Journal of Intelligent & Fuzzy Systems 36 (2019) 2333–2344
DOI:10.3233/JIFS-169944
IOS Press

Fuzzy clustering using salp swarm algorithm for automobile insurance fraud detection

Santosh Kumar Majhi[a,*], Subho Bhatachharya[a], Rosy Pradhan[b] and Shubhra Biswal[a]
[a] *Department of Computer Science and Engineering, Veer Surendra Sai University of Technology, Burla, Odisha, India*
[b] *Department of Electrical Engineering, Veer Surendra Sai University of Technology, Burla, Odisha, India*

Abstract. In this paper, a hybrid fuzzy clustering techniques using Salp Swarm Algorithm (SSA) is proposed. The proposed fuzzy clustering method is used to optimize the cluster centroids obtained as an under sampling method. The performance of the proposed fuzzy clustering method is compared with some well-known clustering algorithms to shows the superiority of the proposed clustering algorithm. In addition, a novel hybrid Automobile Insurance Fraud Detection System is proposed in which undersampling of the majority class is performed by using the proposed fuzzy clustering algorithm which eliminates the outliers from the majority class samples. The balanced dataset for automobile fraud detection obtained after undersampling undergoes classification. Different classifiers used for this purpose are Random Forest Classifier, Logistic Regression Classifier and XGBoost Classifier. The performance of each of the three classifiers is evaluated by considering different performance metrics such as sensitivity, accuracy and specificity. The proposed fuzzy clustering method along with XGBoost outperforms the other methods presented.

Keywords: Fuzzy C-means, salp swarm algorithm, random forest classifier, logistic regression classifier, XGBoost classifier

1. Introduction

With the continuous progress in the field of media and communication networks, many new techniques to commit financial fraud have been developed. Financial fraud can be defined as an act of gaining illegal financial benefit in contrary to law, policy or rule [1]. Financial fraud includes credit card fraud, corporate fraud, insurance fraud and money laundering. Automobile insurance fraud refers to misleading an insurance company by claiming monetary support for vehicular theft or damage with the help of false documents [2]. Automobile insurance fraud has become one of the important concerns for insurance companies and also for the consumers as behavior of the person receiving the compensation in the event of an accident, is not always honest.

The automobile insurance fraud can be done by filling a fake application or by plotting accidents or thefts. The fraud detection becomes difficult in case of false representation of data [3]. Also only a small number of claims have been observed to be illegal making the detection process more difficult. Exact classification of fraudulent instances is important for Automobile Insurance Fraud Detection System (AIFDS). A robust automobile insurance fraud detection system must efficiently differentiate the malicious samples from the normal insurance claims. Moreover, it should minimize the misclassification rate.

[*] Corresponding author. Santosh Kumar Majhi, Department of Computer Science and Engineering, Veer Surendra Sai University of Technology, Burla, Odisha, 768018, India. E-mail: smajhi_cse@vssut.ac.in.

ISSN 1064-1246/19/$35.00 © 2019 – IOS Press and the authors. All rights reserved

In this paper a novel hybrid AIFDS is proposed in which undersampling of the majority class is performed by using salp swarm optimized fuzzy c-means clustering which eliminates the outliers from the majority class samples. Salp Swarm Algorithm (SSA) is used to optimize the cluster centroids obtained from Fuzzy C-means clustering (FCM) as an under sampling method. The undersampling results in a balance and reduceddataset which is used for the training phase of classifiers. In this work Random Forest Classifier, Logistic Regression Classifier and XGBoost Classifier have been used.

The rest of the paper is organized as follows: Section 2 gives a brief description of some related works. Section 3 describes the materials and methods required which discusses about Fuzzy C-Means, Salp Swarm Algorithm and the different classifiers used. Also it gives the details of the dataset and the outlier detection method using the proposed hybrid SSA-FCM approach. Section 4 presents the proposed automobile insurance fraud detection framework. Section 5 analyzes and compares the performance of the classifiers based on different performance parameters which demonstrates the efficiency the proposed method. Finally, Section 6 concludes the work by providing a brief summary.

