ISSN: 0976-3104
ISSUE: Computer and Allied Sciences

ARTICLE

INTELLIGENT FEATURE SELECTION WITH SOCIAL SPIDER OPTIMIZATION BASED ARTIFICIAL NEURAL NETWORK MODEL FOR CREDIT CARD FRAUD DETECTION

Gurumurthy Krishnamurthy Arun*, Kaliyappan Venkatachalapathy

Department of Computer and Information Science, Annamalai University, Chidambaram, TN, INDIA

ABSTRACT

In recent days, credit card fraud remains as an essential problem for theft and fraud commitment by the use of payment cards like credit or debit cards. For resolving this problem, financial industries have started to develop fraud detection algorithms. Data mining and machine learning (ML) approaches can be used to investigate the normal and abnormal patterns along with individual transactions in order to raise an alarm for possible frauds. In this view, this study develops an intelligent Feature Selection (FS) with social spider optimization (SSO) algorithm based artificial neural network (ANN) model called SSO-ANN for credit card fraud detection. The proposed SSO-ANN model involves preprocessing, ant colony optimization (ACO) algorithm based FS and SSO-ANN based classification. The proposed SSO-ANN model has the ability to detect the existence of frauds in credit card payments. The performance of the SSO-ANN model has been tested against two benchmark dataset namely German Credit dataset and Kaggle's Credit Card Fraud Detection dataset. The experimental outcome pointed out that the SSO-ANN model has shown superior results with the maximum classifier accuracy of 93.20% and 92.82% on the German Credit dataset and Kaggle's Credit Card Fraud Detection dataset.

KEY WORDS
Credit card fraud, Data mining, Classification, Machine learning, Feature Selection

INTRODUCTION

Generally, fraud is a mischief and criminal activity which is performed to gain the economical and personal applications. The fraud events can be reduced under the application of two approaches such as, Fraud prevention as well as Fraud detection. Initially, Fraud prevention is defined as proactive model that eliminates the occurrence of fraud. Secondly, fraud detection is required while an illegal transaction is made by a criminal. Specifically, credit card fraud is a general crime that happens often which is carried out by stealing the details of a person. The credit card transactions can be processed in a digital and physical manner [1].

With the improved credit card users, several crime actions were also improved. Even though there is massive number of authentic models, some of the mischief actions are still in progress which is highly complex to investigate. Using the internet, Fraudsters conceals the position and identity details. The credit card fraud has major impact on economic sector. The overall credit card fraud has attained staggering of USD $21.84 billion. The credit card loss, especially for merchants leads in major deviations such as losing the cost details, administrative charges, and so on [2]. As the merchants should tolerate the loss, and few products might have increased price, or offers and incentives were minimized. Hence, it is optional for loss reduction, and well-defined fraud detection model is applied to avoid the fraud cases. Diverse works are proposed on credit card fraud detection.

Received: 2 June 2020
Accepted: 10 July 2020
Published: 20 July 2020

ML and relevant approaches are utilized prominently such as ANN, rule-induction system, Decision Trees (DT), Logistic Regression (LR), and Support Vector Machines (SVM). Such technologies were applied either uniquely or by integrating various models for developing hybrid approaches. In this method, credit card fraud detection is processed using Random Forest (RF), SVM and LR. An Artificial Immune Recognition System (AIRS) is applied for credit card fraud detection as projected in [3]. AIRS is an extended version of reputed AIS scheme, in which negative selection is employed for reaching maximum precision. Hence, the accuracy is improved and minimizes the system response time to greater extent. A tailored Fisher Discriminant function has been utilized for detecting the credit card fraud [4]. The alteration of conventional functions makes highly sensitive instances.

