DOI 10.29042/2018-4358-4363

Helix Vol. 8(6): 4358- 4363

A Survey on Machine Learning Techniques for Insurance Fraud Prediction

[*1]Komal S. Patil, [2]Prof.Anand Godbole

[*1]Department of Computer Engineering, Sardar Patel Institute of Technology, Mumbai, India

Email: pkomals94@gmail.com, anand_godbole@spit.ac.in

Received: 20th September 2018, Accepted: 11th October 2018, Published: 31st October 2018

Abstract

The fraudulent activities are increasing day by day with increase in technology in insurance sector. These fraud cases make shoddy impact on socio-economical system. This paper presents a detail survey of machine learning techniques used in insurance fraud prediction. This paper has disclosed traditional machine learning techniques like supervised and unsupervised learning and also some contemporary methods such as hybrid and ensemble leaning. The approach of the problem changes with the change in dataset hence this paper aims to provide an organized overview of the fraud prediction techniques based on the type of training data provided to the machine learning model.

Keywords

Insurance Fraud; Supervised Learning; Unsupervised Learning; Hybrid Classifiers; Ensemble Classifiers; Bagging; Boosting; Stacking.

Introduction

'Insurer' and 'Insured' are two pillars of insurance industry and the whole business runs due to the utmost faith within both of them. Insurance Fraud occurs when any action performed by either insurer or insured with an aspiration to gain some advantage to which they are not legally permitted or fraud may occur when any one of the party purposely refuses to provide benefits to other party which was legally permitted to them. The main intention behind initiating an insurance fraud is "to appear as conventional and to be proceed and get recompense in routine manner"

The survey of TOI reveals that one in every ten insurance claims is found to be fraud, which means around 10% of total insurance claims are fraud. According to the study of KPMG India Financial Services, insurance is the most vulnerable to fraud than other financial services, the survey says that loss caused due to insurance frauds are over Rs. 30,000crores(approximately $45billion) which is actually 9% of the total amount of insurance industry. The financial survey of Ernst & Young says that premium, claims and third party frauds are the three main fraud risks in insurance sector, from which only fraudulent claims contributes around 50% of the total fraud.

Nowadays almost every organization and agencies have their data stored in their databases; this data could be used for detecting and analyzing fraudulent activities. This hidden knowledge and patterns can be discovered using various machine learning techniques. Machine learning provides wide range of methods and algorithms to handle different types of problems depending on the need of an organization and type of the data hence it is one of the most popularly used technique for classification and prediction of fraud. Machine learning models are trained on the historic data and make predictions based on the prior knowledge extracted from the data, the model keep on updating with the new patterns and knowledge with incoming data. Machine learning is the most trending framework because of its high reliability and compatibility. A single classifier or multi classifier or hybrid model can be used according to need of the problem. Many researchers have used different techniques to deal with fraud cases. This paper gives the generalize overview of all the techniques used for an insurance fraud prediction and detection. The techniques are categorized based on the type of the data and also on the type of problem which is being solved by the model. Some hybrid and ensemble techniques are also disclosed in this paper as these techniques are gaining popularity because of their compatibility with other algorithms which are proved to be more efficient that single classifier models.

Materials and Methods

1. Traditional Machine Learning Methods

Traditionally the machine learning techniques are classified on basis of type of data which will be provided to the model. Therefore based on type of data there are three types of machine learning methods supervised learning, unsupervised learning and semi-supervised learning. Supervised and unsupervised methods are widely used by the researchers for fraud detection while few of them have also used semi-supervised method for fraud prediction.

1.1. Supervised Learning

Supervised learning techniques requires a labeled dataset these labels are nothing but the target variables target variables which discloses whether the particular claim is fraud or not. It means that to use supervised techniques, data should contain previously correctly identified claims based on this data algorithm generalizes the fraud instances and make predictions on new data instances.

For any supervised learning method let *X* be the set of *n* instances. These instances are also represented by a feature vector $x = (x_1, \ldots x_D)$ where D is a dimension of vector *x*. A training dataset is a collection of all such instances $\{x_i\}_{i=1}^{n} = \{x_1, \ldots x_n\}$ which will be given as input to the learning model. Let *Y* be the set of labels these labels are nothing but distinct values of different classes, $y \in \{1, \ldots C\}$ where *C* is number of classes. Now $P(X_i, Y_i)$ is given as probability of any instance *i* for particular class label, then supervised learning trains function *F* such that *f(x)* predicts whether Y_i is correctly labeled for given instance X_i.

