Classification and Fraud Detection in Finance Industry

Akshansh Sinha
Independent
Bangalore

Shivam Mokha
Independent
Bangalore

ABSTRACT

Due to increase of fraud which results in loss of money across the globe, several methodologies and techniques developed for detecting frauds Fraud detection involves analysing the activities of users in order to understand the malicious behaviour of users. Malicious behaviour is a broad term including delinquency, fraud, intrusion, and account defaulting. This paper presents a survey of current techniques used in credit card fraud detection and evaluates a new hybrid approach to identify fraud detection. The paper also discusses popular algorithms used for unsupervised and supervised learning.

General Terms

Pattern Recognition, Machine Learning, Fraud Detection

Keywords

Fraud detection, Data mining, Machine learning, SVM, Genetic Algorithms, Anomalies.

1. INTRODUCTION

With the emerging rise of technology today, the dependency on e-commerce and the online payments has grown exponentially. As the credit card provides convenience to the users but frauds caused due to these activities causes inconvenience. The credit card information is confidential, the bank and the other financial enterprises doesn't want to disclose the information about their customers. Risk management is critical for financial enterprises to survive in such competing industry. The provisional loss arises due to the “bad” accounts bank lends the money to customers who eventually do not have capability to pay back. In the risk management, the chances of false negative (false “good” accounts) could still be high. However, by leveraging their performance such as credit card utilization, payment information, risks can further be managed to control provisional loss. In this paper, a focus on risk management as well as fraud detection is depicted.

1. It shows an interest in classifying if a booked account as a “bad” account within 12 months since booked. Since an internal classification model is already available, with a secondary interest to train a better classifier to outperform the benchmark model.

2. Since there are few research initiatives that implements fraud detection. Concentration on how to optimize fraud detection techniques is brought to light. Since the emergence of many advanced computing and classification systems including the support vector machine and the optimization technique like genetic algorithm shows a greater fluctuation in the implementation of many different technologies due to the accuracy and efficiency it produces. This research uses a hybrid approach of Genetic Algorithm and Support vector machines to perform fraud detection.

2. DATA SET DESCRIPTION

The proposed fraud detection model is based on building up of a classifier through processes including pre-processing, clustering, feature selection and SVM training. The UCSD dataset used in the proposed work consists of two sets of datasets including training set and testing test. It contains 1 lakh transactions of train data and 50000 transactions of test data where there are numerous vague transactions exist. Hence in order to remove such vague transactions the dataset should be subjected to certain processes like pre-processing. Through the pre-processing technique the single as well as anonymous transactions are filtered and it outputs the reduced datasets with relevant transactions only (train set with 21,850 transactions and test set with 9.425 transactions).

The dataset should necessarily contain a class indicating whether each transaction is either legal or fraud. Since the UCSD dataset do not have such a label, it should be labelled to the respective classes by subjecting it to the required process like clustering. The clustering phase actually works on the pre-processed data which is the output of pre-processing phase.

3. DATA PROCESSING

The main aim of the pre-processing module is to preprocess the datasets. It aims to obtain the datasets consisting of only relevant transactions with no unique transactions. The training set and the test set of the UCSD dataset are the inputs of the data pre-processing. This processing is done separately to both the datasets so as to make it into a reduced form removing the unwanted transactions which are mainly the single ones and thereby keeping only necessary number of transactions. The pre-processing is done to the train/test data by listing out the unique customer ID’s. Out of the total number of 19 attributes in both the datasets search for the attribute “custattr1” which refers to the customer ID. For each customer ID, scan the entire dataset for the respective transactions of the same ID. Only if there are more than two transactions exist for a particular customer ID, it is kept in the pre-processed lists of transactions else it will be removed. This procedure is done for train set and test set and thus obtains the pre-processed datasets as outputs. After pre-processing the train set has been reduced to 21,850 numbers of transactions and test set has been reduced to 9,425 numbers of transactions.

Algorithm 1: (Data Pre-processing):

```
Input: Load the train / Test data
Output: Pre-processed list of data
Initialize: Attribute matrix
CustId = find ( fieldTitles , custattr1)
For I =1 to n
If (CustId>2)
Add ( CustId Accepted Set of Users)
End for
End if
```

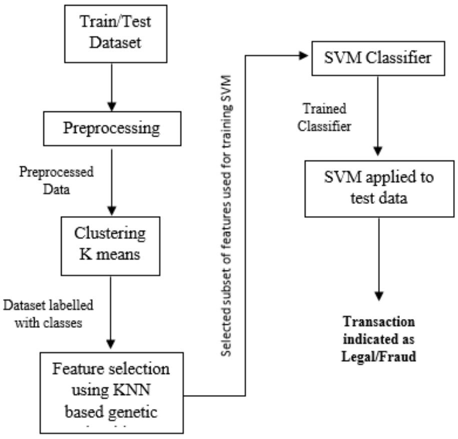

Train/Test Dataset
Preprocessing
Preprocessed Data
Clustering K means
Dataset labelled with classes
Feature selection using KNN based genetic
Selected subset of features used for training SVM
SVM Classifier
Trained Classifier
SVM applied to test data
Transaction indicated as Legal/Fraud