In order to enhance the prediction of credit card frauds, an effective model is presented in [5]. A data set derived from a Turkish bank has been employed. Every transaction was measured as a fraudulent. The misclassification rates are limited with the help of Genetic Algorithm (GA) as well as scatter exploration. The newly presented technique is highly beneficial when compared with existing outcome. An alternate economical loss is a financial statement fraud. The methodologies like SVM, LR, Genetic Programming (GP) and Probabilistic NN (PNN) were applied to find financial statement fraud. A data set with 202 Chinese industries is employed. The t-statistic is applied for feature subset selection, in which 18 and 10 features are decided in 2 phases. The final outcome shows that the PNN performs quite well than GP. A fraud detection technique depends upon the user account's visualization and threshold-type examination as projected in [6]. Then, Self-Organizing Map (SOM) has been utilized as a visualization framework.

***Corresponding Author**
Email: arunnura2370@gmail.com

Hybrid approaches are the integration of several technologies. A hybrid scheme is composed of consisting of the Multilayer Perceptron (MLP), NN, SVM, LR, and Harmony Search (HS) optimization have been applied in [7] for detecting corporate tax evasion. HS is applicable in identifying optimal parameters for classification models. Under the application of data acquired from food and textile applications in Iran, MLP with HS optimization has attained maximum accuracy. The hybrid clustering mechanism with noise prediction ability was utilized in [8] for detecting fraud in lottery and internet games. The training data set undergoes compression over main memory at the time of increasing stored data-cubes. This model has reached maximum detection value with minimum false alarm rate.

The economic crisis can be handled using clustering and classifier ensemble methodologies which forms hybrid approaches finally [9]. The integration of SOM and k-means models are applied in clustering, whereas LR, whereas classification applies MLP, and DT approaches. The SOM and MLP classifier are said to be a remarkable combination, which produces standard accuracy. The concatenation of various models like RF, DR, Roush Set Theory (RST), and Back propagation NN (BPNN) were utilized [10] for developing fraud detection approach in case of corporate financial statements. Organizational financial statements were employed as data set. The final outcome with RF and RST is capable of reaching best classification accuracy.

The model which finds automobile insurance fraud as presented in [11]. A Principal Component Analysis (PCA) based RF method is combined with capable nearest neighbour technique. The traditional classical majority voting in RF has been substituted with effective nearest neighbour module. Overall data sets are employed in the experimental study. The PCA dependent method has generated a maximum classification accuracy and minimum variance, as related with RF and DT methodologies. The concatenation of GA and Fuzzy C-Means (FCM) is effectively applied [12] for exploring the fraud activity involved in automobile insurance sector. The sample records are divided into normal and suspicious classes according to the developed clusters. By removing the original and fraud records, the malicious cases are examined in future with the application of DT, SVM, MLP, and Group Method of Data Handling (GMDH). Hence, SVM performs well by reaching a resourceful specificity and sensitivity rates than other approaches.

The contribution of the study is given as follows. This study introduces a new FS with SSO algorithm based ANN model called SSO-ANN for credit card fraud detection. The proposed SSO-ANN model operates on three different stages namely preprocessing, ACO algorithm based FS(ACO-FS) and SSO-ANN based classification. The proposed SSO-ANN model has the capability of detecting the presence of frauds in credit card payments.

METHODS

The simulation outcome of the SSO-ANN model has been tested against two benchmark dataset namely German Credit dataset (https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)) and Kaggle's Credit Card Fraud Detection dataset (https://www.kaggle.com/mlg-ulb/creditcardfraud).

The block diagram of the proposed SSO-ANN model is demonstrated in Fig. 1. As shown, initially, the input data undergo pre-processing and converts the data into a compatible format. Then, ACO-FS algorithm is executed for selecting the useful count of features from the pre-processed data. Finally, SSO-ANN based classification process takes place to determine the existence of credit card frauds or not.

Preprocessing: Initially, preprocessing of input credit data takes place in two stages namely format conversion and data transformation. During format conversion, the input data in .csv format is converted into .arff format. Next, in the data transformation stage, the numerical values are converted into corresponding categorical values, i.e. values in 0's and 1's are converted into good and bad credits. Once the data is preprocessing, it is sent to the FS process for selecting the desired number of features.

