An Ensemble Random Forest Algorithm for Insurance Big Data Analysis

Weiwei Lin[1], Ziming Wu[1], Longxin Lin[2], Angzhan Wen[1], and Jin Li[3]

***Abstract*—Due to the imbalanced distribution of business data, missing of user features and many other reasons, directly using big data techniques on realistic business data tends to deviate from the business goals. It is difficult to model the insurance business data by classification algorithms like Logistic Regression and SVM etc. In this paper, we exploit a heuristic bootstrap sampling approach combined with the ensemble learning algorithm on the large-scale insurance business data mining, and proposes an ensemble random forest algorithm which used the parallel computing capability and memory-cache mechanism optimized by Spark. We collected the insurance business data from China Life Insurance Company to analyze the potential customers using the proposed algorithm. We use F-Measure and G-mean to evaluate the performance of the algorithm. Experiment result shows that the ensemble random forest algorithm outperformed SVM and other classification algorithms in both performance and accuracy within the imbalanced data, and it is useful for improving the accuracy of product marketing compare to the traditional artificial approach.**

***Index Terms*—Classification algorithms, Ensemble Learning, Random Forest, Big Data, Spark.**

This work was supported in part by the National Natural Science Foundation of China under Grant 61402183, in part by the National Science and Technology Ministry under Grant 2015BAK36B06, in part by the Guangdong Provincial Scientific and Technological Projects under Grant 2017A010101008, Grant 2017A010101014, Grant 2016A010101007, Grant 2016B090918021, and Grant 2014B010117001, in part by the Guangzhou Science and Technology Projects under Grant 201607010048 and Grant 201604010040, in part by the Opening Project of Guangdong Province Key Laboratory of Big Data Analysis and Processing under Grant 2017004, and in part by the Fundamental Research Funds for the Central Universities, SCUT.

Weiwei Lin is with School of Computer Engineering and Science, South China University of Technology, Guangzhou, China and is with Guangdong Key Laboratory of Big Data Analysis and Processing, Guangzhou 510006, China(e-mail: linww@scut.edu.cn).

Ziming Wu is with School of Computer Engineering and Science, South China University of Technology, Guangzhou, China (e-mail: 354242964@qq.com).

Longxin Lin is with College of Information Science and Technology, Jinan University, Guangzhou, China (e-mail: tlinlx@jnu.edu.cn).

Angzhan Wen is with School of Computer Engineering and Science, South China University of Technology, Guangzhou, China (e-mail: 770133694@qq.com).

Jin Li is with School of Computer Science, Guangzhou University, Guangzhou, China (e-mail: jinli71@gmail.com).

Corresponding author: Weiwei Lin (linww@scut.edu.cn) Longxin Lin (tlinlx@jnu.edu.cn)

I. 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.

The main purpose of this paper 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 2 introduces the current research status of imbalanced data classification; Section 3 puts forward the classification model and intelligent recommendation algorithm based on random forest for insurance business data, and analyzes its efficiency; Section 4 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.

II. RELATED WORK

The classification problems for China Life insurance business dataset is based on large data imbalance data classification. In data sets that positive and negative proportion are extremely imbalanced, the model predictions generated directly by logistic regression or SVM are biased in favor of the large proportion. When the positive and negative ratio is 100:1, due to the lack of sufficient support for negative example, classification models all predict positive cases can make the model recall reach 1.0 and precision is greater than 0.99. Therefore, it can be seen that such a model does not have any help in practical applications, such as in earthquake prediction, computer viruses and many other application fields.

Usually, the classification problem for unbalanced data sets has two pretreatment methods. (1) Over sampling, this generates data of the minority class to balance the proportion of data through some specific algorithms; (2) Under sampling, which reduces the proportion of the majority class through some sampling algorithm, so as to balance the number of positive and negative cases of training set [7].