However supervised learning method has few drawbacks. It is difficult to get labeled data always. Organization need to maintain the data labels from the beginning itself, in case if the data is unlabeled it is quite difficult and expensive to give labels to such huge data. Hence for such cases unsupervised learning methods are used. K- Nearest Neighbor (KNN), Naive Bayes, Decision Trees, Support Vector Machine (SVM), Neural Network, Regression are some popular supervised algorithms.

1.2. Unsupervised Learning

Unsupervised learning methods deals with the data where the target variable or the data label is not available. Unsupervised learning finds specific patterns within the data; this method of discovering the particular structures within the regularities of the data is called as density estimation. Clustering is commonly used method for density estimation. In clustering claims which poses similar characteristics are group together assuming that majority of instances are non fraudulent.

For the given training set $\{x_i\}_{i=1}^{n}$ the aim of supervised learning is to separate *n* instances into *k* clusters in such a way that instances within same clusters have same characteristics and instances of different clusters have different characteristics. The number of clusters can be predefined or algorithm itself partition the data into possible clusters based on characteristic of the data. The clusters formed are not necessarily discriminate, the clusters may be overlapped or may not be differentiate properly in such cases there is very thin or there may not be any boundary between the clusters at all. This is the limitation of unsupervised learning due to this the new claim may not be classified properly. Clustering, Association rules, Principal Component Analysis (PCA) are commonly used in unsupervised learning.

2. Hybrid Methods

Every individual learning method has its own benefits and drawbacks hence few researchers started using hybrid approaches for fraud detection. Hybrid learners are nothing but using two different learners together so that flaws of one learner could be overcome by another one. Hybrid methods are designed to perform specific tasks with combination of two or more algorithms these algorithms can be supervised, unsupervised or could be both, most of hybrid methods use combination of supervised and unsupervised methods. Vipula Rawte et al.[4] have used evolving clustering method to first detect the cluster of the disease and then applied SVM to detect whether the particular claim is legitimate or fraud. In most of the hybrid methods clustering is use to identify the position of an instance and then different classifiers are used to classify that particular instance into specific class.

3. Ensemble Learners

Ensemble is a framework of integrating various homogeneous or heterogeneous learners together so as to outperform the model than that of single classifier. The main idea behind constructing an ensemble is to improve the prediction performance. A typical ensemble learner contains the following elements:

1. Training Set: The training dataset for ensemble models need to be labeled always. A training data could be considered as attribute-value vector. The training set can be given as
   $X = \{X_1, X_2 \ldots X_n\}$ where *n* be the number of attributes in a training set and *Y* be the set of target variable.
2. Base Classifier: The base classifiers are nothing but the classification algorithms which are trained on the training set and make their predictions. Each base classifier performs independent of each other. Each base classifiers solves same problem and make individual predictions, predictions made by every classifier may or may not be same this property of an ensemble gives more generalize predictions.
3. Combiner: As the name suggests combiner combines the output from each base classifier and produces new prediction. This combiner could be a function or could be another classifier (meta-classifier).

There are some predefined ensemble models which are given below:

3.1 Bagging

Bagging is most commonly used independent ensemble model. Bagging is a technique which implements similar classifiers on small set of instances and then applies a mean or average of all the predictions. Generally bagging uses different learners on different population. In bagging the training dataset is divided into its subsets and then each training set is given to the individual classifiers. These base classifiers make their predictions and the combiner function makes final prediction by taking mean or average or voting of every base classifier. The base classifiers used in bagging may or may not be same. The use of diverse classifiers produces diverse output and hence the final prediction will be more generalize.

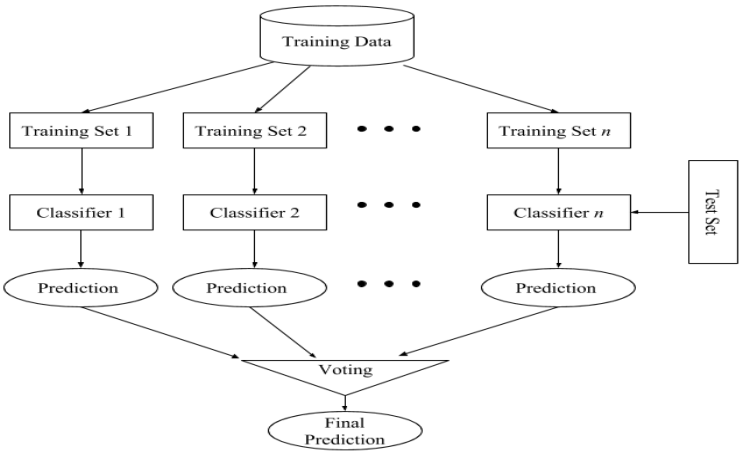

Figure 1: Bagging Model

Wagging is variant of bagging the only difference between bagging and wagging is in wagging all the base classifiers are trained on the entire training dataset and each base classifier carry particular weight for its prediction. With this approach weights can be assigned to the classifiers which give more promising and accurate predictions than other base classifiers. In final prediction, predictions made by the classifiers with high weight have more influence on final prediction. Random forest is another variant of bagging. As the name suggests random forest is the forest of decision trees, these decision trees are built as base classifiers and a final decision tree is formed over all these trees taking inputs from all the base trees.