ACO-FS model: In this section, ACO-FS model has been discussed the way of selecting features from the preprocessed data. The provided feature set of size n, the FS issues is mainly applied to find the least size of s feature subset (where s<n), at the time of retaining maximum accuracy while presenting the actual feature subset. An incomplete does not indicate the ordering among the features of solution. Simultaneously, the next feature has been selected and it is not affected by existing feature which appends the partial solution. Hence, there is no requirement equal size for FS problem [13]. The matching of FS problem to ACO method is comprised with following procedures:

- Graph depiction
- Heuristic popularity
- Pheromone extension
- Solution development

SSO-ANN based Classification Model: In this framework, the processes involved in the SSO-ANN based classification model have been discussed in the following subsections.

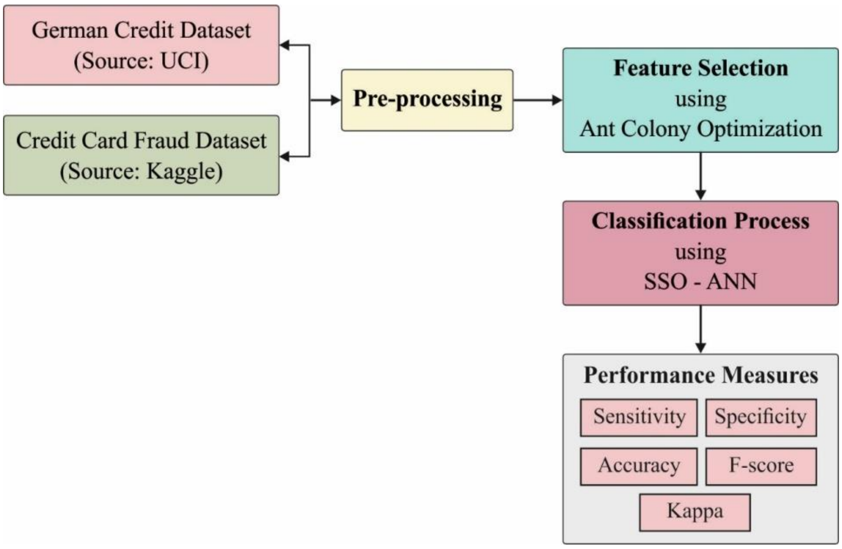

Fig. 1: Block diagram of SSO-ANN model

Structure of ANN: The ANN is defined as a soft computing model which is extensively applied for data processing. It has been evolved from biological nervous systems, like working function of human brain. It is illustrated with respect to weighted directed graphs where node is treated as artificial neurons as well as directed edges among neurons defined weights. It is classified into 2 classes namely, Feed forward and Recurrent networks. Initially, Feed forward networks are defined as static based while recurrent networks are dynamic. It generates a single set of resultant values rather producing a series of values from provided input. The Feedforward networks are generally memory-less and autonomous of existing network. When presenting a novel input pattern, the neuron outcomes are determined [14].

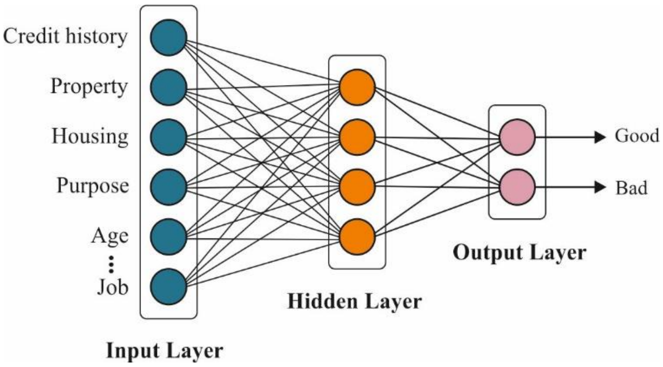

Fig. 2: Structure of ANN

Every neuron in the input and hidden layers are correlated with other neurons of subsequent layer. The neurons present in the hidden layers were utilized for calculating weighted sums of inputs and a threshold. The architecture of MLP includes input layer, hidden layer, and output layer, as depicted in Fig. 2. The input layer shows the parameters of datasets. The performance of hidden layer means the attributes of datasets that is not linearly separated while output layer offers the essential outcome. The final results show the function of neurons using a sigmoid activation function which is depicted by Eq. (1).

