IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727
PP 24-30
www.iosrjournals.org

Designing Hybrid Model for Fraud Detection in Insurance

[1]Ms Janhavi Naik, [2]Dr J.A Laxminarayana
[1](Computer Department, Goa College of Engineering, India)
[2](Computer Department, Goa College of Engineering, India)

Abstract: *Fraud is increasing dramatically with the expansion of all modern technology and global communication systems, resulting in extreme loss to the businesses. Hence, fraud detection has become an important issue to be explored. Fraud detection involves identifying fraud as quickly as possible once it has been perpetrated. Insurance fraud is a criminal offence where one engages in the illicit act of intentionally falsifying a claim in an attempt to defraud an insurance company with the expectation of receiving income when not entitled to payment from the benefits being sought. The hybrid model proposed in this paper uses K-means clustering and SVM classification. K-means clustering (unsupervised technique) provides a simple and easy way to classify a given dataset. Support Vector Machines (supervised machine learning algorithm) is used for classification. Both advantages of these supervised and unsupervised techniques are utilized to detect the fraud. The proposed hybrid model enables data analysts to explore large databases quickly and efficiently.*
Keywords: *Fraud Detection, Hybrid Model, Kmeans clustering, SPSS, Support Vector Machines.*

I. Introduction

Insurance fraud spreads quickly in the domestic and foreign field. Hence we need more efficient and accurate technology to insurance fraud. Fraud involves intentional deception or misrepresentation intended to result in an unauthorized benefit. It is shocking because the incidence of insurance fraud keeps increasing every year. In order to detect and avoid the fraud, data mining techniques are applied. This includes some preliminary knowledge of Insurance system and its fraudulent behaviors, analysis of the characteristics of insurance data.

Data mining which is divided into two learning techniques supervised and unsupervised is employed to detect fraudulent claims. But, since each of the above techniques has its own set of advantages and disadvantages, by combining the advantages of both the techniques, a hybrid approach for detecting fraudulent claims in insurance industry is proposed. This technique will have low time complexity, high recognition rate and high accuracy for benefit of better future.

The Association of Certified Fraud Examiners (ACFE) in paper [5] defined fraud as "the use of one's occupation for personal enrichment through the deliberate misuse or application of the employing organization's resources or assets. In the technological systems, fraudulent activities have occurred in many areas of daily life such as telecommunication networks, mobile communications, on-line banking, and Ecommerce.

The development of new fraud detection methods is made more difficult due to the severe limitation of the exchange of ideas in fraud detection. Data sets are not made available and results are often not disclosed to the public. According to a recent survey in paper [1], it is estimated that the number of false claims is greater when considered to total claims in industry. So, to make insurance industry free from fraud, it is necessary to focus on elimination or minimization of fake claims arriving through insurance.

II. Literature Survey

2.1 Data Mining Approaches

The below mentioned are some of approaches applied for different fraud in insurance systems, credit card fraud and telecommunication frauds.

- **Anomaly Detection:**

It calculates the probability of each claim to be fraudulent by examining the previous insurance claims. The analysts further investigate the cases that have been flagged by data mining model.

- **K-Means Algorithm:**

It takes the parameter k as input, and divides a set of n objects into k clusters such that the resulting intra-cluster similarity is high while the inter-cluster similarity is low. It predefines the number of clusters. This becomes the drawback for clustering new incoming objects since there would be fixed number of clusters.

- **Outlier Detection:**

A baseline of the unusual behavior of usage of service for policyholder is established. Any deviation from this baseline indicates an outlier. It generally results from clustering.

- **Support Vector Machines:**

SVM is fundamentally a classification technique. The system is trained to determine a decision boundary between classes of "legitimate" and "fraudulent" claims. Then each claim is compared with that decision boundary and is placed into either of two classes.

- **Non-Negative Matrix Factorization:**

It is a technique for clustering particular attribute into several clusters according to usage by different policyholder.

2.2 Cluster Analysis

Clustering is an unsupervised learning method unlike the classification method which is generally viewed as a supervised learning technique. It is the process of making a group of abstract objects into classes of similar objects. As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to observe characteristics of each cluster. Suppose we are given a database of 'n' objects and the partitioning method constructs 'k' partition of data. Each partition will represent a cluster and $k \leq n$. It means that it will classify the data into k groups, which satisfy the following requirements. One is that each group contains at least one object and second is that each object must belong to exactly one group.