3.2 Boosting

Boosting is an iterative ensemble model; it is also called as dependent ensemble model. It is an iterative technique which adapts the weight of an observation based on the recently applied classifier. It tries to increase the weight of the observation which was classified incorrectly by the previous classifier. In the first iteration boosting performs normal classification on the given set of data and make predictions. The instances which are misclassified in the first iteration are sent to next iteration with in order to improve the accuracy of weak classifiers. The workings of the iterations are depends on the type of boosting, different types of boosting are gives below:

Ada-Boosting: It is an adaptive boosting technique which uses weighted approach for misclassified instances. In the very first iteration all the instances are assigned equal weights and predictions are made, after first iteration the instances which were misclassified are assigned higher weights than other instances in order to improve the predictability of the model. This process continues till the most accurate prediction model is built.

Gradient Boosting: Gradient boosting runs on the same principal of traditional boosting. In the first iteration a simple classification model is built to predict outcomes over the given set of data**,** and then from the misclassified instances a loss function is plotted. Now the original plot and the error plot are combined together to built a new plot which yields more prediction accuracy than the base predictor, this process of plotting error plot continues till the model finds minimum error points. In this process the error is sequentially reduced after every iteration and hence it produces strong prediction model.

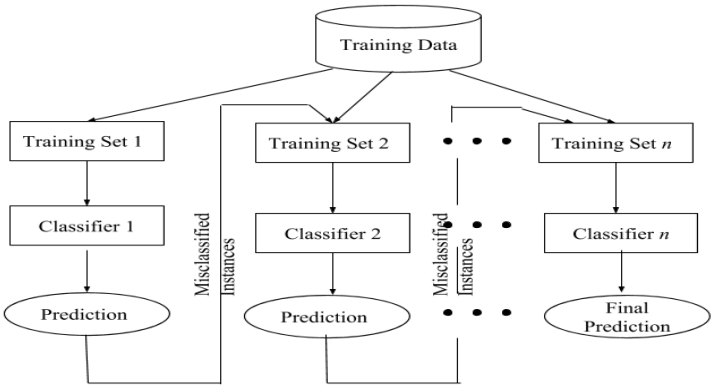

Figure 2: Boosting Model

3.3 Stacking

Stacking is similar to bagging. The basic framework of both models is similar the only difference between stacking and bagging is in stacking a meta classifier or meta learner is used for final prediction. Stacking model is also called as super learner. The generalization ability of stacking is far more than other models. The first layer of stacking is same as bagging where the base classifiers make predictions based on the dataset provided to them, output from all this base classifier is given as an input to the meta classifier. The meta classifier produces output considering new attributes which are predictions made by base classifiers, target variable will remain same as original dataset. While testing the testing instances are initially classified by the base classifiers, predictions made by these classifiers are added as a new attribute to the dataset and it is fed to the meta classifier. The meta classifier combines the different outputs of base classifiers into the final prediction.

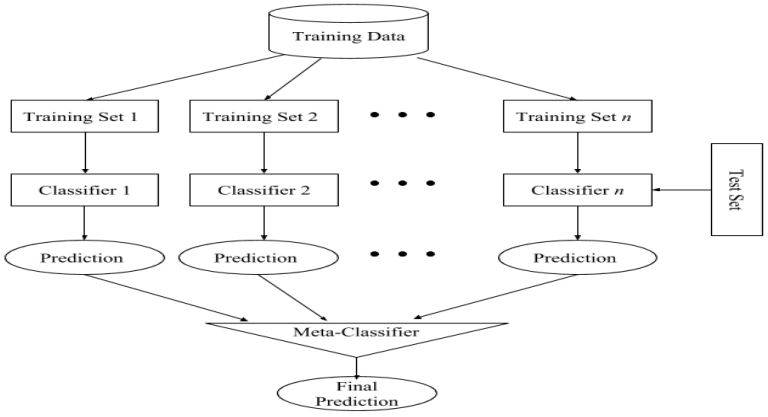

Figure 3: Stacking Model

Results and Discussion

The articles studied and analyzed are given in Table 1. These articles are analyzed based on the methods and algorithms used for the research.