$$p_v = \sum_{u=1}^{n} w_{v,u} xu + \theta v, m_v = f_v(p_v) \qquad (1)$$

where p_v denotes linear integration of inputs $x_1, x_2, \ldots, x_n$, and thethresholdθ_v ,w_{vu} is the association weight among the input x_u, neuron v, and f_v represents activation function of vth neuron, and m_v means the final outcome. A sigmoid function is assumed to be an activation function as provided in Eq. (2).

$$f(z) = \frac{1}{1 + e^{-z}} \qquad (2)$$

The MLP can be trained using BP learning technique that is said to be a Gradient Descent (GD) model for accessing the weights. Every weight vector (w) was loaded using minimum random measures from

IIOAB JOURNAL ISSN: 0976-3104

pseudorandom sequence generator. Hence, it consumes maximum procedures for network training, and modified weights are processed at every step. The mentioned problems can be resolved by using SSO algorithm for computing the optimal value of weight and threshold functions, as SSO have the ability of calculating parallel weight and finds viable solutions.

SSO Algorithm: The SSO model depends upon the cooperative nature of social-spiders as projected by Cuevas [15].Here, search space is referred as communal web and spider's location is the best solution. The most important feature of social-spiders is a female-based population. Here, male spiders are minimum when compared with female in the overall community. The count of females N_f is selected randomly with least proportion$N$, which is determined as:

$$N_f = [(0.9 — rand * 0.25) * N], \qquad (3)$$

where $rand$implies random values among [0, 1]. The value of male spiders N_m is measured using:

$$N_m = N - N_f \, . \qquad (4)$$

All spiders gain a weight according to the fitness rate of attained solution:

$$w_u = \frac{fitness_u - worst}{best - worst}, \qquad (5)$$

where $fitness_u$ refers the fitness value accomplished by estimating u^tspider's position $u = 1, 2, \ldots, N$. The $worst$ and $best$ shows the inferior and superior fitness value of whole population, correspondingly.The communal web is mainly applied for transmission between the colony members. The data undergoes encoding as tiny vibrations are based on the weight as well as distance of spider which has been generated by the given expression:

$$Vb_{u,v} = w_v e^{-d_{u,v}^2}, \qquad (6)$$

where $d_{u,v}$ means the Euclidean Distance among the spider u and v. There are 3 kinds of relationships namely,

- Vibrations $Vb_{u,c}$ which are perceived by spider u by transmitting the data by the member c, and it is closer to u with maximum weight, such as $w_c > w_u$;
- Vibrations $Vb_{u,b}$ which are perceived by spider u which is forwarded by spider b with optimal weight of whole population; and
- Vibrations $Vb_{u,f}$ are perceived by spider u as data is sent by closer female f.

It is clear that weight (w) and bias (b) attributes are constrained with higher influence on ANN function [14]. In this work, the SSO technique is employed for parameter optimization of ANN. The ANN model undergoes training with the parameters present in the social spider. The 10-fold cross-validation (CV) approach is employed for evaluating the fitness function (FF). The FF can be represented as follows.

$$Fitness = 1 - CA_{validation} \qquad (7)$$

$$CA_{validation} = 1 - \frac{1}{10}\sum_{i=1}^{10}\left|\frac{TP}{TP + TN}\right| \times 100 \qquad (8)$$

where, TP and TN represent the number of true as well as false classifications correspondingly.
By using derived fitness function, Eq. (5) becomes

$$w_u = \frac{\left[1 - \left(1 - \frac{1}{10}\sum_{i=1}^{10}\left|\frac{TP}{TP + TN}\right| \times 100\right)_u\right] - worst}{best - worst} \qquad (9)$$

Finally, the above equation is used to determine the weight according to the fitness rate.