Over sampling method is mainly divided into two types, (1) a non-heuristic sampling method which increases minority class samples by random replication, is easy to cause the over fitting of decision boundary; (2) a heuristic sampling method, which is represented by SMOTE algorithm [1], balances the category distribution of original data sets by adding some virtual samples. In recent years, many improved algorithms are proposed base on SMOTE, such as SMOTE-RSB [2] algorithm combined with the theory of RST, which filter the samples from the final sampling result when their similarity is greater than the given threshold; SMOTE-IPF [3] algorithm uses multi noise filter to resample synthetic sample data; SMOTE-FRST[4] algorithm is also combined with the theory of RST, which remove the synthetic samples that are less than the distance threshold; Borderline-SMOTE [10] focuses on the samples of minority class in the decision boundary when the law in the sample of a few samples on the boundary of decision during the sampling.

Sampling method is also divided into non-heuristic method and heuristic method. (1) Non-heuristic method randomly remove the samples of majority class, in order to reduce the degree of imbalance, but this method will usually remove some key samples, which results in under fitting on the boundary of decision because of the lack of key characteristics; (2) The heuristic sampling method usually distinguishes samples based on the recent neighbor algorithm, which divides the samples into safe points, dangerous points and noise points [5]. Representative algorithms include Tomek links [6], compressed nearest neighbor algorithm CNN [8], nearest neighbor removal algorithm NCL [9], and so on. In addition, some research scholars also proposed some novel algorithms such as boosting, bagging and other combination algorithm [11], [12], reverse random under sampling [7] and so on.

In recent years, with the rapid growth in the amount of data, the traditional algorithm cannot deal with the challenges of mass data. In order to cope with the challenges of big data, data mining based on big data technology has become a hot research, some scholars have proposed the classification algorithms for imbalanced data sets based on MapReduce, such as random forest algorithm based on MapReduce [13], [14], which achieves SMOTE over sampling through a MapReduce process, and then trains decision tree for each training data sets in parallel. Random forest belongs to the combination algorithm, having the advantages of automatic balance of error and automatic selection of features. Besides, the algorithm itself is easy to parallelize, so it has very excellent performance in dealing with large-scale imbalanced data classification.

Data mining application based on big data technology cannot get away from the support of computing platform. Apache Hadoop MapReduce is one of the most popular tools for big data computation, but its efficiency is very low when dealing with iterative calculation. The main reason is that it needs to store the intermediate results of each MapReduce to the disk. It is obviously insufficient for the efficiency requirement of real-time query. In order to solve the problem of query delay, the new generation of big data processing tools began to try the pipeline scheduling system and In-Memory architecture. The company Cloudera came up with a real-time query framework named Impala, and put forward the data processing methods of MapReduce combined with Impala. The framework localize the tasks through its scheduling algorithm, which saves Shuffle I/O overhead and improves the iteration efficiency nearly 8 times [19]; Apache Spark is a new generation of big data computing platform based on memory, which can directly cache the intermediate result in the memory area so as to eliminate the process of persistence. The efficiency of Spark is known for its computation efficiency and scheduling model based on DAG. Compared to Hadoop which only provides the MapReduce programming model, Spark provides a variety of RDD programming paradigm. Now, with the memory becomes more and more cheap, more and more computing was moved to Spark, such as machine learning framework Spark MLlib and

graph computing framework GraphX. Because the computation and storage of Spark are kept in memory, so the efficiency of algorithm running on Spark is 10 to 100 times upgrading compared to Hadoop. In the following chapters, we will introduce the implementation of the integrated random forest algorithm based on Spark and its specific application in the data classification of insurance business.

III. MODLING AND ALGORITHM

A. Classification Model

The traditional sales approach of insurance company is promoting the insurance products through the agent. Each insurance salesman develops new customers and maintains old customers on its own strength, and the salesmen choose potential users to sell new products in accordance with the sales experience. The traditional distribution method has many disadvantages. (1) Low efficiency, each salesman contacts the potential customers using the telephone or home visiting one by one; (2) Choosing potential users empirically or using artificial feature selection method; (3) Lack of customer evaluation system, don't know the characteristics influence weight of the potential customers; (4) The data accumulated in this way usually has serious ruinous, indirect influence the accuracy of classification model.

