Empirical analysis of ensemble methods for the classification of robocalls in telecommunications

Meghna Ghosh, Prabu P
Department of Computer Science, Christ (Deemed to be University), India

Article Info

Article history:

Received Oct 25, 2018
Revised Apr 1, 2019
Accepted Apr 8, 2019

Keywords:

Ensemble method
Machine Learning
Random Forest
Robocalls
XGBoost

ABSTRACT

With the advent of technology, there has been an excessive use of cellular phones. Cellular phones have made life convenient in our society. However, individuals and groups have subverted the telecommunication devices to deceive unwary victims. Robocalls are quite prevalent these days and they can either be legal or used by scammers to trick one out of their money. The proposed methodology in the paper is to experiment two ensemble models on the dataset acquired from the Federal Trade Commission (DNC Dataset). It is imperative to analyze the call records and based on the patterns the calls can classify as a robocall or not a robocall. Two algorithms Random Forest and XgBoost are combined in two ways and compared in the paper in terms of accuracy, sensitivity and the time taken.

Corresponding Author:

Meghna Ghosh,
Department of Computer Science,
Christ (Deemed to be University),
Hosur Main Road, Bangalore-560029, India.
Email: meghna.ghosh@cs.christuniversity.in, prabu.p@christuniversity.in

1. INTRODUCTION

The Federal Trade Commission received over 22 million complaints of illegal and unwanted calls, in 2014. Telephone spammers today are leveraging recent technical advances in the telephony ecosystem to distribute massively automated spam calls known as robocalls [1]. A phone call that uses a computerized autodialer to deliver a pre-recorded message at the other end, as if it were from a robot is a robocall. Once viewed as an inconvenience they have reached epidemic proportions. Few robocalls are also considered legal. The calls permitted can be campaigning for candidates, alerting students to campus closures, appointment reminders, flight cancellation etc. An illegal robocall is a non-emergency call containing a pre-recorded message without the consent of the consumer. It can be either from a registered business which contravened the law or from a scammer that pose as a legal organization in order to steal your money, identity or both. Technology has made it easy to find ways scrape personal information on public databases or internet to find the phone numbers and sell them to both legal and illegal spam callers.

In Canada, during the Canadian Federal 2011, in order to reach voters, the political parties legitimately used robocalls. The investigation showed that the robocalls were used to divert the people from casting their ballot by giving them inaccurate information of the changed locations of the poll stations. There has been a steep rise in the automated calls since 2009. According to the FTC report, an agency received over 375,000 complaints about automated robocalls as compared to 2009. The report also stated that the increase in the number of robocalls is due to the free or cheap access to internet calling services which also helps the scammers hide their identity.

Machine Learning is an application of artificial intelligence that provides the system the potential to grasp patterns and learn from data and ameliorate from experience depending on some task, without being explicitly coded. Machine Learning mainly focuses on learning from input data and predicting an outcome along with updating the results with the arrival of new, unseen data. Efficient analysis of fraud in telecommunication with the help of machine learning can help network operators save a money restore the consumer’s confidence in their security. The analysis in machine learning has a lot of steps, the first one being identifying the objective of the task, depending on which the model is chosen. Collection of data is the next important step followed by the pre-processing of data which is the transformations applied to the data. Before feeding the pre-processed data to a particular data it is required to split the data into training and testing data. Few strategies for improving the performance of the data could be supplement with additional data or switch to a different model.

In machine learning, ensemble models use multiple learning algorithms and combine them in order to build an optimal predictive model. Rather than just relying on one algorithm such as a Decision Tree and expecting that the right decision is made at each split, ensemble method suggests assembling a sample of decision trees and make a final predictor based on the aggregation of the results. The paper defines the ensemble methods, bagging, boosting and stacking which is described in detail below.