Sr no	Title of Article	Methods used	Type	Algorithms used
1	Modeling Insurance Fraud Using Using Ensemble Combining Classifier [1]	Amira Kamil et al.	Ensemble	SVM, ANN, DT
2	A novel hybrid under sampling method for mining unbalanced datasets in banking and insurance[2]	G.Ganesh Sundar kumar et al.	Supervised	DT, SVM, LR, PNN, MLP,GMDH, *k*-Reverse Nearest Neighbor
3	Research and Application of Random Forest Model in Mining Automobile	Yaqi Li et al.	Supervised	Random Forest

	Insurance Fraud[3]			
4	Fraud Detection in Health Insurance using Data Mining Techniques[4]	Vipula Rawt et.al	Hybrid	SVM, K-means Clustering
5	Credit Card Fraud Detection: A Hybrid Approach Using Fuzzy Clustering & Neural Network[24]	Tanmay Kumar Behera et al.	Hybrid	Fuzzy Clustering, Neural Network
6	A Hybrid Outlier Detection Algorithm Based On Partitioning Clustering And Density Measures[25]	Hamada Rizk et al.	Hybrid	K-Medoids
7	A Principle Component Analysis-based Random Forest with the Potential Nearest Neighbor Method for Automobile Insurance Fraud Identification[6]	Yaqi Li et al.	Ensemble	Random Forest, Principle Component Analysis , Potential Nearest Neighbor
8	A Case Study of Applying Boosting Naive Bayes to Claim Fraud Diagnosis[9]	Stijn Viaene et al.	Ensemble	Naive Bayes
9	Pattern Discovery on Australian Medical Claims Data—A Systematic Approach[10]	Ah Chung Tsoi et al.	Unsupervised	Clustering , Hidden Markov Model
10	Combining Re-sampling with Twin Support Vector Machine for Imbalanced Data Classification[11]	Lu Cao et al.	Supervised	Twin SVM
11	Fraud Detection and Frequent Pattern Matching in Insurance claims using Data Mining Techniques[12]	Aayushi Verma et. al	Unsupervised	K- Means Clustering
12	Random Rough Subspace based Neural Network Ensemble for Insurance Fraud Detection[14]	Wei Xu et al.	Ensemble	Neural Network, OSS, Resilient back propagation
13	Framework for the Identification of Fraudulent Health Insurance Claims using Association Rule Mining[18]	Saba Kareem et al.	Unsupervised, Supervised	Clustering , Apriori Algorithm, SVM

Table 1: Review Articles

Conclusion

This survey has explored the machine learning techniques used in insurance fraud prediction. Machine learning approach provides the vast range of methods and algorithms for fraud prediction. Supervised and unsupervised learning methods are widely used in combination with other methods to improve the prediction accuracy of the model. Hybrid learning methods provides flexibility to user by blending different algorithms together these techniques have outperformed than that of the traditional learning methods. Ensemble learning is gaining more importance recently due to its reliability and flexibility with different approaches. In few recent studies it is revealed that ensembles not only improve prediction accuracy but they also deal with some chronic machine learning problems such as over-fitting, class imbalance and concept drift. Ensemble models and their applications are tempting because of their generalization ability. Ensembles are expensive to build in terms of both time and resources but this could be seen as one time investment because once the ensemble is assembled it produces highly efficient results.

References

[1] Amira Kamil, Ibrahim Hassan and Ajith Abraham."*Modeling Insurance Fraud Detection Using Ensemble Combining Classification*". International Journal of Computer Information Systems and Industrial Management Applications. ISSN 2150-7988 Volume 8 (2016)

[2] G. Ganesh Sundarkumar, Vadlamani Ravi." *A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance*". Engineering Applications of Artificial Intelligence 37 (2015)

[3] Yaqi Li, Chun Yan,Wei Liu, Maozhen Li "*Research and Application of Random Forest Model in Mining Automobile Insurance Fraud*". International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)(2016)

[4] Vipula Rawte, G Anuradha."*Fraud Detection in Health Insurance using Data Mining Techniques*". 2015 International Conference on Communication, Information and Computing Technology (ICCICT), Jan.16-17

[5] Siddhartha Bhattacharyya , Sanjeev Jha , Kurian Tharakunnel , J. Christopher Westland."*Data mining for credit card fraud: A Comparative Study*". Decision Support Systems 50 (2011)

