Z. A Parallel Random Forest Algorithm for Big Data in a Spark Cloud Computing Environment[J]. IEEE Transactions on Parallel and Distributed Systems, 2017, 28(4): 919-933.

[15] Paul A, Mukherjee D P, Das P, Gangopadhyay A, Chintha A R, Kundu S. Improved Random Forest for Classification[J]. IEEE Transactions on Image Processing, 2018: 1-1.