For a lot of classification models, the distribution balance and correlation of features directly affect the forecast results. Due to the imbalance distribution of the insurance business data feature and the independence between each other, directly using SVM or other strong classifier algorithms such as decision tree etc. will make a serious deviation result. So in handling such kind of problem, we prefer to use boosting algorithms or ensemble algorithms.

The primary problem of imbalance feature dataset classification is data resampling, there are two main types of paradigm through the previous research which called over-sampling and under-sampling. In this article, we used a heuristic under-sampling method which using the bootstrap under-sampling with replace for several times on the original dataset, for each sampling batch we used k-nearest neighbor algorithm to preprocess the candidate data sample set, and then calculated the classification result of each sampling batch. We draw lessons from the method of ensemble learning in the sampling process, used homomorphism integrated learning method to categorize the user feature dataset, unified with the random forest algorithm as the classification kernel model for each data batch. At last, we calculate the classification result for each batch, and integrate the results of each classifier by following model (1). P represents the product purchase probability, w on behalf of the training parameters, x for the purchase characteristics, Φ represents the decision function, N for total voting times.

$$P(y = 1 \mid x, w) = \frac{\sum_{i=1}^{s}\sum_{j=1}^{c}\Phi(y = 1 \mid x_{i,j}, w_{i,j})}{N} \quad (1)$$

B. Ensemble Random Forest Algorithm

The Random Forest is made up of several decision trees, each decision tree will be full growth, it do not need to cut processing, the more tree it has the more accurate the result will be, and it will not over fitting. The random forest algorithm will do the overall estimate, and it has the advantage of automatic feature selection etc. So we have the following main problems to be solved.

(1) Design a bootstrap sampling with replace algorithm for minor class, and then find the k nearest major class neighbor of the minority class samples, this could be parallel processed by taking the advantage of Spark.

(2) Build the random forest classifier for each sampling batch, each classifier only process the training dataset that dispatched on the same node, then using the model to predict the testing dataset in parallel, at last collect the prediction results by Spark Driver.

Algorithm 1 Ensemble Random Forest Algorithm

Input: Training data T, parameters $\{\lambda, \delta, k, s, c\}$
Output: Model with evaluation

1. Ensemble-RF$(T, \lambda, \delta, k, s, c)$
2. **for** $i < -1$ *to* s **do**
3. $(train, test) \leftarrow$ randomSplit(T, λ)
4. $split \leftarrow$ bootstrap$(train, \delta, k)$
5. $model \leftarrow RandomForest.$train$(split, c)$
6. $score \leftarrow$ evaluate$(model, test)$
7. out$[i] \leftarrow (model, score)$
8. **end for**
9. **return** *out*

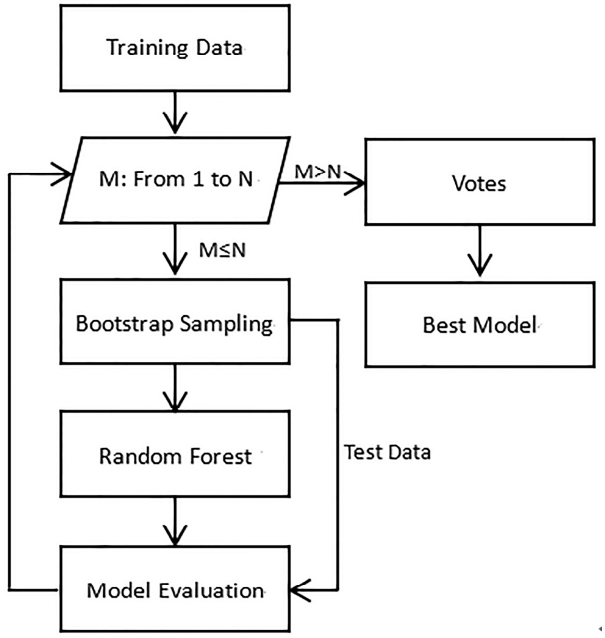


Fig. 1. Ensemble Random Forest algorithm.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2738069, IEEE Access