Fig. 1: The block diagram

4. SYSTEM DESIGN

The proposed fraud detection model is based on building up of a classifier through processes including pre-processing, clustering, feature selection and SVM training. The UCSD dataset used in the proposed work consists of two sets of datasets including training set and testing test. It contains 1 lakh transactions of train data and 50000 transactions of test data where there are numerous vague transactions exist. Hence in order to remove such vague transactions the dataset should be subjected to certain processes like pre-processing. Through the pre-processing technique the single as well as anonymous transactions are filtered and it outputs the reduced datasets with relevant transactions only (train set with 21,850 transactions and test set with 9.425 transactions). Thus, the dataset is now ready for the next phase of the work which is the clustering phase. The dataset should necessarily contain a class indicating whether each transaction is either legal or fraud. Since the UCSD dataset do not have such a label, it should be labelled to the respective classes by subjecting it to the required process like clustering. The clustering phase actually works on the pre-processed data which is the output of pre-processing phase. The train and test datasets are clustered by using K-Means clustering approach. K-Means iteratively clusters the data and finally outputs the two clusters. From the two clusters obtained, take minimum population of the cluster containing the transactions which are to be labelled as fraud and the other cluster containing the population of transactions would be labelled as legal. Thus, the labelled dataset is now ready. It is then subjected to the feature selection process in which the potential attributes are given more preference.

In order to make the final SVM classifier achieve good performance, selection of best combination of features from the total features is vital for training the SVM classifier. Genetic algorithm with K-Nearest Neighbour method is used for selecting the best subset of features. Initial population is randomly generated of genome length 8 which is the total number of attributes as that of the dataset and that is referred to as chromosomes. KNN method is used for fitness function evaluation in which the nearest neighbours of different combinations of attributes are calculated. The next generations are then created by selecting the chromosomes of high fitness values. By applying genetic operators like crossover and mutation next generations are created. The best individuals are selected by using tournament selection method. This is repeated till the best individuals are obtained. This way the informative features are selected for classification and are given as input to the next phase. The last phase comes the construction of SVM classifier which should be trained with the selected features that are obtained. With these specific features SVM classifier is constructed by training it with the train data. Since the data are to be separated only to two classes i.e.; legal or fraud, the default linear kernel function is used which can give good classification performance. As the dataset do not contain larger attributes it is best to use linear kernel for faster training. The input parameters are set to default values. After constructing the SVM classifier the test data can be applied to it and each transaction can be classified as legal or fraud.

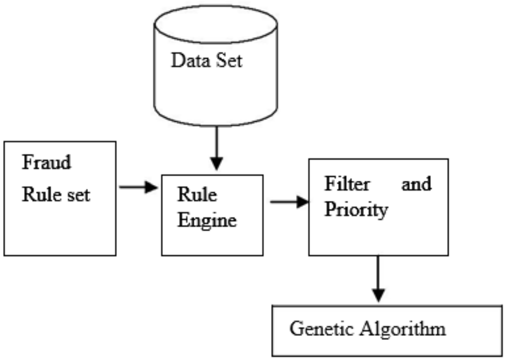


Fig. 2: System design

5. CLUSTERING

Clustering is a process of organizing data into clusters such that there is high intra class similarity and low inter class similarity. The former refers to that how closely the units of same group resemble each other and latter refers to how the units of different groups resembles. Hence it divides the data into groups of similar objects. It is an unsupervised machine learning technique in which the data has been analysed to perform the grouping without having any explicit training. It means that for a given unlabelled dataset it performs the natural grouping of instances without having any sort of learning approach with pre-labelled instances as in supervised learning. To form clusters, it is necessary to check whether the data points are close to each other and it is done by measuring distances among data points. In the proposed work, the clustering comes to role as to label the UCSD transactions. Through clustering the transactions are clustered into two sets. Each transaction is referred to as a data point and the distance among them are found so as to group it. This way the transactions are labelled into corresponding classes, that is, legal or fraud. In the proposed work, the dataset contains total of eight potential attributes which are of Boolean values. By taking random weight value (here value is taken as 2) the dataset is partitioned. h = c/w is calculated in which C=8, where 'C' refers to the total attributes, 'W'= 2 where W is the random weight value. And 'h' is the partition value, the value of h is obtained as 4. The attributes would then be divided into two parts containing each of four attributes out of total eight. The Boolean values of each partitioned block are then converted to its respective decimal values. These are then subjected to perform K Means clustering process. The data points are plotted and the initial centroid are randomly assigned. The distance measure of data points to the centroid is calculated. The distance measure used is the city block distance.