a. Bagging

Bagging is an ensemble method where the model is trained on a number of bootstrapped samples and the final model is the aggregated result of the sample models. Bootstrapping is the process of resampling with replacement where small samples are repeatedly chosen from the original sample. A few duplicate samples exist in a bootstrapped sample as sampling is done with replacement. In the case of a numeric target variable or regression problem the predicted outcome is the average of the models whereas in case of classification the predicted class is defined based on plurality. Bagging or Bootstrap Aggregation generally reduces variance for those algorithms having high variance. The Bagging model is shown in Figure 1. Random Forest is one such algorithm that is an improvement over bagged decision trees. Random Forest is the combination of multiple decision trees in which the sub-trees are learned in such a way so that the resulting predictions have less correlation. The algorithm also helps calculate the error function for a variable at each split. Bagging method is an efficient method when it comes to dealing with overfitting of data along with reducing variance.

b. Boosting

Boosting is an ensemble method in which a set of weak learners is converted to a set of strong learners [2]. A weak learner is a classifier that has very less or slight correlation to the true classification whereas the strong learner is well correlated to the true classification. The bagging model is depicted in Figure 2. In the training phase weights are allocated to each of the resulting models. A higher weight is assigned to the learner with superior classification results as compared to a learner with poor classification results. Boosting optimizes the advantages of a single model and generates a combined model with lower errors. Boosting helps reduce bias. Gradient boosting is one boosting algorithm that grasps the patterns in order to strengthen a model with weak predictions.

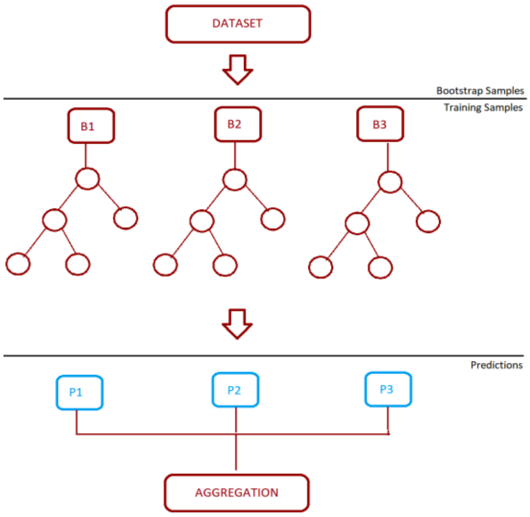

Figure 1. Bagging model

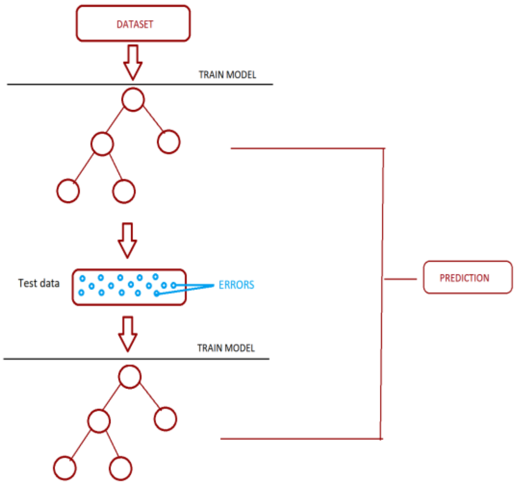

Figure 2. Boosting model

c. Stacking

Stacking is a way to ensemble multiple regression and classification problems which introduces the concept of a meta-learner. Stacking involves training a set of base learners on the data. The predictions obtained from the base learners are taken as an input for the meta-classifier which gives the final predicted output. Unlike bagging and boosting, stacking is not so widely used. It helps in improving the overall performance and we often end up with a model that is better than an individual intermediate model.

d. Related work

Qayyum et al. [3] in the paper, has focused on detecting fraudulent behavior in the field of subscription fraud. Time series data is taken as input on which neural network is implemented. The paper proposed a multilayer perceptron which is based on an input, hidden and output layer. It is used to associate a fraud rating to a subscriber based on his/her calling profile. Two techniques are developed, one in which the users are allowed to choose their own fraud features which they think best fraud they are dealing with, another in which weights priorities are associated based on the user's input to give more priority to one fraud compared to others. A single threshold value is used by the current neural network and the weights are adjusted based on the error in prediction. However, for better classification to output neutron could be used. One output neuron can be assigned for detecting fraudulent behavior and the other for non-fraudulent behavior.