III. Proposed work

3.1 Problem Definition

Given a set of records, the main problem is to correctly classify the fraudulent and legitimate records. Hence when supervised techniques are used, they have training dataset where it cannot classify claims of policyholder whether true or false correctly. In case of unsupervised techniques it doesn't have class labels and training set, hence correct classification is not achieved. Hence both techniques suffered disadvantages in their methodology.

3.2 Proposed Approach

3.2.1 Supervised Learning:

Supervised Methods uses pre-defined class labels. In the context of insurance fraud detection the class labels are "legitimate" and "fraudulent" claims. The training dataset is used to build the proposed model. Any new claim can be compared with the already trained model to predict its class.

3.2.2 Unsupervised Learning:

Unsupervised learning has no class labels. It focuses on finding those instances which show unusual behavior. It can discover both old and new types of fraud since they are not restricted to the fraud patterns which already have pre-defined class labels like supervised learning techniques do.

3.2.3 Hybrid Model :

The proposed hybrid model for detecting insurance frauds is built using both supervised and unsupervised techniques. For this, chosen methods are as follows:-

1. K-Means Clustering Method
2. Support Vector Machine (SVM).

IV. Detailed methodology

4.1 K-means Clustering

It is one of the simplest unsupervised learning algorithms for clustering problem. It follows a way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster.

These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as bar centers of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.

The algorithm is stated as follows. Our starting point is a large set of data entries and a ***k,*** defining the number of centers.

1 – The first step is to choose randomly ***k*** of our points as partition centers.

2 – Next, we compute the distance between every data point on the set and those centers and store that information.
3 – Supported by the last step calculations, we assign each point to the nearest cluster center. Thus, we get the minimum distance calculated for each point, and we add that point to the specific partition set.
4 – Update cluster center positions.
5 – If the cluster centers change, repeat the process from **2**.

4.2 Support Vector Machines

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.

Support Vectors are simply the co-ordinates of individual observation. Support Vector Machine is a frontier which best segregates the two classes (hyper-plane/ line). For this need is to remember a thumb rule to identify the right hyper-plane: "Select the hyper-plane which segregates the two classes better". SVM has a feature to ignore outliers and find the hyper-plane that has maximum margin. Hence, we can say, SVM is robust to outliers. In SVM, it is easy to have a linear hyper-plane between these two classes. For adding this feature, SVM has a technique called the kernel trick. These are functions which takes low dimensional input space and transform it to a higher dimensional space i.e. it converts not separable problem to separable problem, these functions are called kernels. It is mostly useful in non-linear separation problem. It does some extremely complex data transformations and works really well with clear margin of separation and it is effective in high dimensional spaces.

It is effective in cases where number of dimensions is greater than the number of samples. It uses a subset of training points in the decision function (called support vectors), so it is also memory efficient. The classifier that is created by this module is useful for predicting between two possible outcomes that depend on continuous or categorical predictor variables.
For working of SVM there are two steps which designed.

1) Training (Preprocessing Step):

• Define two class labels as legitimate or fraudulent.
• Classify claims into two classes using the training data set.
• Choose support vectors and find the maximum marginal hyper plane that separates the claims into two classes.

2) Classification:

• Identify the new incoming claims into either legitimate or fraudulent class.

Given a set of training examples labeled as belonging to one of two classes, the SVM algorithm assigns new examples into one category or the other. The examples are represented as points in space, and they are 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.

The feature space that contains the training examples is sometimes called a hyperplane, and it may have many dimensions. Support vector machines are among the earliest of machine learning algorithms, and SVM models have been used in many applications, from information retrieval to text and image classification.

