Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 10, October 2021: 1723 - 1726

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 datasets with the leading machine learning algorithms Random Forest ad SVM (support vector machine).

Keywords:MachineLearning,FinanceSector,RiskAssessment,FraudDetection,Accuracy,AlgorithmParameters

I. INTRODUCTION

Machinelearninghasgreatinfluenceonfinancesectorwhichincludesawiderangeofcompaniesandorganizationsinvolvedwithmoney,likemoneylending,investing,insuringandsecuritiesissuanceandtradingservices.Machinelearning (ML) can be used to find the interesting and usefulinformation from the data. It can be applied on importantprocesseslikeriskmanagementandfraudpredictions.Appropriatedecisionsshouldbetakenthroughoutthesestagesbythedecisionmaker to avoid thegreatloss.ML can contribute well for the appropriate decision-making process by learning the machine with available datasetandbytrainingthemachinewithefficientmachinelearning algorithms. If the available data set contains theclassificationofeachinstance,thensupervisedlearningalgorithmisused.Ifthedatasetdoesn’tcontainthe classification, then unsupervised methods are used and if thedata set gives classification for only some instance, then themachine have to extract the rule through its experience andreinforcedtechniquesareused.Inthispapersupervisedlearning algorithms are used because of the classification isalready given in the data set. Algorithms perform differentlyfor the different data set. The reasons are the size of data set,numberofattributes,imbalanceproblem,missingvaluesandvaluetype ofdata set.

II. RELATEDWORK

‘DataMining:CurrentApplications&Trends’by[1]SedhantSethisaysthat,largeamountofdataisavailable, butthese data has no use until it is changed into some usefulinformation.Thisinformationcanbeextractedfromtheavailable raw data and this information is required to beprocessedandscannedfortakingusefulandaccuratedecisions and predictions (or forecasting). This paper alsodescribes the different applicable areas where data miningcanbeusededucation,banking,retailindustry,telecommunication,forecasting,scienceandengineering,web mining, fraud detection, intrusion detection, financialdataanalysis, businessanalyticsetc. In 'Implementation of Data Mining Techniques inUpcoding Fraud Detection in the Monetary Domains',[2]Dr.MrsAnanthiSheshasayeeandSuryaSusanThomasgivesaninsightintothevariousdatami ningTechniqueswhichareefficientindetectingupcodingfraudsespeciallyinthehealthcareinsurance sectorinIndia. QiangLiuinhispaper,'ASurveyonSecurityThreats and Defensive Techniques of Machine Learning: AData Driven View',[3] addresses security threats of datamining techniques like and give a systematic survey on themfrom two aspects, the training phase and the testing/inferringphase[8].Theycategorizecurrentdefensivetechniquesofmachinelearningintofourgroup s:securityassessmentmechanisms,countermeasuresinthetrainingphase,thoseinthetestingorinferringpha se,datasecurity andprivacy.Finally, they provide five notable trends in the research onsecuritythreatsanddefensivetechniquesofmachinelearning, whichareworthdoingin-depthstudiesinfuture.

III. ApplyingMachineLearningAlgorithms

Three leading classification algorithms are used for trainingpurpose. Random Forest, SVM (Support vector machine),ANN(Artificialneural network).

1)RandomForest

Random forest is a supervised learning algorithm which willworkswellforclassificationandregressionproblems.Random tree is the collection of trees which is called forestmostly trained with the "bagging" method[]. Random forestbuilds multiple decision trees and merges the result of eachtree to get an accurate prediction. The tree consists of a rootnode and child nodes. Each internal node represents the testonthefeaturesandthebranchesrepresentstheoutcomeofthetestandleafnoderepresentsalabeloraparticu larnumberoffeature.

Forclassificationproblem,featurevectorisrandomlytakenasinputandclassifiesitwitheverytreeint heforestandoutputstheclasslabelthatreceivedthemajorityof"votes". If this is used for regression, the output will be theaverage of the outputs over all the trees in the forest. All thetrees are using same parameters but performed on differenttraining set. Feature vector for input is selected randomlywith replacement using bootstrap method. The classificationerror is estimated internally during the training. The trainingisdoneusingrandomlyselectedfeaturesusingsamplingwithreplacement, some vectors are left out. This is called oobdata(out of bag).The classification error is calculated usingthisoob. The parametersofrandomforestare[5]:

1) Max_depth:depthofthetree
2) Min-sample_count:Minimumsamplecountneededattheleafnode
3) Max_categories:valueofacategoricalvariabletofindthesuboptimal split
4) Calc_var_importance:calculatetheimportanceofvariable
5) Nactive_vars:sizeofrandomlyselectedfeatures
6) Max_num_of_trees_in_the_forest:maximumnumberoftreesinthe forest.
7) Forest_accuracy:sufficientaccuracy(OOBerror)
8) Termcrit_type:learningterminationcriteria

Randomforestcanavoidoverfittingproblembecause the training is done using sampling. Moreover, it canidentifythemostimportantfeaturefromthetrainingset.Randomforestcanavoidoverfittingproblembec ause the training is done using sampling. Moreover, it canidentifythemostimportantfeaturefromthetrainingset.