City block distance measure also known as Manhattan distance or absolute value distance. It examines the absolute differences of coordinates of a pair of objects. Performing city block distance measure between data points and centroids. It clusters the data point to the centroids where the distance found to be minimum. New centroids are then calculated for each cluster and the process is repeated until there is no convergence. Finally, the two clusters are obtained with different population. The minimum population cluster would be taken as fraud class and the other population of cluster would be taken as legal class.

Algorithm 2 Clustering

```
Input: Set of Pre-processed datasets
Output: population of cluster1 and cluster2
Initialize w= [2], c=8
Compute h= c/ w[i]
$X_1$ = bin to dec (1 to h) $X_2$ = bin to dec (h+1 to c)
Input for K Means: $X_1$, $X_2$
Initialize centroids: $c_1 c_2 \ldots\ldots c_k$, number of clusters = 2
While (no change in mean)
    For j =1 to k
        Compute $a_j – b_j$
        $C_j$ = new mean ($c_1 c_2 \ldots\ldots\ldots c_k$)
        Print C1, C2
    End for
End while
Fraud (min (class1, class2))
```

6. FEATURE SELECTION USING GENETIC ALGORITHM

The feature selection refers the task of identifying and selection a useful subset of features to be used to represent patterns from a larger set of features. Feature selection plays a vital role in optimizing the performance of the classifier. Training a classifier with many attributes is tedious and that increases the computational costs. Hence choosing the attributes that potentially contributes to the class is relevant and so does the feature selection. This work proposes a method using genetic algorithm to identify subset of features combinations from the set of attributes so as to improve the classification accuracy. These combinations of features are used to train SVM and finally classifier is prepared for classification. GA based KNN approach i.e.; genetic algorithm with K nearest neighbour approach have been used to select out the feature combination from the list of random combinations.

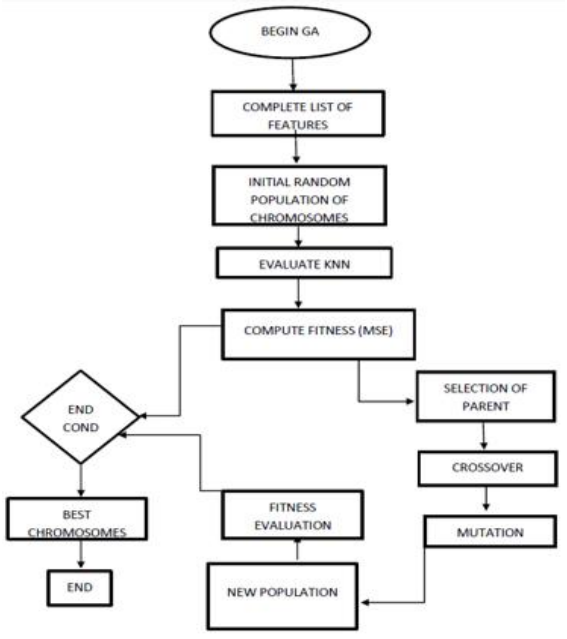


Fig. 3: Genetic Algorithm

Genetic algorithms are evolutionary algorithms which aim at obtaining better solutions as time progresses. Since their first introduction by Holland, they have been successfully applied to many problem domains from astronomy to sports, from optimization to computer science, etc. They have also been used in data mining mainly for variable selection and are mostly coupled with other data mining algorithms. The algorithm begins with multi-population of randomly generated chromosomes. These chromosomes undergo the operations of selection, the attributes of the dataset TS, PT, TR, PU, PM, PV, AS AND SS are the features. These 8 attributes are the total features that are taken for GA. The random population is created of genome length 8 which refers to the set of combinations. The gene value ‘1’ depicts that crossover and mutation. Crossover combines the information from two parent chromosomes to produce new individuals, exploiting the best of the current generation, while mutation or randomly changing some of the parameters allows exploration into other regions of the solution space. Natural selection via the fitness function assures that only the best fit chromosomes remain in the population to mate and produce the next generation. Upon iteration, the genetic algorithm converges to a global solution. In the proposed work, the chromosomes are the particular feature indexed by the position of 1 is selected. Otherwise (if it is 0) the feature is not selected for chromosomal evaluation. For example, from the figure the first row has the values 0 1 1 1 0 1 1 0 which implies the features 2 3 4 6 7 are selected.