V. Hybrid Model

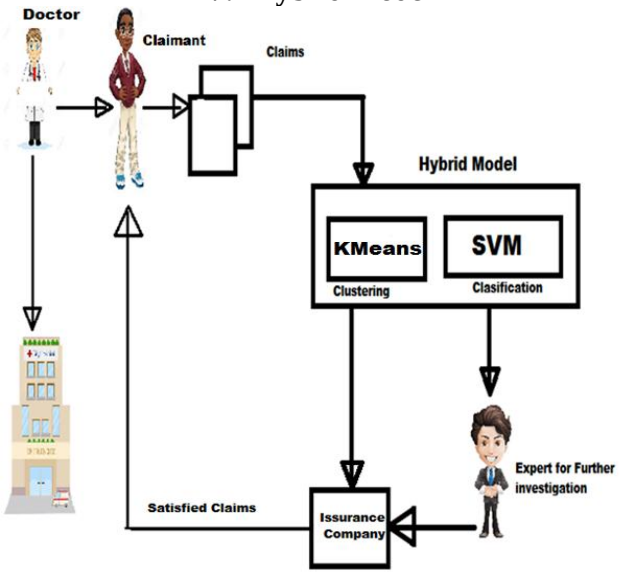

Figure 5: Block Diagram for Hybrid Model Designed

Classes are meaningful to humans and it can be easily used for pattern classification and hence A claim will be classified as a legitimate claim if it follows a similar pattern to the legitimate behavior else it will be classified as an illegitimate.

Advantage provided by unsupervised technique is that it aims to detect anything which does not abide by the normal behavior. Lack of direction, hence it can find patterns that have not been noticed previously. Hence by combining the advantages of both the techniques, a hybrid approach for detecting fraudulent claims in an insurance industry is proposed.

VI. Implementation

6.1 Construction of Dataset

Insurance Data is collected from well known reputated company containing the details of policy holders. Using the data, detailed simplified dataset is manually prepared by careful selection of attributes. A proper dataset consisting of 1000 records is been constructed.

Data preprocessing is done, by considering only the attributes necessary for fraud detection. Once preprocessed data is available, further step is conversion of text data to numeric data. First step is clustering process, hence for Kmeans clustering the dataset (numeric dataset) will be given as the input. Hence for conversion, SPSS tool is used.

	policy no	App no	Annual P	Naac	Mode	Product	Reg /Sing	own name	adv code	Prev Med	Health	Approved	Reason	Fraud
2	4070594	A3487527	12004	4201	MONTHLY	DEP	REG	PURSHOT	AV9045	NO	fine	YES	complete	no
3	4071921	A3487520	3998	1000	HALF Y	TRP	REG	DINESH	AV9045	NO	good	yes	complete	no
4	4081241	A3487529	50000	4000	YEARLY	PPR	REG	MENINO	55359	YES	ok	yes	dealth	no
5	4083421	A3487252	12046	4216	MONTHLY	DEP	REG	ARUN	AL6207	NO	very good	no	Health P	yes
6	4118269	C3831908	6072	122	MONTHLY	BEP	SINGLE	RINA	AX6056	NO	fine	YES	complete	no
7	4121584	C3831908	6000	100	MONTHLY	BEP	SINGLE	LINGARAJ	AX6056	NO	good	YES	financial	no
8	4243981	A3487253	2012	4204	MONTHLY	DEP	REG	MATHAI	AV9045	YES	good	no	complete	yes
9	4268156	C3832425	6801	2040	MONTHLY	BEP	SINGLE	AKSHAY	AV9045	NO	good	YES	complete	no
10	4276623	A3788163	13828	1383	MONTHLY	SLH	REG	ALKA	AX6013	NO	very good	no	Health P	yes
11	4285737	A4033857	12176	4262	HALF Y	DRC	REG	GANGASIF	AZ2682	NO	ok	YES	financial	no
12	4294180	A4033858	12043	4095	MONTHLY	DRC	REG	BANI	AZ2682	NO	very good	no	Health P	yes
13	4293062	A4033857	18047	6316	MONTHLY	DEP	REG	VIJAY	AV9045	YES	very good	no	Health P	yes
14	4303440	A4033842	12026	4209	MONTHLY	DEP	REG	SANKET	AY0614	NO	fine	YES	complete	no
15	4306857	A4033857	34452	9647	QUARTER	DRT	REG	SATISH	AX5764	YES	very good	no	complete	yes
16	4310995	A4033859	12109	3875	MONTHLY	DRC	REG	IJAJUDEEN	AZ2682	NO	good	YES	complete	no
17	4315578	A4035151	12521	4007	MONTHLY	DRC	REG	DEVIDAS	AZ2682	NO	good	no	Health P	yes
18	4325671	A4035151	12040	4214	MONTHLY	DEP	REG	RAHUL	AZ2682	NO	very good	no	Health P	yes
19	4329002	A3600987	12000	900	YEARLY	FRD	REG	SUPRIYA	AY0614	YES	ok	YES	complete	no
20	4336995	A3600987	50000	3750	YEARLY	FRD	REG	KAUSHAL	AY0614	NO	very good	no	Health P	yes
21	4341569	A4062809	40115	8023	YEARLY	DRT	REG	SMITA	AY0614	NO	good	YES	complete	no
22	4348442	A4063247	12117	4241	HALF Y	DEP	REG	JYOTI	AZ7432	NO	fine	YES	complete	no
23	433173	A3918600	12005	4202	HALF Y	DEP	REG	SAWANT	AV9045	NO	ok	YES	financial	no