[6] Yaqi Li, Chun Yan, Wei Liu, Maozhen Li."*A Principle Component Analysis-based Random Forest with the Potential Nearest Neighbor Method for Automobile Insurance Fraud Identification*". Applied Soft Computing Journal

[7] Stijn Viaene, Mercedes Ayuso , Montserrat Guillen ,Dirk Van Gheel, Guido Dedene."*Strategies for detecting fraudulent claims in the automobile insurance industry*". European Journal of Operational Research 176 (2007)

[8]Michal Wozniak, Manuel Grana, Emilio Corchado.: "*A Survey of Multiple Classifier Systems as Hybrid Systems Information Fusion*" (2013)

[9] Stijn Viaene, Richard A. Derrig, and Guido Dedene."*A Case Study of Applying Boosting Naive Bayes to Claim Fraud Diagnosis*." IEEE Transactions On Knowledge And Data Engineering, Vol. 16, No. 5, May 2004.

[10] Ah Chung Tsoi, Shu Zhang, and Markus Hagenbuchner. "*Pattern Discovery on Australian Medical Claims Data: A Systematic Approach.*" IEEE Transactions On Knowledge And Data Engineering, Vol. 17, No. 10, October 2005.

[11] Lu Cao,Hong Shen."*Combining Re-sampling with Twin Support Vector Machine for Imbalanced Data Classification*". International Conference on Parallel and Distributed Computing, Applications and Technologies 2016 17.

[12] Aayushi Verma, Anu Taneja, Anuja Arora."*Fraud Detection and Frequent Pattern Matching inInsurance claims using Data Mining Techniques*". Proceedings of 2017 Tenth International Conference on Contemporary Computing (IC3), 10-12 August 2017, Noida, India

[13] Lior Rokach. "*Ensemble-based classifiers*". Springer Science+Business Media B.V. 2009

[14] Wei Xu, Shengnan Wang, Dailing Zhang, Bo Yang. "*Random Rough Subspace based Neural Network Ensemble for Insurance Fraud Detection*". Fourth International Joint Conference on Computational Sciences and Optimization 2011.

[15] Yi Peng, Gang Kou, Alan Sabatka, Zhengxin Chen, Deepak Khazanchil, Yong Shi3 "*Application of Clustering Methods to Health Insurance Fraud Detection*". 1-4244-0451-7/06/$20.00 C2006 IEEE.

[16] Dr.M.S. Anbarasi, S. Dhivya. "*Fraud Detection Using Outlier Predictor In Health Insurance Data*". International Conference On Information, Communication & Embedded Systems (Icices 2017).

[17] Riya Roy , Thomas George K."*Detecting Insurance Claims Fraud Using Machine Learning Techniques*". International Conference on circuits Power and Computing Technologies [ICCPCT] 2017.

[18] Saba kareem Dr. Rohiza Binti Ahmad Dr. Aliza Binit Sarlan."*Framework for the Identification of Fraudulent Health Insurance Claims using Association Rule Mining*". IEEE Conference on Big Data and Analytics (ICBDA)2017.

[19] Chun Yan, Yaqi Li "*The Identification Algorithm and Model Construction of Automobile Insurance Fraud Based on Data Mining*". Fifth International Conference on Instrumentation and Measurement, Computer, Communication and Control 2015.

[20] Stijn Viaene, Richard A. Derrig, Bart Baesens, Guido Dedene "*A Comparison Of State-Of-The-Art Classification Techniques For Expert Automobile Insurance Claim Fraud Detection*". The Journal of Risk and Insurance, 2002, Vol. 69, No. 3, 373-421.

[21] E.W.T. Ngai , Yong Hu , Y.H. Wong , Yijun Chen , Xin Sun "*The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature*." Decision Support Systems 50 (2011) 559–569.

[22] S. Viaene, G. Dedene, R.A. Derrig "*Auto claim fraud detection using Bayesian learning neural networks*". Expert Systems with Applications 29 (2005) 653–666.

[23] Richard A. Bauder, Taghi M. Khoshgoftaar. "*Medicare Fraud Detection using Machine Learning Methods*". 16th IEEE International Conference on Machine Learning and Applications 2017.

[24] Tanmay Kumar Behera, Suvasini Panigrahi."*Credit Card Fraud Detection: A Hybrid Approach Using Fuzzy Clustering & Neural Network*". Second International Conference on Advances in Computing and Communication Engineering 2015.

[25] Hamada Rizk, Sherin Elgokhy, Amany Sarhan."*A Hybrid Outlier Detection Algorithm Based On Partitioning Clustering And Density Measures.*" 978-1-4673-9971-5/15/$31.00 ©2015 IEEE.

