Mining Bigdata in Context with Random Forest using Machine Learning

Pratibha[1], B. Sowjanya[2], S. Mounasri[3]
1 Department of Computer Science and Engineering
Sridevi Women's Engineering college, Hyderabad, India.
2 Department of Computer Science and Engineering
Sridevi Women's Engineering college, Hyderabad, India.
3 Department of Computer Science and Engineering
Sridevi Women's Engineering college, Hyderabad, India.

Abstract— Due to the imbalanced distribution of business data missing user features and other purpose, usage of big data methods on commercial facts tends to deviate from the business goals. It is difficult to classical the insurance commercial facts through classification algorithms such as logistic regression and support vector machine svm . We exploit a heuristic bootstrap selection method combined with the ensemble learning algorithm on the large scale insurance business data mining and propose an ensemble random forest algorithm that uses the parallel computing capability and memory cache we composed the insurance commercial facts from china life insurance Company to evaluate the probable consumers via the suggested algorithm. We custom f measure and g mean to estimate the recital of the algorithm. Testing outcome indicates that the ensemble random forest algorithm outstripped svm and further classification algorithms in both performance and accuracy in the imbalanced data and it is useful for improving the accuracy of product marketing compared to the customary simulated method.

Index Terms— SVM, RFA, F- measure, G- mean.SS

1 Introduction

With the arrival of the era of big data, the third industrial revolution represented by information technology opened a new chapter. Big data technology was widely applied. In the academic community, the respected journals "Nature" and "Science" have respectively launched big data issues named "Big Data" and "Deal With Data", which discuss a variety of problems encountered in big data technology from the Internet technology, economics, supercomputing, biological sciences, medicine and many other aspects. In the industry, whether gene sequencing, biological medicine and other life sciences, or banking, insurance and other traditional Financial sector, are all driven by big data technology to enter a new round of science and technology competition. Big data technology does not only create significant value but also promote the change and progress of traditional industries.

The traditional marketing method of selling insurance is mainly based on off-line sales business. Insurance salesmen sell the company's products by calling or visiting the customers. This blind marketing way has achieved good results in the past, which maintained the company sales performance for a long time through widespread sales. With the gradual opening of the insurance industry, a large number of private insurance companies enter the market, which forms a healthy competitive environment and constantly promote the reform of the insurance industry. On the other hand, people's willingness to purchase insurance gradually increased, the potential insurance customers are rapidly expanding. According to statistics, the success rate of the traditional telephone sale is less than one thousandth, and the insurance sales rate of a senior insurance salesmen can reach about two percent, but this is obviously very inefficient. Therefore, how to better accurately understand the users' purchase intention has become a very urgent need for the insurance company.

With the development of big data technology, the traditional financial services industry is eager to find a breakthrough driven by the big data wave. Achieving targeted marketing has become the primary objective of many financial industries, and financial big data has become one of the hot spots in the social development of today. Data mining combined with big data technology has become a support technology of traditional financial and insurance industry transformation. Due to the lack of purpose and innovation of traditional marketing methods, the poorly organized insurance business data and obscure customers' purchasing characteristics directly lead to a serious imbalance in the category of product data, which bring difficulties to user classification and recommendation of insurance products.

Classification of imbalanced data sets has puzzled many researchers. In real life, we could not get the expected distribution of data because of various reasons, especially in some cost sensitive business scenarios. For the unbalanced distribution of data in the same sample space, we usually choose some resampling methods which sacrifice some features to construct relatively balanced training data sets. In addition, we can also construct the virtual samples to balance the data distribution. As a result, we improve the recognition rate of the minority class that is recall rate but sacrifice the precision of the classification model.

1.1 **Machine learning**

Machine learning is a field of computer science that provides computer systems the capability to "study" (i.e., progressively improve performance on a specific task) through data, without being unambiguously programmed.

The name Machine learning was coined in 1959 by Arthur Samuel. Developed from the reading of pattern recognition and computational learning scheme in artificial intelligence, machine learning discovers the learning and building of algorithms that can acquire since and make forecasts on facts – such algorithms overcome subsequent firmly static program instructions by creating data-driven forecasts or results, over structuring a model from sample inputs. Machine learning is active in a variety of computing tasks where designing and programming open algorithms with good performance is challenging or infeasible.

1.2 **Objective**

The main purpose is providing a novel classification model for traditional insurance business data base on the background of insurance industry reform, combined with the big data technologies. This paper does not only provide a good strategy for the orientation of precise marketing of insurance products, but also has a very good reference for the classification of imbalanced data sets. This paper is organized as follows. Section II introduces the current research status of imbalanced data classification; Section III puts forward the classification model and intelligent recommendation algorithm based on random forest for insurance business data, and analyzes its efficiency; Section IV applies the proposed algorithm to the insurance product business data of China Life insurance company and successfully analyze potential customers and the distribution of their major characteristic.