Figure 6.1: Manually prepared Real time dataset

6.2 SPSS

SPSS (Statistical Package for the Social Sciences) has now been in development for more than thirty years. It provides extensive data management functions, along with a complex and powerful programming language. It uses both a graphical and a syntactical interface. It provides dozens of functions for managing, analyzing, and presenting data. The data used can be ranging from simple integers or binary variables to multiple response or logarithmic variables. It consists of three windows:

I. Data Editor
II. Viewer or Draft Viewer which displays the output files
III. Syntax Editor, which displays syntax files

Data View window, which displays data from the active file in spreadsheet format.
Variable View window, which displays metadata or information about the data in the active file, such as variable names and labels, value labels, formats, and missing value indicators.

6.3 Data conversion

The whole text data is converted to numeric values. Mostly data consists of categorical data, hence values are assigned for each value of different attribute so as to get a proper numeric dataset for Kmeans implementation. Later the same will be useful for support vector machines. This will prove efficient while dealing with classification.

	policyno	AnnualP	Naac	Mode	Product	RegSing	advcode	PrevMed	Health	Approved	Reason	Fraud
1	4070594.00	12004.0	4201.0	10	02	25	60	20	50	15	35	0
2	4071921.00	3998.0	1000.0	12	02	25	60	20	52	15	35	0
3	4081241.00	50000.0	4000.0	13	03	25	61	21	51	15	36	0
4	4083421.00	12046.0	4216.0	10	02	25	62	20	53	16	37	1
5	4118269.00	6072.0	122.0	10	03	30	63	20	50	15	35	0
6	4121584.00	6000.0	100.0	10	03	30	63	20	52	15	38	0
7	4243981.00	2012.0	4204.0	10	02	25	60	21	52	16	35	1
8	4268156.00	6801.0	2040.0	10	03	30	60	20	52	15	35	0
9	4276623.00	13828.0	1383.0	10	04	25	64	20	53	16	37	1
10	4285737.00	12176.0	4262.0	12	05	25	65	20	51	15	38	0
11	4294180.00	12043.0	4095.0	10	05	25	65	20	53	16	37	1
12	4293062.00	18047.0	6316.0	10	02	25	60	21	53	16	37	1
13	4303440.00	12026.0	4209.0	10	02	25	66	20	50	15	35	0
14	4306857.00	34452.0	9647.0	11	09	25	67	21	53	16	35	1
15	4310995.00	12109.0	3875.0	10	05	25	65	20	52	15	35	0
16	4315578.00	12521.0	4007.0	10	05	25	65	20	52	16	37	1
17	4325671.00	12040.0	4214.0	10	02	25	65	20	53	16	37	1
18	4329002.00	12000.0	900.0	13	04	25	66	21	51	15	35	0
19	4336995.00	50000.0	3750.0	13	04	25	66	20	53	16	37	1
20	4341569.00	40115.0	8023.0	13	05	25	66	20	52	15	35	0
21	4348442.00	12117.0	4241.0	12	02	25	68	20	50	15	35	0
22	433173.00	12005.0	4202.0	12	02	25	60	20	51	15	38	0
23	4345290.00	99116.0	19823.0	13	09	25	66	21	51	15	38	0

Figure 6.3: Numeric Dataset

6.4 Kmeans Clustering

In this scheme the input data is classified into specified number of groups. It is unsupervised learning approach used when there is no prior knowledge about particular class of observations in a dataset. This scheme classifies n data points into pre-specified k clusters. The data will be grouped into k-clusters according similarities among the cluster.