To solve the above problems we need to design a global partition algorithm, put the data items which have the same key into same partition, and put the partitions that have to rely on into the same node. The ensemble random forest algorithm has to sampling the origin dataset several times, and building random forest is a processing of iteration to promoting for user feature information gain, so we choose to implement the whole algorithm over Apache Spark platform, take the advantage of the Spark's efficient parallel computing to accelerate the calculation of the algorithm. The design procedure of the algorithm is shown in Fig. 1, the pseudo code as shown in Algorithm 1, symbol *s* represents partition number, *c* represents the decision tree number of the random forest, *k* is the ratio of positive and negative cases, λ is the sampling threshold, δ represents the search radius of nearest neighbor, the *bootstrap* process according to the *sampling* process, and the *evaluate* process represents the accuracy of the classification model.

Algorithm 2 Bootstrap Sampling based on KNN

Input: Dataset T, parameters {δ, k}
Output: Sample Batch
1. **bootstrap**(T,δ,k)
2. $minor \leftarrow broadcast(T.minor)$
3. $dist \leftarrow T.major.\text{map}(l \Rightarrow \text{dist}(l, minor))$
4. $candidate \leftarrow dist.\text{filter}(l \Rightarrow l.distance < \delta)$
5. $batch \leftarrow \text{sample}(candidate, k)$
6. **return** *batch*

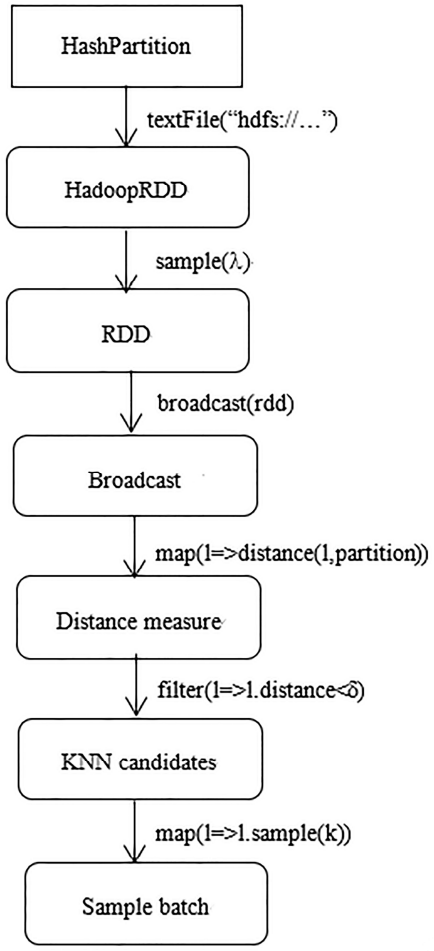


Fig. 2. Bootstrap sampling based on k-NN.

The bootstrap sampling method based on k nearest neighbor algorithm put the minor class samples into a broadcast variable, which will be cached in every node by Spark. We take $\lambda \in (0,1)$ as the threshold of sampling ratio, $\delta \in (0,\infty)$ as the search radius of k nearest neighbor algorithm, then we find the k major samples in every partition for each minor sample, put the rest samples into the test dataset, the design of sampling algorithm is shown as Fig. 2, the pseudo code is shown in Algorithm 2.

Spark MLlib provides a variety of ensemble learning algorithms including the random forest and boosting trees etc., most of the algorithms have been optimized for the DAG computing model of Spark, we should combine the sampling method and ensemble algorithms or other weak learning methods to implement the classification model of the imbalance distribution feature dataset. So we modified the random forest algorithms implemented in Spark MLlib, and put our bootstrap sampling method into the classification model.

