Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 10, October 2021: 1666-1670

Comparative study of various Machine Learning Algorithms using Finance Industry

T.LOGESWARI

Associate Professor
Department of Computer Science
New Horizon College
Bangalore

Abstract—. Finance sector is the wealth backbone of any country, so risk assessment and fraud detection have great importance. Risk assessment is the process of identifying vulnerabilities to an organization by identifying risk involved in each and every new plans, policies or investments. This paper concentrates on risk level detection of loan application and insurance claim and suggests a predictive model for risk assessment and fraud detection using three efficient machine learning algorithms after applying under sampling technique on data and compares the accuracy difference of them, on imbalanced and resampled data sets with the leading machine learning algorithms Random Forest ad SVM (support vector machine).

Key words: Machine Learning, Finance Sector, Risk Assessment, Fraud Detection, Accuracy, Algorithm Parameters

I. INTRODUCTION

Machine learning has great influence on finance sector which includes a wide range of companies and organizations involved with money, like money lending, investing, insuring and securities issuance and trading services. Machine learning (ML) can be used to find the interesting and useful information from the data. It can be applied on important processes like risk management and fraud predictions. Appropriate decisions should be taken throughout these stages by the decision maker to avoid the great loss.ML can contribute well for the appropriate decision-making process by learning the machine with available data set and by training the machine with efficient machine learning algorithms. If the available data set contains the classification of each instance, then supervised learning algorithm is used. If the data set doesn't contain the classification, then unsupervised methods are used and if the data set gives classification for only some instance, then the machine have to extract the rule through its experience and reinforced techniques are used. In this paper supervised learning algorithms are used because of the classification is already given in the data set. Algorithms perform differently for the different data set. The reasons are the size of data set, number of attributes, imbalance problem, missing values and value type of data set.

II. RELATED WORK

'Data Mining: Current Applications & Trends' by[1] Sedhant Sethi says that, large amount of data is available, but these data has no use until it is changed into some useful information. This information

can be extracted from theavailable raw data and this information is required to be processed and scanned for taking useful and accurate decisions and predictions (or forecasting). This paper also describes the different applicable areas where data mining can be used-education, banking, retail industry, telecommunication, forecasting, science and engineering, web mining, fraud detection, intrusion detection, financial data analysis, business analytics etc. In 'Implementation of Data Mining Techniques in Upcoding Fraud Detection in the Monetary Domains',[2] Dr.Mrs Ananthi Sheshasayee and Surya Susan Thomas givesan insight into the various data mining Techniques which are efficient in detecting upcoding frauds especially in the healthcare insurance sector in India. Qiang Liu in his paper,' A Survey on Security Threats and Defensive Techniques of Machine Learning: A Data Driven View',[3] addresses security threats of data mining techniques like and give a systematic survey on them from two aspects, the training phase and the testing/inferring phase[8]. They categorize current defensive techniques of machine learning into four groups: security assessment mechanisms, counter measures in the training phase, those inthe testing or inferring phase, data security and privacy. Finally, they provide five notable trends in the research on security threats and defensive techniques of machine learning, which are worth doing in-depth studies in future.

III. Applying Machine Learning Algorithms

Three leading classification algorithms are used for training purpose. Random Forest, SVM (Support vector machine), ANN (Artificial neural network).

1) Random Forest

Random forest is a supervised learning algorithm which will works well for classification and regression problems. Random tree is the collection of trees which is called forest mostly trained with the "bagging" method[]. Random forest builds multiple decision trees and merges the result of each tree to get an accurate prediction. The tree consists of a root node and child nodes. Each internal node represents the test on the features and the branches represents the outcome of thetest and leaf node represents a label or a particular number offeature.

For classification problem, feature vector is randomly taken as input and classifies it with every tree in the forest and outputs the class label that received the majority of "votes". If this is used for regression, the output will be the average of the outputs over all the trees in the forest. All the trees are using same parameters but performed on different training set. Feature vector for input is selected randomly with replacement using bootstrap method. The classification error is estimated internally during the training. The training is done using randomly selected features using sampling with replacement, some vectors are left out. This is called oob data(out of bag).The classification error is calculated using this oob. The parameters of random forest are[5]:

1) Max_depth : depth of the tree
2) Min-sample_count : Minimum sample count needed atthe leaf node
3) Max_categories : value of a categorical variable to findthe suboptimal split
4) Calc_var_importance : calculate the importance ofvariable
5) Nactive_vars : size of randomly selected features

6) Max_num_of_trees_in_the_forest: maximum number oftrees in the forest.