The chromosomes refer to the random population of combinations created of genome length 8 indicating the population containing eight attributes. These are one set of input. Training data of eight attributes and the class vector constructed using K Means are the other set of inputs. Aim of KNN is to select out the combination having great significance contributing to the class. For each combination selected from the random population a new class vector is created by finding its nearest neighbours with respect to the train data. For each combination, the columns in the train data according to the value of given combination are taken. Hence a set of transactions for a given combination is obtained. Euclidean distance measure is computed on each row of transaction in the combination set of data to the original training data. K value taken as smallest indicates the probability of having effective results. Here k value is taken as 3. Hence 3 nearest neighbours are computed for each row of data. The class that is found to have the majority vote among the neighbours is assigned to the selected transaction in the combination set. This way KNN is applied to all rows of transactions and a new class vector is thus obtained for all rows of transaction in the selected combination.

Fitness function evaluation is done by finding MSE (Mean Squared Error). This is done by comparing the newly obtained class vector to the already available class vector.

MSE (Mean Squared Error) = $1 \div n \, \Sigma \, (y' - y)^2$

Where y' refers to the new obtained class vector and y refers to the already available class vector. This process is done for all other combinations in the random population. Fitness values for each combination are found out. The combination with lower fitness values are passed to the next generation. The new set of chromosomes is found by applying crossover and mutation.

For instance: 1 1 0 1 0 0 1 0 is the chromosome it denotes the attributes 1 2 4 and 7 are selected. Then the columns in the training dataset according to these values are taken. KNN is applied to it. Euclidean distance between each row of transaction in the combination set of data.

Nearest three neighbours are taken and the class corresponding to the majority is assigned to the data. This is done for all rows in the combination set and obtains a class vector of $y^{'}$. MSE is calculated for $y^{'}$ and the already available class vector y. Fitness score is hence obtained. Fitness scores for set of combinations are obtained in a similar way and the lower fitness value combinations are passed to the next generation to perform mutation and crossover. Tournament selection method is used here to select the chromosomes due to its simplicity, speed and efficiency. This selection method makes sure that no worst individuals passed to the next generation.

Crossover:

Uniform crossover is applied in which the random bits of parents are selected. This way child is created. For example:

Suppose parents selected are p1 and p2

P1: 1 2 5 6 8 11001101
P2: 1 2 4 5 6 11011100
Applying crossover
1 1 0 0 1 1 0 1
1 1 0 1 1 1 0 0
Result: 1 1 0 1 1 1 0 1 (1 2 4 5 6 8)

P1 and P2 of combinations 1 2 5 6 8 and 1 2 4 5 6 are selected which are then subjected to uniform crossover where the random bits of p1 and p2 are exchanged to form the resultant output as 1 1 0 1 1 1 0 1 which is of the combination 1 2 4 5 6 8.
Mutation:
The random bits of the crossover output are changed in some position to form the mutated output.

7. CLASSIFICATION USING SUPPORT VECTOR MACHINE

A SVM classifier classifies data by finding the best hyperplane that separates all data points of one class from those of the other class. The best hyperplane for an SVM means the one with the largest margin between the two classes. Margin means the maximal width of the slab parallel to the hyperplane that has no interior data points. The support vectors are the data points that are closest to the separating hyperplane; these points are on the boundary of the slab. As with any supervised learning model, first train a support vector machine, and then cross validate the classifier. Use the trained machine to classify (predict) new data. In addition, to obtain satisfactory predictive accuracy, various SVM kernel functions are used, and must tune the parameters of the kernel functions. SVM is trained with the featured set of indices and the class vector. Train data refers to the data to be given for training which the transactions of the dataset are and the class information refers to the class type i.e. legal or fraud. Training is the process of taking content that is known to belong to specified classes and creating a classifier on the basis of that known content. The training is performed by giving the input arguments. The input is given as training data with each row corresponds to the transaction and each column represents the attributes. The next argument is the grouping variable which refers to the class vector indicating legal or fraud. To map the data to feature space for finding hyperplanes the kernel functions are must. There are many kernel functions are available including linear, Gaussian and radial basis function. The default one is linear and in the proposed work linear kernel has been used as the dataset attributes are of small and it can faster do the training compared to nonlinear kernels. As the proposed work consists of only 2 classes (legal/fraud) it is best to use linear kernels for linearly separable data. Linear kernel finds the dot product so as to find the maximal hyperplane for separating the data. Training is the process of taking content that is known to belong to specified classes and creating a classifier on the basis of that known content. The training is performed here by giving the input arguments. The input is given as training data with each row corresponds to the transaction and each column represents the attributes. The next argument is the grouping variable which refers to the class vector indicating legal or fraud. To map the data to feature space for finding hyperplanes the kernel functions are must.
Svmtrain() function is used for training. The function takes the inputs as train data and the class. An SVM classifier is built and this classifier is then used for classifying the new incoming data. Svmclassify() function is used for classifying the test data. The function takes the input as the classifier and the test data.