2)SVM(SupportVectorMachine)

It is a supervised and binary classifier which which train thelabeleddataandoutputsalinewhichseparatestheinstances[4].

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

It checks whether the data is linearly separable ornot. If it is not linearly separable, the data is converted into ahigh dimensional area and outputs the hyperplane which canplacebetweenthe twoclasses.

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

Support vector are the points nearest to the line or hyperplane,thepointsinthe data set.
Even thoughitisabinaryclassifier,itcanbeused for classifying more than two classes. This can be used forbothclassificationandregressionproblems.Itcanbeusedforlarger data sets as the training time with SVMs can be high.Theparametersare[]:

1) C:regularizationparameteroferrorterm
2) Kernel:kerneltypetobeusedintheclassifier.Itcanbelinear, polynomial, sigmoid, precomputed or callable(defaultis‘rbf’)
3) Degree : degree of poly(ignored by others, default valueis3)
4) Gamma:kernelcoefficientofrbf

Two types of training is possible in SVM. One isusualtrainandtheotheroneis‘auto’type.‘Auto’typegivesmoreaccuracybecauseinautotype,firstaparticularnumberof instances are taken and and do the classification. Then thegammavalueofthatoutputistakenandagaintrainusingthewhole data. SVM can’t read continous values and performpoorfor imbalanceddata.

V. EXPERIMENTALRESULTS.

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	199	100
	No of approval class in test set	68	68	171	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	3	5	5	5
	Accuracy	75%	77%	87%	75%

RANDOM FOREST			
Fraud Money Transaction	Total Records	(Undersampled) 3845	(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:Randomforestonloan riskassessmentdataset
Fig.2:Randomforestonfraudmoneytransactiondata set

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%

Fig.3:SVMonloanriskassessmentdataset

SUPPORT VECTOR MACHINE (SVM)			
Fraud money transaction	Total Records	(undersampled) 3845	(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.4:SVMonfraudmoneytransactiondataset

V. CONCLUSION

Theexperimentwasrepeatedbyusingoriginaldatasetand under sampleddataset.Classifieralgorithmsgiveshigh accuracy for undersampled data sets. Then the parameters ofalgorithmsarechangedandrepeatedtheexperimenttogetthemaximum accuracy. Random forest is performing efficientlyforallthecases.Itgivesanaccuracyof77%fortheloanriskassessment data setwhen the depth and sample count isadjustedto10and5andtheeventrateis30%,whereasSVMshows accuracy less than this. It gives a maximumaccuracy of 98 % on fraud money transaction data set whenthe data is undersampled with event rate 26 %. Again itproducesonly40%accuracyonfraudmoneytransactiondatasetwitheventrateofeachclasslessthan20%,b utthehighest accuracythantheothertwoalgorithms.So,itcanbeconcludedthat,classesintheavailabledatasetshouldbala nce with each other. Mostly the event rate should begreater than 25%, then only algorithms will provide betteraccuracy.Randomforestcanbeconsideredasthebestalgorithm for imbalanced data set and can be used as anefficient algorithm for risk assessment and fraud detectionpredictioninfinancesector.

REFERENCES

[1] SidhanthSethi,DheerajMalhothra,NehaVarma,“Dataminingcurrent ApplicationsandTrends”,IJIET,Vol-6, Issue-4,2016

[2] Dr.Mrs.AnanthiSheshasayee,SuryaSusanThomas,”Implementation of Data Mining techniques in upcodingfrauddetectioninthemonetarydomains”,ICIMA(International Conference on Innovative Mechanism forIndustry Applications), 2017

[3] Qian Liu, Pan Liu, WentaoZao, Wei Ca , Shui yu,Vectror C.M , Leug, “A Survey on Security Threats andDefensiveTechniquesofMachineLearning:A DataDrivenView”,IEEE 2016

[4] HasanAnsariArief,PutriSaptawati,YudistiraDwi,WoudhanaAnsar,“FraudDetectionbasedonDataMiningan IndonesianE-procuremntsystem(SPSE)”, IEEE2016

[5] R.Meenatkshi,Sivaranjari,“FraudDetectiononFianancialStatementusingDataMiningTechniquesandPerformanceAnalysis”, 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_services_with_sas.pdf

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