In first step k-clusters need to be defined. In Second step randomly centroid for each cluster will be choose. Centroid of particular cluster means mean value of that cluster. In third step distance of data from centroid of each cluster need to be computed.

Data will be grouped into particular cluster, according to minimum distance of data from centroid of cluster. Next time again centroid will be recomputed for each cluster because different values come in cluster. These steps will be repeated until there is no change in the output. Following is the flowchart designed for Kmeans clustering.

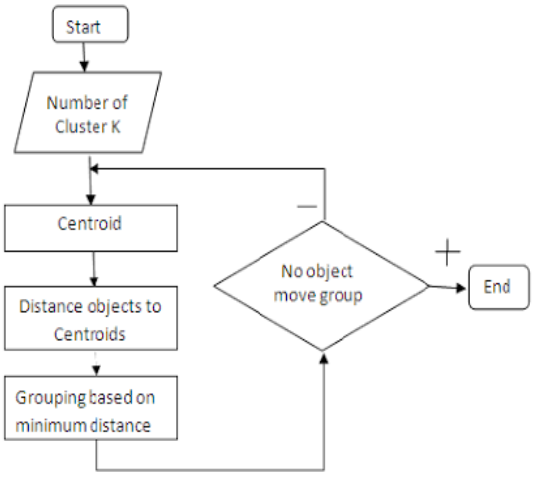

Figure 6.4: Flowchart for Kmeans Clustering

6.5 Support Vector Machines

Different linear SVM classifier on a 2D projection of the dataset is proposed. We only consider the features of this dataset. The objective of a Linear SVC (Support Vector Classifier) is to fit to the data you provide, returning a "best fit" hyperplane that divides, or categorizes, the data. From there, after getting the hyperplane, we then feed some features to our classifier to see what the "predicted" class is. This makes this specific algorithm rather suitable for our uses.

Different packages need to be installed like Numpy, Pandas, Sklearn, SVM and Matplot. Pandas package is required for data analysis. In the process of modeling logistic regression classifier, first we are going to load the dataset (CSV format) into pandas data frame and then we play around with the loaded dataset.

Numpy package is for performing the numerical calculation. Plotly package for visualizing the data set for better understanding. Matplotlib here is not truly necessary for Linear SVC. The reason why we're using it here is for the eventual data visualization. Sklearn package is for modeling the machine learning algorithms. Train test split method to split the dataset into the train and test dataset. Logistic Regression method is used for modeling the logistic regression classifier. Metrics method can be used for calculating the accuracy of the trained classifiers.

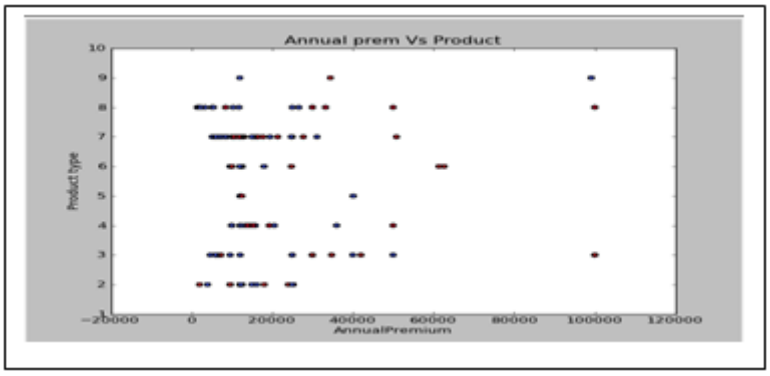

Figure 6.5: Features plotted of insurance dataset
(Product type VS Annual Premium)

VII. Conclusion And Future Work

Through all the papers referred during literature survey, Clustering techniques and Support Vector Machine Techniques were chosen for project. Kmeans Clustering method under unsupervised techniques is chosen for partitioning datasets. In supervised learning techniques, Support vector machines, is one of best classification tool that can be used for distinguishing datasets.

Kmeans Clustering algorithm is designed. Classifiers are built. Hence combining the advantages of both learning techniques, Supervised and unsupervised a hybrid model is designed. Dataset is built using policy holder records and attributes are chosen. Text to data conversion is done by studying SPSS tool.