C. Algorithm Improvements

Ensemble random forest algorithm based on Spark can optimize the imbalance classification problems from two aspects.

(1) Over-sampling algorithms based on SMOTE can generate a lot of virtual samples, which would increase the computing load, and the SMOTE based algorithms is suitable for the balanced classification algorithms like SVM, logistic regression and other strong classifier, but it's useless for ensemble learning algorithms like random forest etc. We used the heuristic under-sampling algorithm like bootstrap sampling method based on KNN rather than random sampling, which will find a better sample subset by searching for security samples near the classification boundary.

(2) Ensemble random forest algorithm based on the Spark platform uses the hash partition algorithm to divide the original dataset into several disjoint subsets, and uses broadcast variable to cache the minor class samples into each node, which could be more efficiency than MapReduce approach. The RDDs in the process of build random forest will be cached automatically by DAG Scheduler which could be rapidly reused in the following calculation.

D. Model Evaluation

Evaluation for classification algorithms mainly includes the mixed matrix analysis of predicted results. The main evaluation indicators for ensemble random forest includes:

TABLE I
HYBRID MATRIX

	predicted positive	predicted negative
condition positive	TP	FN
condition negative	FP	TN

Precision-Recall, F-measure or G-mean etc. In this paper, we compare the ensemble random forest algorithm with other classification algorithms from three aspects: the running time, F1 measure and G mean as well.

F measure contains two sub indicators; classification precision is the fraction of predicted instances that are relevant, while recall is the fraction of relevant instances that are predicted. Precision and recall are not necessarily but mutually restricted in large scaled data set processing. Normally, when you need to get higher recall, the model should predict more positive samples, and the precision may be associated with decline.

$$recall = \frac{TP}{TP + FN} \tag{2}$$

$$precision = \frac{TP}{TP + FP} \tag{3}$$

The F_b both take into account the recall and precision, the coefficient of b on behalf of the weight of recall, usually we use F1 to denote the accuracy of classification model. G-mean is the geometric mean of precision and recall. We use both F1 and G-mean as the evaluation predictors of ensemble random forest algorithm in the following experiment, and encapsulated them into the evaluate function of Algorithm 1.

$$F_b = \frac{(1+b^2) \times recall \times precision}{b^2 \times recall + precision} \tag{4}$$

$$G\text{-}mean = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}} \tag{5}$$

E. Algorithm Analysis

The computational complexity of the ensemble random forest algorithm could be estimated as follows: let the *s* represents sampling batches, *c* as the average decision tree node number of random forests, *m* on behalf of the feature number of each decision tree, if the average length of sampling bathes is *n*, then the computational complexity of ensemble random forest is going up to $O(s \times c \times m \times n \times \log(n))$ which is an estimation of non-parallel algorithms.

The computational complexity of serial ensemble random forest algorithm is unpredictable in the situation of large-scale business data, but as the previous introduced, the parallel implement of ensemble random forest based on Spark is considerable efficiency, the parallel algorithm used Spark RDD as the storage of the sampling batches, which means all data would be cached into memory and involved in the calculation directly. We broadcast the minor class samples to reduce the communication overhead, and build the random forest in parallel to accelerate the modeling. Let *k* as the number of partitions, the length of each sampling batch should be pruned to $O(s \times c \times m / \mathrm{k})$, and then the complexity of the Spark based ensemble random forest algorithm would be $O(s \times c \times m \times n \times \log(n) / \mathrm{k}^2)$. The actual operation efficiency will be higher by the help of DAG Scheduler.

IV. EXPERIMENTS AND RESULTS

Our experiment was performed on a five-node Linux cluster. Each node has two 1.35GHz Intel core CPU, 8 GB memory, with Centos 6.5 operation system installed and Apache Spark 1.5 deployed. We extracted more than 500,000 customer purchase behavior data from China Life Insurance Company over the past three years. The positive cases are 20787 only accounted for 4.1% of the total data, and the features of the data is highly out of balance. We choose 16 available user features by computing the information gain, and preprocessing the business data by setting segmentation points for discrete values or setting the threshold value for sequence values.