Hilas et al. [4] have presented two clustering algorithms to identify user behavior profiles to detect fraudulent activities in a telecommunication organization. Unsupervised learning deals with unlabelled data and helps find groupings in the data. The partitioning and hierarchical clustering are used in the present work. As the main representative the k-means algorithm is applied and the hierarchical clustering is used in the agglomerative clustering. In the paper, the well established unsupervised learning techniques are applied on telecommunications data. The techniques help comprehend on a fraudulent behaviour from a legitimate user's behavior. Raw usage data must be transformed into appropriate user profiles. Some of the challenges faced were construction and selection. It is concluded from the analysis that, as regards user profile building, accumulated characteristics of a user yield better discrimination results. However, in order to preserve online detection ability aggregating user's behavior was avoided for larger periods. Misclassification in clustering occurred due to mixed types of behavior.

Phua et al. [5] the paper surveys the various technical articles in automated fraud detection for a period of 10 years. It defines and formalizes the types and subtypes of fraud. This research paper presents the techniques along with their problems. The main objective is to define current challenges in this domain for the different large data streams. It compares and summarizes the various data mining based fraud detection techniques. The different data mining techniques are selected depending on the practical issues of operational requirements, resource constraints. Graph-theoretic anomaly detection and Inductive Logic Programming are some of the commercial fraud detection techniques. Non-linear supervised algorithms which are complex, such as, support vector machines and neural networks is given more importance. In the long term, less complex algorithms such as naïve Bayes and logistic regression will produce faster results if not better. Other related data mining techniques covered by survey papers include outlier detection, skewed/imbalanced/rare classes, sampling, cost-sensitive learning, stream mining, graph mining, and scalability. From this paper, it is concluded that unsupervised approaches that can contribute to future fraud detection research include actual monitoring systems and text mining from law enforcement and semi-supervised and game-theoretic approaches from intrusion and spam detection.

Xu et al. [6] worked on the detection of fraud in the 3G telecommunication network. A rough fuzzy set based approach was used. The 3G network is analyzed including the subscription and superimposed fraud. The profiles and various parameters are defined in order to present the framework. Citi FMS, a rule-based system, was developed for the detection of abnormalities and alarm.

Farvaresh, H. et al. [7] aims at identifying the subscription fraud by analyzing the user profiles in the paper. A variety of hybrid algorithms are applied to the dataset acquired from the Telecommunication Company of Tehran. In clustering SOM and K means is combined whereas in classification SVM, decision trees, neural networks, bagging, boosting and various other ensembles were applied. The results revealed that SVM among single classifiers and boosting ensemble had higher accuracy as compared to the other algorithms.

In the paper, the author discusses the role of neural networks for pattern recognition in the prevention of telecommunication fraud. Akhter et al. [8] has collected data on fraudulent and non-fraudulent calls which are preprocessed for suitable neural network learning. A model is built from the preprocessed data which incorporates various patterns of fraudulent behavior. The combination of neural, rule-based, case-based technologies provide a fraud detection rate superior to that of conventional systems and the multi-stream analysis capability makes it extremely accurate. Due to the inherent ability to adapt along with the speed and efficiency, Artificial Neural Network is a better method for detecting telephone fraud.

Subscription Fraud, Call Forwarding, Calling Bypass, Roaming Fraud, and Cloning Fraud is the different types of fraud in telecommunication. Adebisi et al. [9] developed a model that detects telecommunication fraud based on a neural network ensemble method. A random rough subspace based neural network ensemble method was employed in the development of the model. The model was designed to detect subscription fraud. It presents the development of patterns that portrays the customer's behavior focusing on the identification of non-payment events. Rules were formed based on the information interrelated with other features. This lead to faster predictions to prevent revenue loss for the company. The results showed that neural network classifier 1 gave 7 wrong classifications, the second classifier gave 12 wrong classifications, the third classifier gave 1 wrong classification and the fourth classifier gave 7 wrong classifications. The neural network ensemble outperformed further enhancing the efficiency of the proposed model.