RESULTS

This section describes the function of the SSO-ANN approach on two benchmark dataset. The dataset applied, performance metrics and the results are explained in the following sections.

Dataset used: The dataset used for assessing the experimental values of the SSO-ANN model are German Credit [16] and Credit card fraud detection [17] dataset. Firstly, the number of instances in the German Credit dataset is 1000 credit applicants with 20 attributes. The number of classes is two including good and bad credits. The number of instances in good credit is 700 and the remaining 300 instances come under bad credit. Secondly, the credit fraud detection dataset includes the credit card transactions by cardholders in Europe in September, 2013. This dataset includes a total of 2,84,315 transactions with 30 attributes. This dataset also comprises the instances under good (1) and bad (0) credit classes. The information related to the dataset is given in Table 1.

Table 1: Dataset Description

Descriptions	German Credit Dataset	Credit Fraud Detection Dataset
Source	UCI	Kaggle
# of instances	1000	284807
# of attributes	20	30
# of class	2	2
Classes: Good/Bad	700/300	284315/492

Table 2 provides the FS results offered by the ACO algorithm on two dataset along with its best cost. The table values pointed out that 4features were chosen by ACO algorithm on the German Credit dataset with the best cost of 0.14. Besides, the number of features chosen in Credit Fraud Detection dataset is 13 with the best cost of 0.832.

Table 2: ACO based Selected Features and its Best Cost

Dataset	Selected Features	Best Cost
German Credit	9,2,1,4,7	0.140
Credit Fraud Detection	1,2,3,5,6,8,9,11,16,19,21,24,28	0.832

Table 3 provides the comparative study of the classifier outcome offered by different classification models on German Credit dataset in terms of several measures. The table depicts that the DT model has demonstrated lower classifier outcome with the sensitivity and specificity of 76.75% and 52.61% respectively. Simultaneously, the RBF Network has surpassed DT model by attaining the sensitivity and specificity of 78.58% and 59.55% respectively. In the same way, the MLP model has outperformed the earlier models with the sensitivity and specificity of 79.72% and 55% respectively. Along with that, the LR model has obtained even better outcome with the sensitivity and specificity of 79.82% and 60.74% respectively. On continuing with, the ACO-DC model has shown moderate results with the sensitivity and specificity of 77.93% and 69.87% respectively. Besides, the NGSAII has tried to show acceptable classifier outcome with the sensitivity and specificity of 89% and 89% respectively. At the same time, the SMOPSO model has showcased near optimal sensitivity and specificity of 90% and 90% respectively. However, the proposed SSO-ANN model has outperformed all the compared methods with the maximum sensitivity and specificity of 93.88% and 91.43% respectively.

Table 3: Performance Evaluation of Various Classifiers on German Credit Dataset

Classifier	Sensitivity	Specificity	Accuracy	F-score	Kappa
SSO-ANN	93.88	91.43	93.20	95.21	83.49
MLP	79.72	55.00	72.80	80.84	33.98
RBFNetwork	78.58	59.55	74.30	82.58	34.10
LR	79.82	60.74	75.20	82.99	37.50
DT	76.75	52.61	71.20	80.41	26.93
ACO-DC	77.93	69.87	76.60	84.74	36.13
NSGA II	89.00	89.00	85.10	-	-
SMOPSO	90.00	90.00	92.30	-	-
PSO-SVM	-	-	81.50	-	-
GA-SVM	-	-	80.50	-	-
SVM	-	-	77.50	-	-

ABC-SVM	-	-	84.00	-	-
SVM-IGDFS	-	-	82.80	-	-
SVM-GAW	-	-	80.40	-	-
KNN-IGDFS	-	-	70.20	-	-
KNN-GAW	-	-	75.80	-	-
NB-IGDFS	-	-	77.30	-	-
NB-GAW	-	-	76.80	-	-

Fig. 3 shows the analysis of the results attained by SSO-ANN model in terms of accuracy. The figure showed that the proposed SSO-ANN model has showcased effective classifier outcome by offering a maximum accuracy of 93.20% on the applied German Credit dataset.