8. FEATURES AND PRE-PROCESSING

There are about 120 features per snapshot in the credit bureau data, resulting in 120 x 3 = 360 features; and 60 features per snapshot in the customer experience data, resulting in 60 x 3 = 180 features; and 300 features in the consumer purchase behaviour data. Therefore, there are a total of 840 features in the raw training/test samples.
Since there are many available features in the samples, a decision is performed for variables selection before training the models. As there are three sources of dataset, a choice is made to conduct variable selection separately. For credit bureau data, only current snapshot features (120 features) in variables is selected while creating additional trend features is described in the next section. For customer experience data, a summary of statement-wise features, either taking the max (indicator features) or mean (continuous features) is taken to reduce the features from 180 to 60 before the variable selection process.
For quick variable selection, the stepwise redresser function in software R is used. Also, to create an additional random variable (uniformly distributed) in the training sample and include it in variable selection stepwise regression function can be used. In the stepwise output, the variables are ranked by their predictive importance to the target. To determine the variable selected, 2 rules are applied: First, for credit bureau variables, one aims to select at most top 50 variables; while for purchase behaviour and customer experience data, one aims to select at most top 25 variables. Based on this understanding, bureau variables generally have stronger predictive power to risks than the other two types of data, thus the decision to keep more variables for bureau data. Secondly, only variables that are ranked above the random variable will be kept in the selected variables list. It is believed that any of the variables that are ranked under the random variables are more likely random noises. By applying rules above, a selection of a total of 81 features for model training purposes is done: 50 features from bureau data, 18 features from transaction and 13 features from digital data.

8.1. Additional features creation

As described above, for bureau data, one only chooses the current snapshot features into the variable selection; however, one can also think the trend of some features may have incremental values to predict risks, for example, the outstanding utilization trend (outstanding balance divided by

credit line), the payment ratio trend (payment divided by last statement outstanding balance), the FICO trend and etc. One can also create some indicators based on previous statements information such as if an account has made purchase or not, or if an account has made any payment or not and etc. As a result, a total of additional 12 features were created.
Therefore, there are a total of 93 features in the dataset for training purposes.

8.2. Feature treatment

One can apply simple missing treatment and capping/flooring to the features selected. If a feature has missing values, the missing observations will be imputed by medians; and the feature is capped by its 99th percentile and floored at 1st percentile.

9. TECHNIQUES AND ALGORITHMS

9.1. Logistic Regression

In statistics, logistic regression, or logit regression, or logit model is a regression model where the dependent variable (DV) is categorical. This article covers the case of binary dependent variables—that is, where it can take only two values, such as pass/fail or win/lose. Cases with more than two categories are referred to as multinomial logistic regression, or, if the multiple categories are ordered, they are referred as ordinal logistic regression.

The binary logistic model is used to estimate the probability of a binary response based on one or more predictor (or independent) variables (features). As such it is not a classification method. It could be called a qualitative response/discrete choice model in the terminology of economics.

Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution. Thus, it treats the same set of problems as probit regression using similar techniques, with the latter using a cumulative normal distribution curve instead. Equivalently, in the latent variable interpretations of these two methods, the first assumes a standard logistic distribution of errors and the second a standard normal distribution of errors.

Logistic regression can be seen as a special case of generalized linear model and thus analogous to linear regression. The model of logistic regression, however, is based on quite different assumptions (about the relationship between dependent and independent variables) from those of linear regression. In particular the key differences of these two models can be seen in the following two features of logistic regression. First, the conditional distribution (y|x) is a Bernoulli distribution rather than a Gaussian distribution, because the dependent variable is binary. Second, the predicted values are probabilities and are therefore restricted to (0.1) through the logistic distribution function because logistic regression predicts the probability of particular outcomes.

Logistic regression is an alternative to Fisher's 1936 classification method, linear discriminant analysis. If the assumptions of linear discriminant analysis hold, application of Bayes' rule to reverse the conditioning results in the logistic model, so if linear discriminant assumptions are true, logistic regression assumptions must hold. The converse is not true, so the logistic model has fewer assumptions than discriminant analysis and makes no assumption on the distribution of the independent variables.

9.2. Random Forest

A random forest is an ensemble (i.e., a collection) of unpruned decision trees. Random forests are often used when a very large training dataset and a very large number of input variables (hundreds or even thousands of input variables) is present. A random forest model is a classifier that consists of many decision trees and outputs the class that is the mode of the class output by individual trees [7].