Numeric dataset is prepared, so that it can be used for Kmeans clustering. Kmeans clustering algorithm is designed. Clusters are created for dataset. Also the same numeric dataset is applied for the support vector machines classification. Different features of the dataset are plotted against one another so as to get comparison and visualization of the data. These features give the exact picture of all numeric data and their target class. The classifier can be plotted on training dataset and thus later can be predicted for testing dataset.

Future work can be exploring the other methods in supervised and unsupervised learning techniques. Later the comparison can be done on different methods so as to get efficient methodology in wide range of techniques. Also instead of two methods, three methods could be used for future work.

References

[1] Song Chen , Aryya Gangopadhyay ,"Health Care Fraud Detection with Community Detection Algorithms", Information Systems University of Maryland Baltimore Country Email: song8@umbc.edu, Information Systems University of Maryland Baltimore Country Email: gangopad@umbc.edu.

[2] Chun Yan, "The Identification Algorithm and Model Construction of Automobile Insurance Fraud Based on Data Mining", College of Mathematics and Systems Science Shandong University of Science and Technology Qingdao of Shandong Province.

[3] Yufeng Kou, Chang-Tien Lu, Sirirat Sinvongwattana Yo-Ping Huang," Survey of Fraud Detection Techniques ",Dept. of Computer Science Virginia Polytechnic Institute and Engineering and State University Tatung University Falls Church, VA 22043, USA Taipei, Taiwan 10451 Email: yphuang@cse.ttu.edu.tw

[4] Hyun-Chul Kim, Shaoning Pang, Hong-Mo Je, Daijin Kim, Sung Yang Bang ,"Pattern Classification Using Support Vector Machine Ensemble",Department of Computer Science and Engineering, Pohang University of Science and Technology San 3 I , Hyoja-Dong, Nam-Gu, Pohang, 790-784, KOREA { grass,snpang,invu7 I ,dkim,sybang} @postech.ac.kr.

[5] J. Nagi, K. S. Yap, S. K. Tiong, Member, IEEE, S. K. Ahmed, Member, IEEE, A. M. Mohammad.,"Detection of Abnormalities and Electricity Theft using Genetic Support Vector Machines ".
[6] Jawad Nagi, Keem Siah Yap, Sieh Kiong Tiong, Member, IEEE, Syed Khaleel Ahmed, Member, IEEE, and Malik Mohamad ,"Nontechnical Loss Detection for Metered Customers in Power Utility Using Support Vector Machines".
[7] Yi Peng1, Gang Kou1, Alan Sabatka2, Zhengxin Chen1, Deepak Khazanchi1, Yong Shi1, "Application of Clustering Methods to Health Insurance Fraud Detection", Peter Kiewit Institute of Information Science, Technology & Engineering, University of Nebraska, Omaha, NE 68182, USA.
[8] Fabida A.1, Jasila E.K.2, "Fraud Detection in Health Insurance Using Expert Re-referencing" Department of Computer Science and Engineering , MES College of Engineering, Kuttippuram Kerala, India.
[9] Punam Devidas Bagul, Sachin Bojewar, Ankit Sanghavi , "Survey on Hybrid Approach for Fraud Detection in Health Insurance", Dept. of Computer Science and Engineering, ARMIET, Sapgaon, Mumbai University, India.
[10] Surbhi Agarwal, Santosh Upadhyay, "A Fast Fraud Detection Approach using Clustering Based Method", Mewar University, Chittorgarh Rajasthan Gr.Noida, Uttar Pradesh.
[11] ZHEXUE HUANG,"Extensions to the *k*-Means Algorithm for Clustering Large Data Sets with Categorical Values", CSIRO Mathematical and Information Sciences, Australia. Email: huang@mip.com.au.
[12] http://dataaspirant.com/2017/01/25/svm-classifier-implemenation-python-scikit-learn/
[13] http://scikit-learn.org/stable/tutorial/statistical_inference/supervised_learning.html
[14] http://scikit-learn.org/stable/datasets/
[15] https://docs.python.org/2/tutorial/modules.html
[16] https://pythonprogramming.net/linear-svc-example-scikit-learn-svm-python/
[17] http://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html
[18] http://opencv-python tutroals.readthedocs.io/en/latest/py_tutorials/py_ml/py_svm/py_svm_basics/py_svm_basics.html#svm-understanding