2 What is Random Forest Algorithm?

What is Random Forest Algorithm? Random forest algorithm is a supervised classification algorithm. As the term propose, this algorithm produces the forest with a number of trees. In general, the more trees in the forest the more robust the forest looks like. In the similar mode in the random forest classifier, the higher the number of trees in the forest provides the high accuracy outcomes.

2.1 Why Random forest algorithm?

- The same random forest algorithm or the random forest classifier can use for both classification and the regression task.
- Random forest classifier will handle the missing values.
- When we have more trees in the forest, random forest classifier won't overfit the model.

2.2 How Random forest algorithm works?

The pseudocode for random forest algorithm can divided into two phases.

- Random forest creation pseudocode.
- Pseudocode to achieve prediction from the produced random forest classifier.

2.3 Random Forest pseudocode:

Arbitrarily select k features from total m features Where k << m.

1. Among the k features compute the node d by the finest split fact.

2. Split the node into daughter nodes using the finest split.

3. Repeat 1 to 3 steps until l number of nodes has been reached.

4. Form forest by iterating steps 1 to 4 for n number times to produce n number of trees.The creation of random forest algorithm starts with arbitrarily choosing k features due to total m features. In the image you can observe that we are randomly taking features and observations.

Now the following stage we are exhausting the arbitrarily selected k features to discover the source node by using the finest split method.

The next stage we will be calculating the daughter nodes using the same best split approach. Will the first 3 stages until we form the tree with a root node and having the target as the leaf node.

Lastly we repeat 1 to 4 stages to generate n arbitrarily produced trees. This arbitrarily produced tree forms the random forest.

Random forest prediction pseudocode

To achieve prediction by the skilled random forest algorithm uses the below pseudocode.

1. Takes the examination features and practice the rubrics of all arbitrarily generated decision tree to predict the outcome and stores the predicted outcome (target)
2. Calculate the votes for each predicted target.
3. Consider the high voted predicted objective as the last prediction from the random forest algorithm.

To achieve the prediction using the skilled random forest algorithm we requisite to permit the test features over the rubrics of each arbitrarily created trees. suppose let s say we formed 100 random decision trees to from the random forest.

Every random forest will predict dissimilar target result for the similar test feature. Then by considering each predicted target votes will be calculated. suppose the 100 random decision trees are prediction some 3 unique targets x y z then the votes of x is nothing but out of 100 random decision tree how many likewise for other 2 targets y z . if x is getting high votes.

Let s say out of 100 random decision tree 60 trees are predicting the target will be x. then the final random forest profits the x as the predicted target. The random algorithm used in extensive varieties applications. In this article we are going address few of them.

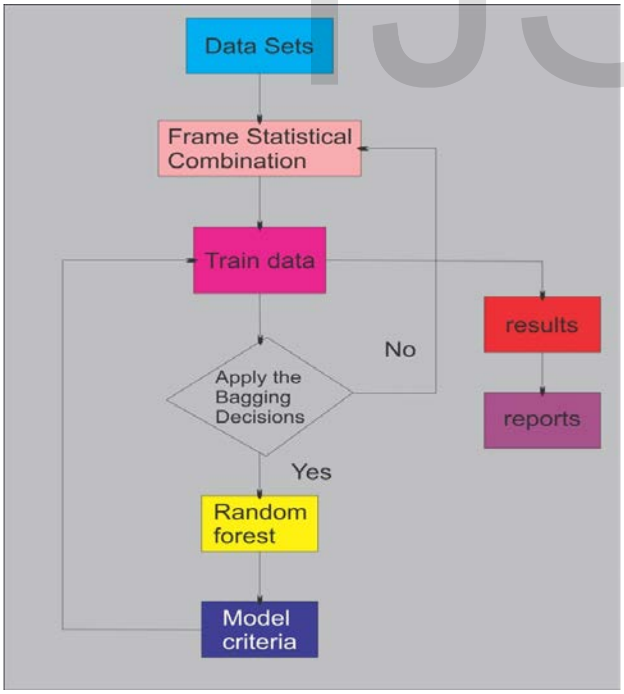
Below are some the application where random forest algorithm is widely used.