Each tree is constructed using the following algorithm:

1. Let the number of training cases be N, and the number of variables in the classifier be M.
2. A told the number 'm' input variables to be used to determines the decision at a node of the tree; thus, m should be much less than M.
3. Choose a training set for this tree by choosing N times with replacement from all N available training cases (i.e. take a bootstrap sample). Use the rest of the cases to estimate the error of the tree, by predicting their classes.
4. For each node in the tree, randomly choose m variables on which to base the decision at that node. Calculate the best split based on these m variables in the training set.
5. Each tree is fully grown and not pruned (as may be done in constructing a normal tree classifier).

The advantages of random forests are:

1. For many data sets, it produces a highly accurate classifier.
2. It handles a very large number of input variables.
3. It can balance error in the class population of unbalanced data sets.
4. It computes proximities between cases, useful for clustering, detecting outliers, and (by scaling) visualizing the data.
5. Using the above, it can be extended to unlabelled data, leading to unsupervised clustering, outlier detection and data views.
6. Learning is fast.

9.3. Gradient Descent

Gradient descent is based on the observation that if the multivariable function $F(X)$ is defined and differentiable in a neighbourhood of a point **'a'**, then $F(X)$ decreases fastest if one goes from **'a'** in the direction of the negative gradient of F at **'a'**, $-\nabla F(a)$. It follows that, if gradient descent is based on the observation that if the multivariable function $F(X)$ is defined and differentiable in a neighbourhood of a point **'a'**, then $F(X)$ decreases fastest if one goes from **'a'** in the direction of the negative gradient of F at **'a'**, $-\nabla F(a)$. It follows that, if

$$b = a - \gamma \nabla F(a)$$

for γ small enough, then $F(a) \geq F(b)$. With this observation in mind, one starts with a guess X_0 for a local minimum of F, and considers the sequence X_0, X_1, X_2…. such that

$$X_{n+1} = X_n - \gamma_n \nabla F(X_n).$$

Since, $F(X_1) \geq F(X_2) \geq \ldots\ldots$, so hopefully the sequence (Xn) converges to the desired local minimum. Note that the value of the step size γ is allowed to change at every iteration. With certain assumptions on the function *F* (for example, *F* convex and $-\nabla F$ Lipschitz) and particular choices of γ (e.g., chosen via a line search that satisfies the Wolfe conditions), convergence to a local minimum can be guaranteed. When the function *F* is convex, all local minima are also global minima, so in this case gradient descent can converge to the global solution.

This process is illustrated in the picture to the right. Here *F* is assumed to be defined on the plane, and that its graph has a bowl shape. The blue curves are the contour lines, that is, the regions on which the value of *F* is constant. A red arrow originating at a point shows the direction of the negative gradient at that point. Note that the (negative) gradient at a point is orthogonal to the contour line going through that point. One sees that gradient descent leads to the bottom of the bowl, that is, to the point where the value of the function *F* is minimal.

9.4. Support vector machines

In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyse data and recognize patterns, used for classification and regression analysis. Given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

In addition to performing linear classification, SVMs can efficiently perform a nonlinear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. [9]

10. RESULTS AND CONCLUSION

10.1. Fraud Detection

The experiment was carried out in Intel Core i7 processor with 8GB RAM implemented in R. In order to evaluate the proposed model, UCSD-FICO Data mining contest 2009 data set is used. The competition was organized by FICO, the leading provider of analytics and decision management technology, and the University of California, San Diego (UCSD). The dataset is a real dataset of e-commerce transactions and the objective was to detect anomalous e-commerce transactions. The train dataset contains 100000 transactions and the test data consists of 50,000 transactions. The dataset contains 19 fields including class labels—amount, hour1, state1, zip1, custAttr1, field1, custAttr2, field2, hour2, flag1, total, field3, field4, indicator1, indicator2, flag2, flag3, flag4, flag5, and Class. It is found that custAttr1 is the account/card number and custAttr2 is e-mail id of the customer. Both these fields are unique to a particular customer and thus decided to keep only custAttr1. The fields total and amount as well as hour1 and hour2 are found to be the same for each customer and thus removed total and hour2. Similarly, State1 and zip1 are also found to be representing the same information and thus removed state1. All other fields are anonymized and therefore decided to keep them as they are. Thus, the final dataset contains 15 fields—amount, hour1, zip1, custAttr1, field1, field2, flag1, field3, field4, indicator1, indicator2, flag2, flag3, flag4, flag5. Data pre-processing is done on all these fifteen attributes of both training s well test set. The pre-processed datasets are then clustered to get the label class and this results clusters as 19,605 of class1 and 11,670 of class 2 in which class2 population is taken as fraud. After clustering the feature selection using genetic algorithm based K Nearest Neighbour approach has been done. [11]