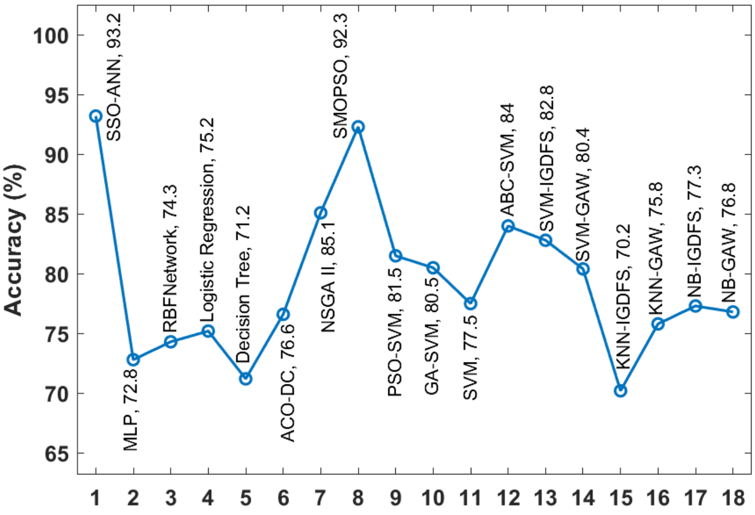

Fig. 3: Accuracy analysis of various models on German Credit dataset

A comparative study of the classifier results provided by diverse classification methods on Credit Card Fraud Detection dataset with respect to various measures is given in [Table 4]. The table showcases the sensitivity and specificity analysis of the SSO-ANN technique on the sampled Credit Card Fraud Detection dataset. The table implies that the NB model has illustrated least classifier outcome with the sensitivity of 62.40%. Concurrently, the DT has surpassed NB method by accomplishing the sensitivity and specificity of 67.83% and 68.51% correspondingly. Along with that, the RF model has an outstanding performance when compared with alternate approaches with the sensitivity of 67.89%. In line with this, the LR model has attained manageable outcome with the sensitivity and specificity of 76.52% and 78.70% respectively. On the same way, the RBF Network model has showcased gradual outcome with the sensitivity and specificity of 79.41% and 80.49% respectively. Meantime, the MLP model has demonstrated closer optimal sensitivity and specificity of 81.39% and 82.40% correspondingly. Therefore, the presented SSO-ANN model has performed effectively than compared methods with the greater sensitivity and specificity of 90.47% and 92.57% respectively.

Table 4: Performance Evaluation of Various Classifiers on Credit Card Fraud Detection Dataset

Classifier	Sensitivity	Specificity	Accuracy	F-score	Kappa
SSO-ANN	90.47	92.57	92.82	89.20	85.20
MLP	81.39	82.40	82.53	80.13	73.50
RBFNetwork	79.41	80.49	80.86	79.68	68.71
Logistic Regression	76.52	78.70	78.92	76.01	65.89
Decision Tree	67.83	68.51	69.93	68.82	63.60
Random Forest	67.89	-	91.96	78.11	-
Naïve Bayes	62.40	-	83.00	74.20	-

After examining the above-mentioned figures and tables, it is evident that the SSO-ANN algorithm has found to be an effective tool for credit card fraud detection.