7) Forest _accuracy : sufficient accuracy(OOB error)

8) Termcrit_type : learning termination criteria

Random forest can avoid overfitting problembecause the training is done using sampling. Moreover, it canidentify the most important feature from the training set. Random forest can avoid overfitting problembecause the training is done using sampling. Moreover, it canidentify the most important feature from the training set.

2) SVM (Support Vector Machine)

It is a supervised and binary classifier which which train the labeled data and outputs a line which separates theinstances[4].

$$W=x^2 + y^2 \text{ ---------- } 1$$

It checks whether the data is linearly separable or not. If it is not linearly separable, the data is converted into ahigh dimensional area and outputs the hyperplane which canplace between the two classes.

$$f(x) = \beta_0 + \beta^T x \text{ ---------- } 2$$

Support vector are the points nearest to the line or hyperplane, the points in the data set.

Even though it is a binary classifier, it can be used

for classifying more than two classes. This can be used for both classification and regression problems. It can be used forlarger data sets as the training time with SVMs can be high. The parameters are[]:

1) C : regularization parameter of error term

2) Kernel : kernel type to be used in the classifier .It can belinear, polynomial, sigmoid, precomputed or callable(default is ‘rbf’)

3) Degree : degree of poly(ignored by others, default valueis 3)

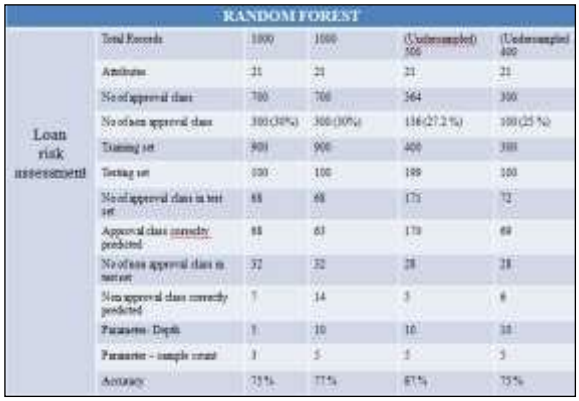
4) Gamma : kernel coefficient of rbf

Two types of training is possible in SVM. One is usual train and the other one is ‘auto’ type. ‘Auto’ type givesmore accuracy because in auto type, first a particular numberof instances are taken and and do the classification. Then thegamma value of that output is taken and again train using thewhole data. SVM can’t read continous values and perform poor for imbalanced data.

V. EXPERIMENTAL RESULTS.

Experiment was done by performing the above explained three algorithms on risk assessment(2 data set) and fraud detection data set(1 data set). Before applying the algorithm the data set is well processed and cleaned. All the missing values are changed, continuous values are converted into discrete and resampling was done for solving the imbalanced data set problem[6]. Undersampling technique is used

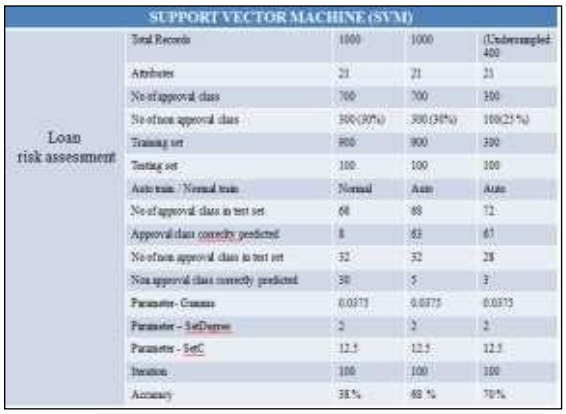
to solve the imbalanced data problem. Then the performance of three algorithms for the three data sets are compared. The accuracy difference on balanced and unbalanced data is well studied. Then the maximum accuracy obtained by each algorithm on the basis of parameter change is examined. As a conclusion the best algorithm for handling risk assessment and fraud detection is suggested[7].



RANDOM FOREST					
Loan risk assessment	Total Records	1000	1000	(Undersampled) 500	(Undersampled 400
	Attributes	21	21	21	21
	No of approval class	700	700	364	300
	No of non approval class	300 (30%)	300 (30%)	136 (27.2 %)	100 (25 %)
	Training set	900	900	400	300
	Testing set	100	100	100	100
	No of approval class in test set	68	68	172	72
	Approval class correctly predicted	68	63	170	69
	No of non approval class in test set	32	32	28	28
	Non approval class correctly predicted	7	14	3	6
	Parameter- Depth	5	10	10	10
	Parameter – sample count	5	5	5	5
	Accuracy	75%	77%	67%	75%