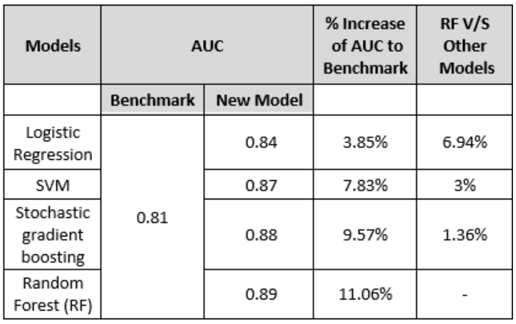
Feature selection using genetic algorithm is done in order to select the significant combination of features that contributes most to the class. 8 features of the dataset are selected and the random combinations of it are made. KNN is applied to the initial population, the original train data and the already available class vector. The fitness values are also obtained corresponding to each combination by finding it mean squared error. At each generation, the best chromosomes are passed to the next generation by applying genetic operators. The genetic algorithm terminates as it reaches the maximum number of generations. The best and mean values should be close to each other for accurate genetic algorithm result. The best fitness shows the best fitness value that the chromosome should have and the mean fitness refers to the average of mean fitness values of all generations.

10.2. Classification

In the research, detection of credit card fraud mechanism was presented to find and examine the result based on the principles of the mentioned algorithm. In the research hybrid algorithm has been used to execute credit card fraud detection and evaluate how credit card fraud impact on financial institution as well as merchant and customer, fraud detection technique by hybrid algorithm. The Genetic algorithms are evolutionary algorithms in which the aim is to obtain the better and optimal solutions. In the study fraud detected and fraud transactions are generated with the given sample data set. If this algorithm is applied into bank credit card fraud detection system, the probability of fraud transactions can be predicted soon after credit card transactions by the banks. And a series of anti-fraud strategies can be adopted to prevent banks from great losses before and reduce risks.

One can calculate areas under ROC curve on the test sample after four models are trained. The AUCs from benchmark and four models are provided in table. Since random forests method generates the largest incremental values to the benchmark. Also, ROCs are as shown in Figure 4.

Table 1: The AUCs of benchmark model and new models



Models	AUC		% Increase of AUC to Benchmark	RF V/S Other Models
	Benchmark	New Model		
Logistic Regression	0.81	0.84	3.85%	6.94%
SVM		0.87	7.83%	3%
Stochastic gradient boosting		0.88	9.57%	1.36%
Random Forest (RF)		0.89	11.06%	-

The obtained results for the four models are shown in Table 1 in terms of AUC metric computed on the test sample. The best result is achieved by the random forest model which outperforms logistic regression (improvement of 6.94%), SVM (improvement of 3.00%) and stochastic gradient boosting method (improvement of 1.36%), while all models can beat the benchmark model. It is expected that all models could beat the benchmark, since the benchmark model only applied the logistic regression on credit bureau data. The new logistic regression model has smallest incremental value in terms of AUC to the benchmark, while SVM, stochastic gradient boosting and random forest have substantial improvement to the benchmark.

One can also check the top 10 variables that have most predictive power to forecast the likelihood to be charged-off. They follow into three categories: balance utilization, FICO and the carried total/highest balance, all of which are business intuitive. For example, the higher utilization indicates the higher risk to be "bad" account; high FICO score means the low risk population, and carrying higher balance explains the customers tend to revolve their balance and are lack of capabilities to pay off and thus are riskier.

Although the sophisticated machine learning methods can beat the benchmark (logistic regression) in terms of model performance, it may give rise to a problem with respect to model implementation and interpretability. The credit card industry is highly regulated by OCC (Office of the Comptroller of the Currency), so companies are required to provide transparency of model structure and interpretation to regulators. However, random forest and stochastic gradient boosting, as ensemble methods, are quite complicated as trees are ensemble together and thus encounter difficulties for interpretation. In addition, their implementation into the internal production system could also be challenging.

11. FUTURE

Although, machine learning techniques to train better models can outperform the benchmark model, however, considering the difficulties of model interpretability and implementation, in the future, more research needs to be done on:

a. Find creative solutions to provide enough transparency of ensemble methods to regulators & enable implementation in the production system.

b. Improve logistic regression under the framework of gradient boosting methods so that the trained model can have both improved model performance and clear interpretation.

Since the existing models lacked high accuracy due to unavailability of publicly available datasets and therefore a hybrid approach is used for detecting frauds on unsegregated dataset of UCSD dataset in the form of a hybrid approach involving genetic algorithm and support vector machines. Since the proposed model, UCSD-FICO Data mining contest 2009 data set is used. The proposed fraud detection model are to be evaluated using anonymised dataset and able to handle class imbalance. The proposed model involves the stages like pre-processing, clustering, feature selection using genetic algorithm and finally support vector machine classification. These stages are successfully implemented to the dataset and created a good model for detecting fraud. SVM shows good accuracy in the proposed approach by classifying the test data to fraud and legal respectively.