We separately used the SMOTE based algorithms and ensemble random forest algorithm to analysis the business data, then we choose traditional SVM, Logistic Regression and Random Forest algorithms to compare with the ensemble random forest algorithm. Their running times and evaluated results are shown in table II, which shows that the ensemble random forest algorithm has higher operation efficiency than most of the strong classification algorithms, and it is suitable for imbalanced classification model. And the table II also shows that, strong classifiers failed to find the decision boundary due with the imbalance distribution of user features, the SMOTE sampling preprocessing is useful but it will spend a longer time to build the correct model on the large number of virtual samples, the ensemble random forest take fewer running time to get more accurate results because the sampling batch is parallel processed by executors, and it's also performed well than MapReduce approaches due to the memory cache mechanism of Spark.

TABLE II
PERFORMANCE COMPARISON OF ENSEMBLE RANDOM FOREST WITH OTHER ALGORITHMS

Algorithm	Time(s)	F1	G-mean
LR	58.7	0.22	0.24
RF	41.6	0.28	0.24
SVM	66	0.18	0.21
LR-SMOTE	102.1	0.25	0.28
RF-SMOTE	67.9	0.28	0.26
SVM-SMOTE	140.5	0.22	0.25
ERF(6seg)	92.1	0.36	0.44

Algorithms like SVM and LR etc. is useless in the classification of imbalance distribution feature dataset, as the Fig. 3 showing below, we tested the ensemble random forest algorithm by difference number of features, and most of the features has a serious imbalance distribution, the ERF algorithm has a better performance when the feature number reaches 16, and it also has a better performance than other algorithms when the number of features in a reasonable range.

We also tested the impact of tree numbers on random forest algorithms, experiment shows that the algorithm could be stable when the tree number reaches 25 or more, Fig. 4 also shows that the bootstrap sampling algorithm has a good performance for imbalance classification algorithms.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2738069, IEEE Access

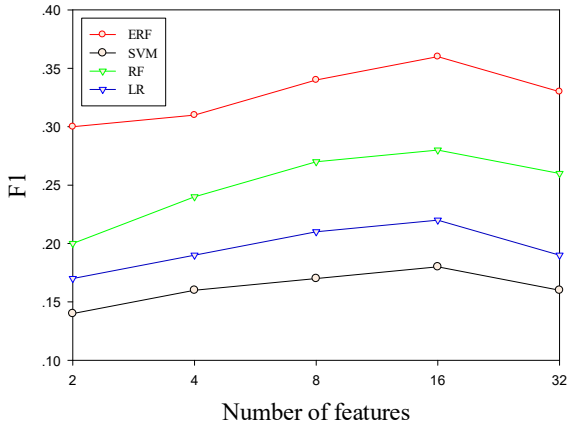


Fig. 3. Impact of feature number on ERF and other algorithms.

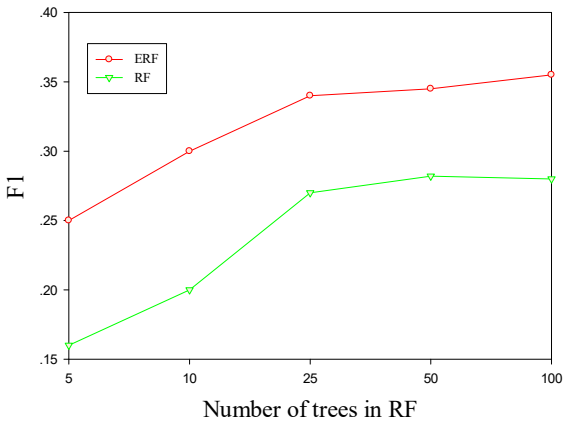


Fig. 4. Impact of tree number on ERF and RF.