CONCLUSION

This study has introduced a new FS based classification model called SSO-ANN for credit card fraud detection. Initially, the input data undergo preprocessing to transform the data into a compatible format. Then, ACO-FS algorithm is executed to select the useful number of features from the preprocessed data. Finally, SSO-ANN based classification process takes place to determine the existence of credit card frauds or not. The performance of the SSO-ANN model has been tested against two benchmark dataset namely German Credit dataset and Kaggle's Credit Card Fraud Detection dataset. The experimental outcome pointed out that the SSO-ANN model has shown superior results with the maximum classifier accuracy of 93.20% and 92.82% on the German Credit dataset and Kaggle's Credit Card Fraud Detection dataset. In future, the performance of the SSO-ANN model can be improved by the use of clustering techniques.

CONFLICT OF INTEREST

There is no conflict of interest.

ACKNOWLEDGEMENTS

None

FINANCIAL DISCLOSURE

None

REFERENCES

[1] Adewumi AO, Akinyelu AA. [2017] A survey of machine-learning and nature-inspired based credit card fraud detection techniques, International Journal of System Assurance Engineering and Management, 8:937–953.

[2] Quah JT, Sriganesh M. [2008] Real-time credit card fraud detection using computational intelligence, Expert Systems with Applications, 35(4):1721–1732.

[3] Halvaiee NS, Akbari MK. [2014] A novel model for credit card fraud detection using Artificial Immune Systems," Applied Soft Computing, 24:40–49.

[4] Mahmoudi N, Duman E. [2015] Detecting credit card fraud by modified Fisher discriminant analysis, Expert Systems with Applications, 42(5):2510–2516.

[5] Duman E, Ozcelik MH. [2011] Detecting credit card fraud by genetic algorithm and scatter search, Expert Systems with Applications, 38(10):13057–13063.

[6] Olszewski D. [2014] Fraud detection using self-organizing map visualizing the user profiles, Knowledge-Based Systems, 70:324–334.

[7] Rahimikia E, Mohammadi S, Rahmani T, Ghazanfari M. [2017] Detecting corporate tax evasion using a hybrid intelligent system: A case study of Iran, International Journal of Accounting Information Systems, 25:1–17.

[8] Christou IT, Bakopoulos M, Dimitriou T, Amolochitis E, Tsekeridou S, Dimitriadis C. [2011] Detecting fraud in online games of chance and lotteries, Expert Systems with Applications, 38(10):13158–13169.

[9] Tsai CF. [2014] Combining cluster analysis with classifier ensembles to predict financial distress Information Fusion, 16:46–58.

[10] Chen FH, Chi DJ, Zhu JY. [2014] Application of Random Forest, Rough Set Theory, Decision Tree and Neural Network to Detect Financial Statement Fraud–Taking Corporate Governance into Consideration, In International Conference on Intelligent Computing, 221–234.

[11] Li Y, Yan C, Liu W, Li M. [2017] A principle component analysis based random forest with the potential nearest neighbor method for automobile insurance fraud identification, Applied Soft Computing, to be published, DOI: 10.1016/j.asoc.2017.07.027.

[12] Subudhi S, Panigrahi S. [2017] Use of optimized Fuzzy C-Means clustering and supervised classifiers for automobile insurance fraud detection," Journal of King Saud University-Computer and Information Sciences, to be published, DOI: 10.1016/j.jksuci.2017.09.010.

[13] Uthayakumar J, Metawa N, Shankar K, Lakshmanaprabu SK. [2020] Financial crisis prediction model using ant colony optimization. International Journal of Information Management, 50:538-556.

[14] Gambhir S, Malik SK, Kumar Y. [2017] PSO-ANN based diagnostic model for the early detection of dengue disease. New Horizons in Translational Medicine, 4(1-4):1-8.

[15] Cuevas E, Cienfuegos M, ZaldíVar D, Pérez-Cisneros M. [2013] A swarm optimization algorithm inspired in the behavior of the social-spider. Expert Systems with Applications, 40(16):6374-6384.

[16] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

[17] https://www.kaggle.com/mlg-ulb/creditcardfraud