Cox et al. [10] in the paper domain-specific interfaces are built for telephone fraud detection. Human recognition skills exceed automated mining algorithms. Exploiting people's ability to deal with visual representations we may revolutionize the way we understand a large amount of data. Different views of the same data can be interlinked. The visualization approach to detecting calling fraud involves a display of calling activity that displays the unusual patterns and with the help of one or more drill-down views, suspicious patterns may be further investigated. Pattern recognition is performed on the AMA call records. The scatterplot and barplot show the various anomalies. The advantages of visual data mining lie in the fact that people excel at detecting patterns, the dynamic nature of fraud makes it a challenging detection problem for static algorithms. In this paper two representations, one for calling communities and the other for showing individual calls have proven to be effective to detect fraud in telecommunications. People complement machines and better exploit the capabilities for knowledge discovery.

Wu et al. [11] identifies the standard features of fraudulent behavior of customers in telecom industry systematically. The outliers in data are identified by the clustering techniques. The work in gives definition of target customers who are maliciously based on these specific methods are proposed to build, evaluate, and apply the model for identifying fraudulent behavior. Kohonen neural network and clustering algorithm are efficiently used for the detection of outliers.

2. RESEARCH METHOD

The study focused on four different parts which are, acquisition of data, data pre-processing, development of the model by combining bagging and boosting algorithms using voting classifier and development of a stacking model.

a. Data source

The dataset used for the experiment is acquired from the Federal Trade Commission, Do Not Call (DNC) Reported Calls Data. The dataset contains information about the robocalls that were reported to the Federal Trade Commission. The dataset includes information about the phone number originating the unwanted call, the date and time the call was made, the date and time the complaint was made, the consumer's city, state and area code and the subject of the calls. The dataset also contains a column of data stating whether the call is a robocall/recorded message or not.

b. Data pre processing

The given dataset is already congregated in the form of columns with each row having a unique identity. The first step in the preprocessing of data includes handling the missing values in the dataset. The missing values can be imputed using statistics (mean, median, mode) of each of the columns or by using a constant value. Feature Selection is another vital step in the pre-processing of data which includes choosing the subset of features that are relevant to the predictive modeling problem. A set of six features are selected for the experiment. Table 1 shows the list of descriptors used for the purpose of designing the classification model. The data acquired from the month of July is used as the training data. The model is trained on the training data. The data acquired from the month of August is used as the test data. The model is fit on the test data to get the desired outcomes.

Table 1. Feature descriptors

Field Name	Description	Data Type
Company_Phone_Number	Telephone number originating the call.	Int64
Violation_Date	The date and time the call was made.	Object
Consumer_State	The consumer's state locations.	Object
Consumer_City	The consumer's city locations.	Object
Consumer_Area_Code	The area code of the consumer's city.	Int64
Subject	The subject of the call.	Object

c. Random Forest

Random Forest is a supervised learning technique that essentially uses an ensemble of decorrelated decision trees. The basic idea of Random Forest is that they use bagging or bootstrap aggregating. It uses sampling with replacement to create multiple datasets and on each of the dataset, the decision tree algorithm is applied. Each tree comes up with a prediction which could be a classification or a regression. Random Forest consists of a collection of tree-structured classifiers $\{h(x, \Theta k)\ k=1, 2, \ldots.\}$ where Θk are iid(independent identically distributed random vectors and each tree casts a vote for the most favored class [12]. The results are finally aggregated.

On approach is random subspace method where at each split the algorithm is provided with a subset of features on which it can be split. It randomly selects the subsets that are given at each step. There is also the random split selection in which a data point is selected randomly on which the split can occur. During the tree building process using bootstrap samples a small part of the original instances are left out. The set of instances is called the OOB (Out Of Bag) data which is used for the error estimation in individual trees. Random Forest helps increase classification accuracy by averaging the noisy and unbiased models and building a model with low variance.