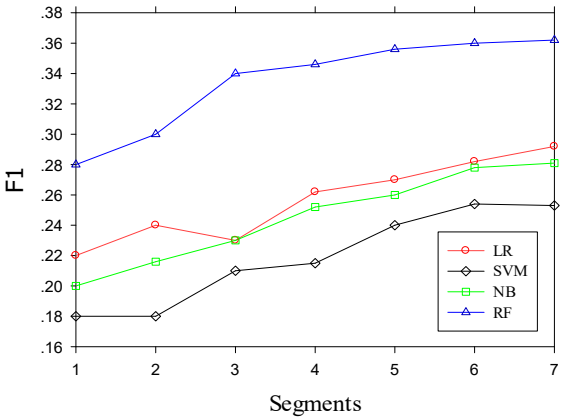


Fig. 5. Effect of bootstrap sampling algorithm to other classification model.

To prove the efficiency of the bootstrap sampling algorithm we changed the classification algorithm to the strong classification algorithms in our model, and compared with random forest algorithm, the result is shown in Fig. 5, which shows that the random forest is the most suitable algorithm of our model, and that also proved the bootstrap sampling method is useful in imbalance classification problems.

Finally we used the ensemble random forest algorithm to analysis the customer feature data derived from China Life Insurance Company, the experiment was performed to choose the potential user by ensemble random forest which compared with the traditional artificial approach, then compared with the real sales data, the error analysis was shown in Fig. 6 which shows that the recall of traditional approach is only 4%, and it failed to distinguish the purchase ability of the potential users; the ensemble random forest algorithm predicted 60831 potential customers and 24% of them buy the insurance product indeed, the recall reached up to 30.9% within the range of prediction interval between 60% and 90%, the error analysis shows that the ensemble random forest could find the purchase feature of potential users effectively and improve the accuracy of product marketing.

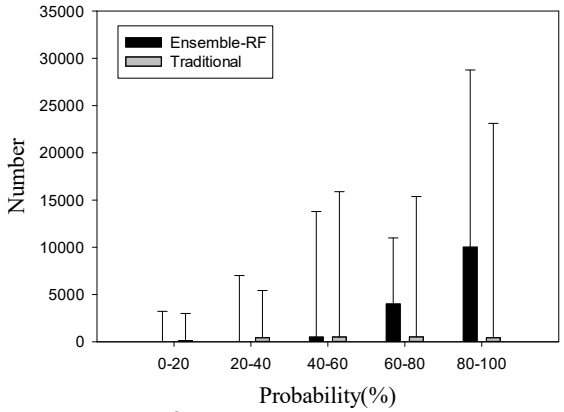


Fig. 6. Proportion of predicted potential users.

V. CONCLUSIONS AND FUTURE WORK

This paper 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 data mining algorithms. 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. Our further work includes: (i) exploring the proposed algorithm to different types of big data analytics; (ii) combining deep learning [20] into the proposed algorithm to improve the accuracy of prediction based on big data analysis.

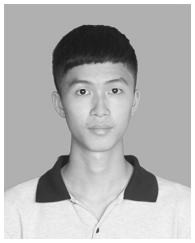
REFERENCES

[1] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique [J]. Journal of artificial intelligence research, 2002: 321-357.

[2] Ramentol E, Caballero Y, Bello R, et al. SMOTE-RSB*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory [J]. Knowledge and information systems, 2012, 33(2): 245-265.