RANDOM FOREST			
Fraud Money Transaction	Total Records	(Undersampled) 3843	(original data) 5999
	Attributes	10	10
	No of nonfraud class	2861	5948
	No of fraud class	984 (26 %)	52 (.87 %)
	Training set	3008	5000
	Testing set	837	999
	No of nonfraud class in test set	701	987
	nonfraud class correctly predicted	701	987
	No of fraud class in test set	136	12
	fraud class correctly predicted	121	0
	Parameter- Depth	10	10
	Parameter – sample count	5	5
	Accuracy	98 %	98 %

Fig. 1: Random forest on loan risk assessment data set Fig. 2: Random forest on fraud money transaction dataset



SUPPORT VECTOR MACHINE (SVM)				
Loan risk assessment	Total Records	1000	1000	(Undersampled 400
	Attributes	21	21	21
	No of approval class	700	700	300
	No of non approval class	300 (30%)	300 (30%)	100 (25 %)
	Training set	900	900	300
	Testing set	100	100	100
	Auto train / Normal train	Normal	Auto	Auto
	No of approval class in test set	68	68	72
	Approval class correctly predicted	8	63	67
	No of non approval class in test set	32	32	28
	Non approval class correctly predicted	30	5	3
	Parameter- Gamma	0.0375	0.0375	0.0375
	Parameter - SetDegree	2	2	2
	Parameter - SetC	12.5	12.5	12.5
	Iteration	100	100	100
	Accuracy	38 %	68 %	70%

SUPPORT VECTOR MACHINE (SVM)			
Fraud money transaction	Total Records	(undersampled) 3843	(original data) 5999
	Attributes	10	10
	No of approval class	2861	5948
	No of non approval class	984 (26 %)	52 (.87 %)
	Training set	3008	5000
	Testing set	837	999
	Auto train / Normal train	Auto	Auto
	No of approval class in test set	701	987
	Approval class correctly predicted	701	987
	No of non approval class in test set	136	12
	Non approval class correctly predicted	0	0
	Parameter- Gamma	0.0325	0.0325
	Parameter - SetDegree	2	2
	Parameter - SetC	12.5	12.5
	Iteration	100	100
	Accuracy	83	83

Fig. 3: SVM on loan risk assessment data set Fig. 4: SVM on fraud money transaction data set

V. CONCLUSION

The experiment was repeated by using original data set and under sampled data set. Classifier algorithms gives high accuracy for undersampled data sets. Then the parameters of algorithms are changed and repeated the experiment to get the maximum accuracy. Random forest is performing efficiently for all the cases. It gives an accuracy of 77 % for the loan risk assessment data set when the depth and sample count is adjusted to 10 and 5 and the event rate is 30 %, whereas SVM shows accuracy less than this. It gives a maximum accuracy of 98 % on fraud money transaction data set when the data is undersampled with event rate 26 %. Again it produces only 40% accuracy on fraud money transaction data set with event rate of each class less than 20 %, but the highest accuracy than the other two algorithms. So, it can be concluded that, classes in the available data set should balance with each other. Mostly the event rate should be greater than 25%, then only algorithms will provide better accuracy. Random forest can be considered as the best algorithm for imbalanced data set and can be used as an efficient algorithm for risk assessment and fraud detection prediction in finance sector.

REFERENCES

[1] Sidhanth Sethi, Dheeraj Malhothra, Neha Varma, " Datamining current Applications and Trends" , IJIET, Vol-6, Issue-4, 2016

[2] Dr.Mrs.Ananthi Sheshasayee, Surya Susan Thomas,"Implementation of Data Mining techniques in upcoding fraud detection in the monetary domains", ICIMA (International Conference on Innovative Mechanism for Industry Applications), 2017

[3] Qian Liu, Pan Liu, Wentao Zao, Wei Ca , Shui yu, Vectror C.M , Leug, "A Survey on Security Threats andDefensive Techniques of Machine Learning: A Data Driven View", IEEE 2016

[4] Hasan Ansari Arief, Putri Saptawati, Yudistira Dwi,Woudhana Ansar, " Fraud Detection based on Data Mining an Indonesian E-procuremnt system(SPSE)",
IEEE 2016

[5] R.Meenatkshi, Sivaranjari, " Fraud Detection on Fianancial Statement using Data Mining Techniques andPerformance Analysis", IJITA, 2016

[6] http://www.hackerearth.com/blog/machine- learning/simple-tutorial-svm-parameter-tuning-python- r/

[7] https://www.capgemini.com/wp- content/uploads/2017/07/fraud_solution_for_financial_s ervices_with_sas.pdf

[8] https://www.cbn.gov.ng/out/2016/mpd/understanding% 20monetary%20policy%20series%20no%2040.pdf