d. XGBoost

Gradient Boosting is a supervised learning algorithm that works well for both classification and regression problems. Gradient Boosting aims at building a predictive model. The predictive model is an ensemble of various weak learners such as decision trees. The goal is to minimize the error function for optimization purposes. Gradient descent is one way of minimizing the loss function by updating the predictions based on a learning rate. XGBoost system is available as an open source package [13]. XGBoost is a powerful, optimized gradient boosting library that provides parallel tree boosting. XGBoost is a robust library and is efficacious in model performance and execution speed.

e. Voting classifier and stacking classifier

Two methods are implemented on the data and the results procured from the experiments are compared in terms of accuracy. The first methodology involves applying the Random Forest and XGBoost algorithm on the data. The results are then integrated with the help of the Voting Classifier. Xueyi Wang entitled “A New Model for Measuring the Accuracies of Majority Voting Ensembles “a new model called COB (core, outlier, and boundary) which quantitatively measures the accuracies of majority voting ensembles for binary classification [14].The Voting Classifier creates an ensemble by combining two standalone models. The Voting Classifier combines the predictions from the bagging and boosting algorithm. It provides us with the average of the predictions. A requisite for an ensemble of classifiers is to be precise than the individual classifiers is if the independent classifiers are diverse and accurate [15]. The following results were obtained on the successful build of the ensemble which is shown in Table 2.

Table 2. Classification report

	Precision	F1-score	recall	support
N	0.34	0.03	0.05	7095
Y	0.70	0.98	0.82	16723
Avg/total	0.59	23818	0.59	23818

3. RESULTS AND ANALYSIS

The evaluations of the Voting Classifier model and Stacking Classifier models are compared. The interpretation of results that the Voting Classifier model is better than the Stacking Classifier in terms of accuracy, time and generalization error. There is no overfitting in the data. Comparing the true negative and true positive of the confusion matrix derived from both the models, the accuracy for Voting Classifier was calculated to be 70.35% as opposed to the Stacking Classifier which showed the accuracy of only 61.7%. The time taken by the Voting Classifier was also considerably low, 1.9888 secs as compared to Stacking Classifier which takes 11.885 secs. Therefore, the hybrid approach of combining the boosting and bagging algorithm using the aggravation technique outperformed the stacking method. Tables 3 and 4 show the confusion matrix for the proposed models. Comaprison of model accuracy as shown in Figure 3.

Table 3. Confusion matrix of voting classifier

		Predicted	
Observed	Positive	Positive	Negative
		209	6886
	Negative	413	16310

Table 4. Confusion matrix of stacking classifier

		Predicted	
Observed	Positive	Positive	Negative
		301	8574
	Negative	538	14385

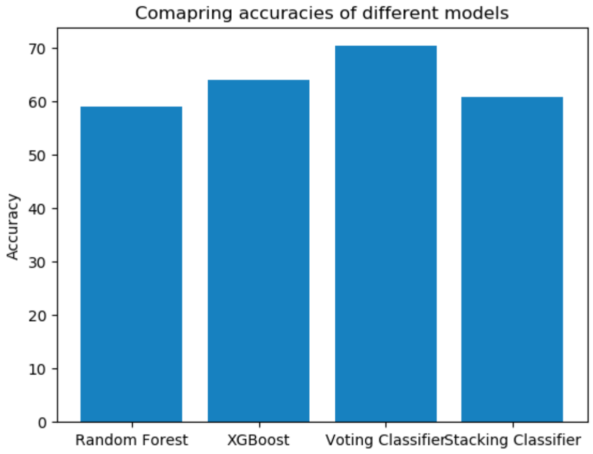

Figure 3. Comaprison of model accuracy

4. CONCLUSION

The main focus of the paper was to come up with a hybrid approach of combining the two defined ensembles, Bagging and Boosting. The potentialities of the learning models were explored. The conjunction of Bagging and Boosting and Boosting algorithms decrease the bias and variance from the dataset which leads to a more efficient model with less generalization error. The Voting Classifier classified the result more effectively with the model having higher true negative rates as compared to true positive. There are various other ensembles that could be combined into a hybrid model to check the proficiency of the model.