- **Banking**
- **Medicine**
- **Stock Market**
- **E-commerce**

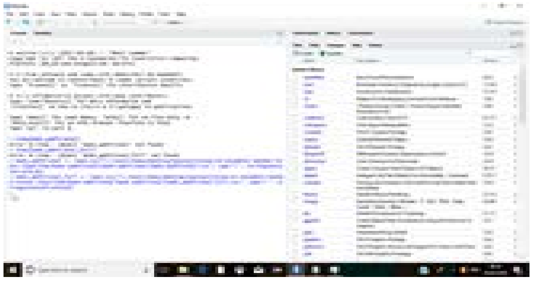
2.4 Advantages of random forest algorithm

- The over fitting problem will never come when we practice the random forest algorithm in several classification problem.
- The similar random forest algorithm can be used for both classification and regression task.
- The random forest algorithm can be used for feature engineering. This means identifying the most important features out of the available features from the training dataset.

2 Architecture Diagram:



3 Results



- Initially we check whether our packages are available or not. if not available the we should packages.
- We should derive the path of our data.

4 Conclusions

we analyzed the imbalance distribution of insurance business data concluded the preprocessing algorithms of imbalance dataset proposed an ensemble random forest algorithm based on apache spark which can be used in the large scaled imbalanced classification of insurance business data the experiment result showed that the ensemble random forest algorithm is more suitable in the insurance product recommendation or potential customer analysis than traditional strong classifier like SVM and logistic regression etc. the proposed bootstrap under sampling algorithm combined with the KNN could be used into preprocessing of imbalanced classification algorithms. The ensemble learning algorithms combined with bootstrap sampling preprocessing could reduce the learning process further and it also has a good reference to other imbalanced although the proposed ensemble random forest algorithm is used to analyze insurance big data in this paper it can also be applied to big data analytics for internet of things finance and mobile internet.

4.1 Future Enhancement

- Exploring the proposed algorithm to different types of big data analytics
- Combining deep learning into the proposed algorithm to improve the accuracy of prediction based

.

5 References

[1] K. W. Bowyer, L. O. Hall, and W. P. Kegelmeye, ``SMOTE: Synthetic minority over-sampling technique,'' J. Artif. Intell. Res., vol. 16, no. 6, pp. 321357, 2002.

[2] E. Ramentol, Y. Caballero, R. Bello, and F. Herrera, ``SMOTE-RSB: A hybrid preprocessing approach based on oversampling and under sampling for high imbalanced data-sets using SMOTE and rough sets theory,'' Knowl. Inf. Syst., vol. 33, no. 2, pp. 245265, 2012.

[3] J. A. Sáez, J. Luengo, J. Stefanowski, and F. Herrera, ``SMOTE-IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering,'' Inf. Sci., vol. 291, pp. 184203, Jan. 2015.

[4] E. Ramentol, N. Verbiest, R. Bello, Y. Caballero, C. Cornelis, and F. Herrera, ``SMOTE-FRST: A new resampling method using fuzzy rough set theory,'' in Proc. 10th Int. FLINS Conf. Uncertainty Modelling Knowl. Eng. Decision Making, pp. 800805, 2012.

[5] G. Ping and O. Y. Yuan-You, ``Classification research for unbalanced data based on mixed-sampling,'' Appl. Res. Comput., vol. 32, no. 2, pp. 379381, 2015.

[6] I. Tomek, ``Two modifications of CNN,'' IEEE Trans. Syst., Man Cybern., vol. 6, no. 11, pp. 769772, Nov. 1976.

[7] M. A. Tahir, J. Kittler, and F. Yan, ``Inverse random under sampling for class imbalance problem and its application to multi-label classification,'' Pattern Recognit., vol. 45, no. 10, pp. 37383750, 2012.

[8] F. Angiulli, ``Fast condensed nearest neighbor rule,'' in Proc. 22nd Int. Conf. Mach. Learn. ACM, 2005, pp. 2532.

[9] J. Laurikkala, Improving Identification of Difficult Small Classes by Balancing Class Distribution. Berlin, Germany: Springer, 2001.

[10] H. Han, W. Y. Wang, and B. H. Mao, ``Borderline-SMOTE: A new oversampling method in imbalanced data sets learning,'' in Advances in Intelligent Computing. Berlin, Germany: Springer, 2005, pp. 878887.

[11] J. H. Friedman and P. Hall, ``On bagging and nonlinear estimation,'' J. Stat. Planning Inference, vol. 137, no. 3, pp. 669683, 2007.

[12] S. Hido, H. Kashima, and Y. Takahashi, ``Roughly balanced bagging for imbalanced data,'' Stat. Anal. Data Mining, vol. 2, nos. 56, pp. 412426, 2009.

[13] R. S. Del, V. López, J. M. Benítez, and F. Herrera, ``On the use of MapReduce for imbalanced big data using random forest,'' Inf. Sci., vol. 285, pp. 112137, Nov. 2014.

[14] R. C. Bhagat and S. S. Patil, ``Enhanced SMOTE algorithm for classification of imbalanced big-data using random forest,'' in Proc. IEEE Int. Adv. Comput. Conf. (IACC), Jun. 2015, pp. 403408.