Risk management is critical for a credit card company to survive in such competing industry. In addition to operational expenses, provisional loss is a major driver to a company's expense. The provisional loss arises due to the "bad" accounts booked – bank lends the money to customers who eventually do not have capability to pay back. In the risk management, there are generally two stages a company can take to manage and control credit risks. The first stage occurs when booking a customer. An aggressive underwriting strategy could book and approve high risk population who seeks for credit card; while a conservative policy may only focus on upmarket and affluent population. As expected, the first strategy could generate both high revenues (interests charged) and high expenses due to bad accounts booked, resulting in trivial incremental net income; and the second strategy could generate not only low revenues but also low losses, resulting in incremental net income be trivial as well. There is always trade-off for different strategies in terms revenue generation and loss control. Finding an optimal strategy is often difficult and needs to be adjusted accordingly due to internal or external factors such as macroeconomic change, for example, almost all credit companies suppressed their approval rates for high risk segments of population and incurred huge financial loss during 2008/2009 economic depression. The second stage happens in customer management after the customer is booked. Although booked customers pass the first screen of risk control, the chance of false negative (false "good" accounts) could still be high. However, in the second stage, by leveraging their performance such as credit card utilization, payment information, risks can further be managed to control provisional loss. In the research, one focuses on the second stage of risk management, and also how one can classify a booked account that is supposed to be a "bad" account within 12 months of being booked. Since an internal classification model is already available, thus a secondary interest is to train a better classifier to outperform the benchmark model.

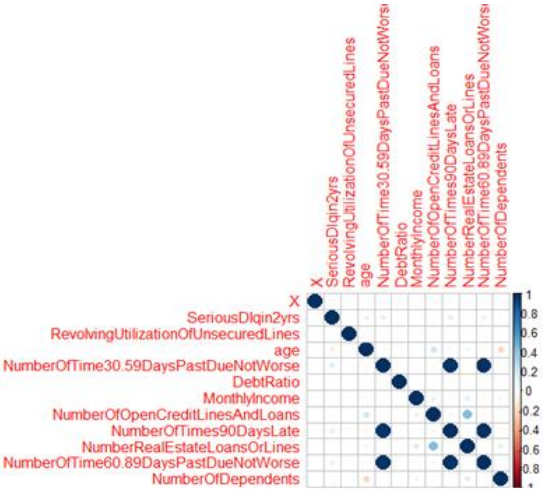


Fig 5: Logit regression

12. REFERENCES

[1] S.Benson Edwin Raj, A. Annie Portia, ―Analysis on Credit Card Fraud Detection Methods‖, IEEE International Conference on Computer, Communication and Electrical Technology, IEEEMarch 2011.

[2] M.Hamdi Ozcelik, Mine Isik, ―Improving a credit card fraud detection system using Genetic algorithm‖, IEEE International Conference on Networking and Information Technology, IEEE2010.

[3] Genetic algorithms for credit card fraud detection by Daniel Garner, IEEE Transactions May 2011.

[4] Research on credit card fraud detection model based on dis-tance sum IEEE 2009 International Joint Conference on Artificial Intelligence.

[5] Credit card fraud detection using neural network, Raghavedra Patidar, Lokesh Sharma, ISSN: 2231-2307, Volume, Issue-NCAI211, JUNE2011.

[6] Bradley, P. A. (1997). The use of the area under the ROC curve in the evaluation of machinelearning algorithms. Pattern Recognition 30(7): 1145-1159.

[7] Breiman, L. (2001). Random forest. Machine Learning 45 (1): 5-32.

[8] Cortes, C. and C. Vapnik. (1995). Support vector networks. Machine Learning 20 (3): 273-297.

[9] Friedman, J. H. (1999). Stochastic gradient boosting. Technical report, Dept. of Statistics, Stanford University.

[10] McCullah, P. and J. A. Nelder. (1989). Generalized linear models. Second edition. London:Chapman and Hall.

[11] R Core Team. (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at http://www.R-project.org/.

[12] Salford Systems. Available at http://www.salford-systems.com/products/spm.

[13] Siddiqi, N. (2006). Credit risk scorecards: developing and implementing intelligent credit scoring. John Wiley & Sons, Inc. Hoboken, NJ.

[14] Thomas, L. C., D. B. Edelman, and J. N. Crook. (2002). Credit scoring and its applications.SIAM. Philadelphia, PA.

[15] Bentley, P., Kim, J., Jung. G. & J Choi. 2000. Fuzzy Darwinian Detection of Credit Card Fraud, Proc. of 14th Annual Fall Symposium of the Korean Information Processing Society.

[16] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules in large databases,” in Proceedings of the 20th International Conference on Very Large DataBases, pp.487–499, 1994.

[17] Min Pei, Erik D Goodman, William F Punch and Ying Ding, “Genetic Algorithms for Classification and Feature Extraction”.