ACKNOWLEDGEMENTS

The author would like to thank guide and mentor, Prof. Prabu P, for his constant support throughout the creation of this research document. Other facilitators of the university have also been a most helpful ally in structuring the data and understanding the problem domain. The paper would not have been complete without the massive contributions by all the faculties and resources in leu of Christ (Deemed to be University), for both research and review. The documents presented and persistent in the paper were worked on a Macbook Air.

REFERENCES

[1] Huahong Tu, Adam Doupé, Ziming Zhao, Gail-Joon Ahn, "SoK: Everyone Hates Robocalls: A Survey of Techniques against Telephone Spam," *IEEE Symposium on Security and Privacy*, Arizona State University, pp.320-338, August 2016.

[2] Robert E. Schapire, "The Boosting Approach to Machine Learning An Overview," *Nonlinear Estimation and Classification, Springer*, pp. 1-23 2003.

[3] Sameer Qayyum, Shaheer Mansoor, Adeel Khalid, Khushbakht, Zahid Halim and A.Rauf Baig, "Fraudulent Call Detection For Mobile Networks," *International Conference on Information and Emerging Technologies*, Islamabad, Pakistan, pp. 1-5 2010.

[4] Constantinos S. Hilas1, Paris A. Mastorocostas, Ioannis T. Rekanos, "Clustering of Telecommunications User Profiles for Fraud Detection and Security Enhancement in Large Corporate Networks: A case Study," *Applied Mathematics & Information Sciences an International Journal*, vol. 9, no. 4, pp. 1709-1718, 2015.

[5] Clifton Phua, Vincent Lee, Kate Smith Ross Gayler, "A Comprehensive Survey of Data Mining based Fraud Detection Research," *School of Business Systems, Faculty of Information Technology,* Melbourne, Australia, pp. 1-14, March 2007.

[6] W.Xu, Y. Pang, J. Ma, S. Wang, G. Hao, S. Zeng, Y. Qain, "Fraud detection in telecommunication: a rough fuzzy set based approach," *International Conference of Machine Learning and Cybernetics*, vol. 3, pp. 1777-1787, July 2008.

[7] Farvaresh, H. and Sepehri, "A Data Mining Framework for Detecting Subscription Fraud in Telecommunication," *Engineering Applications of Artificial Intelligence, vol.* 24, no. 1, pp. 182–194, 2011.

[8] Mohammad Iquebal Akhter, Mohammad Gulam Ahamad, "Detecting Telecommunication Fraud using Neural Networks through Data Mining," *International Journal of Scientific & Engineering Research*, vol. 3, no. 3, pp. 1-5, March 2012.

[9] Fayemiwo Michael Adebisi and Olasoji Babatunde, “Fraud Detection in Mobile Telecommunications,” *International Journal of Innovative Research in Science Engineering and technology*, vol. 3, no. 4, pp. 11612-11620, April 2014.

[10] Kenneth C Cox, Stephen G, Graham J Wills, “Brief Application Description Visual Data Mining: Recognizing Telephone Calling Fraud,” Data Mining and Knowledge Discovery, vol 1, no. 2, pp. 225–231, June.

[11] S. Wu, N. Kang, L. Yang, “Fraudulent Behavior Forecast in Telecom Industry Based on Data Mining Technology,” *Communications of the IIMA*, vol. 7, no. 4, pp. 1-6, 2007.

[12] Vrushali Y Kulkarni, Pradeep K Sinha, “Effective Learning and Classification using Random Forest Algorithm,” *International Journal of Engineering and Innovative Technology*, vol. 3, no. 11, pp. 267-273, May 2014.

[13] Tianqi Chen, Carlos Guestrin, “XGBoost: A Scalable Tree Boosting System,” ACM SIGKDD, pp. 785-794, August 13 - 17, 2016.

[14] Xueyi Wang “A New Model for Measuring the Accuracies of majority voting ensembles,” *IEEE World Congress on Computational Intelligence*, 2012.

[15] Sarwesh Site, Dr. Sadhna K. Mishra, “A Review of Ensemble Technique for Improving Majority Voting for Classifier,” *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, no.1, pp. 177-180, January 2013.