[3] Sáez J A, Luengo J, Stefanowski J, et al. SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering[J]. Information Sciences, 2015, 291: 184-203.
[4] Ramentol E, Verbiest N, Bello R, et al. SMOTE-FRST: a new resampling method using fuzzy rough set theory[C]//10th International FLINS conference on uncertainty modelling in knowledge engineering and decision making (to appear). 2012.
[5] GU Ping, OU YANG Yuan-you. Classification research for unbalanced data based on mixed-sampling[J]. Application Research of Computers, 2015, 32(2): 379-381.
[6] Tomek I. Two modifications of CNN[J]. IEEE Trans. Syst. Man Cybern. 1976, 6: 769-772.
[7] Tahir M A, Kittler J, Yan F. Inverse random under sampling for class imbalance problem and its application to multi-label classification[J]. Pattern Recognition, 2012, 45(10): 3738-3750.
[8] Angiulli F. Fast condensed nearest neighbor rule[C]//Proceedings of the 22nd international conference on Machine learning. ACM, 2005: 25-32.
[9] Laurikkala J. Improving identification of difficult small classes by balancing class distribution[M]. Springer Berlin Heidelberg, 2001.
[10] Han H, Wang W Y, Mao B H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning[M]//Advances in intelligent computing. Springer Berlin Heidelberg, 2005: 878-887.
[11] Friedman J H, Hall P. On bagging and nonlinear estimation [J]. Journal of statistical planning and inference, 2007, 137(3): 669-683.
[12] Hido S, Kashima H, Takahashi Y. Roughly balanced bagging for imbalanced data [J]. Statistical Analysis and Data Mining, 2009, 2(5-6): 412-426.
[13] Del Río S, López V, Benítez J M, et al. On the use of MapReduce for imbalanced big data using random forest [J]. Information Sciences, 2014, 285: 112-137.
[14] Bhagat R C, Patil S S. Enhanced SMOTE algorithm for classification of imbalanced big-data using Random Forest[C]//Advance Computing Conference (IACC), 2015 IEEE International. IEEE, 2015: 403-408.
[15] Liaw A, Wiener M. Classification and regression by randomForest[J]. R news, 2002, 2(3): 18-22.
[16] Goldston D. Big data: Data wrangling [J]. Nature, 2008, 455(15).
[17] Reichman O J, Jones M B, Schildhauer M P. Challenges and opportunities of open data in ecology [J]. Science, 2011, 331(6018).
[18] TU Xin-li, LIU Bo, LIN Wei-wei. Survey of big data[J]. Application Research of Computers, 2014, 31(6): 161.
[19] Beibei Li, Bo Liu, Weiwei Lin, and Ying Zhang. Performance Analysis of Clustering Algorithm under Two kinds of Big Data Architecture. Journal of High Speed Networks, 2017, 23(1): 49-57.
[20] Ping Li, Jin Li, Zhengan Huang, Tong Li, Chong-Zhi Gao, Siu-Ming Yiu, Kai Chen. Multi-key privacy-preserving deep learning in cloud computing. Future Generation Computer Systems, 2017. DOI: 10.1016/j.future.2017.02.006.

Weiwei Lin received his Ph.D. degree in Computer Application from South China University of Technology, China, in 2007. He is currently an Associate Professor with the Department of Computer Science and Engineering, South China University of Technology. He has published more than 60 papers in refereed journals and conference proceedings. His research interests include distributed system, cloud computing and big data.



Ziming Wu received the B.Eng. degree in computer science and technology from South China University of Technology, China, in 2017, where he is currently pursuing the M.S. degree with the Department of Computer Science and Engineering. His research interests include cloud computing and big data.

Longxin Lin received his Ph.D. degree in Computer Application from South China University of Technology, China, in 2008. He is currently an Associate Professor with the Department of Information Science and Technology, Jinan University. His research interests include distributed system and video analysis.

Angzhan Wen received his B.S. degree in computer science and technology from South China Agricultural University, in 2005. He is currently pursuing the M.S. degree with the Department of Computer Science and Engineering, South China University of Technology. His current research focuses on cloud computing and big data.

Jin Li received his B.S. in Mathematics from Southwest University, China, in 2002 and the Ph.D. degree in Information Security from Sun Yat-Sen University, China, in 2007. He is currently a Professor in Guangzhou University. He has published over 60 research papers in refereed international conferences and journals and has served as the program chair or program committee member in many international conferences.