2. Related works

A hybrid approach for stacking and bagging of meta-classifiers has been proposed by Phua et al. [4]. The class imbalance problem was overcome by performing under-sampling over the majority class. The balanced data is fed to basic classifiers like Naïve Bayes, C4.5 and Backpropagation Neural Networks. A Bayesian Dichotomous Logit model was applied to a Spanish automobile dataset by Bermudez et al. [5]. Thismethod calculates probability of imbalanced dataset using asymmetric links. Subelj et al. [3] proposed a graph based social network model which applies an iterative assessment algorithm to assign a suspicion score to each data point in the graph. The malicious claims are analyzed by the edges with the neighbouring nodes. Xu et al. [6] have used an ensemble learning technique based on rough set based neural networks. Here the whole data has been divided into non-overlapping rough subspaces and then neural network has been applied at each space. Tao et al. [7] used Fuzzy SVM which assigns a dual membership value for each fraud instance with respect to the sample mean vector. Based on the membership values the classification was done. In paper [8] oversampling of minority class is done by Synthetic Minority Oversampling Technique (SMOTE). Sundarkumar and Ravi [9] proposed another undersampling and outlier detection based on k-reverse nearest neighborhood (kRNN) and one-class SVM (OCSVM). As a result, the outliers and the noisy data were easily detected. Then basic models were applied to the pruned dataset. Subudhi and Panigrahi [10] proposed yet another efficient approach for outlier detection in the majority class. They applied Genetically Optimize Fuzzy C-Means clustering over the majority class. FCM helps in finding meaningful clusters by assigning Fuzzy membership values. The Euclidian distance of each data point was calculated from the cluster centres. If the distance was more than a threshold value, then the data point was treated as an outlier and hence removed. The cluster centres obtained from FCM were optimized using Genetic Algorithm. The irrelevant data points in a dataset are responsible for decreasing the efficiency of a classifier [11].

Fuzzy C-means is a faster clustering algorithm but it gets trapped in the local minimums easily [12]. Taherdangkoo and Bagheri [13] proposed a hybrid FCM and Stem cells algorithm (SCA) so that SCA can solve the local optima problem caused by FCM. For every iteration SC-FCM implements FCM to the cells to improve the fitness value. In the recent years different variants of Particle Swarm Optimization (PSO) have been proposed to improve the performance of the optimization method. Esmin et al. [14] have given a review of implementation of PSO and the variants of PSO in clustering high dimensional data. For linearly non-separable and complex datasets PSO has been considered as more suitable to find the centroids of clusters. For better management and characterization of uncertainty in data, a clustering algorithm termed as interval valued possibilistic fuzzy c-means (IPFCM) has been proposed by Yong Xia et al. using two fuzzifiers [15]. Here an interval valued fuzzy set is introduced into possibilistic fuzzy c-means method to overcome the drawbacks of IFCM and IPCM.

Hassanzadeh and Meybodi have proposed a firefly optimization based clustering algorithm and compared the obtained results with PSO, K-means, and K-PSO considering standard datasets [16]. Han et al. [17] proposed a clustering algorithm based on the Bird Flock Gravitational Search Algorithm (BFGSA). The algorithm is compared with the GSA, the Artificial Bee Colony (ABC), the Firefly Algorithm (FA), K-means and different variants of Particle Swam Optimization such as NM-PSO, K-PSO, K-NM-PSO, and CPSO. The experimental results shows better performance of BFGSA over other compared algorithms. Firouzi et al. introduced a hybrid Simulated annealing and ant colony optimization based data clustering algorithm [18]. Kao et al. has implemented the hybrid of PSO and K-means [19] and compared with Genetic Algorithm based clustering [20] and hybrid of GA and K-means [21] to improve the clustering result. This algorithm has a better convergence characteristic with a few numbers of evaluations. However, the main drawback is having the overlapped data points. A new data clustering approach is proposed in paper [22] based on PSO integrated with the kernel density estimation (KDE). KDE is used to improve the balance between exploitation and exploration. The hybridization of improved PSO and genetic algorithm (GA) along with K-means algorithm improves the convergence speed as well as helps to find the global optimal solution. In the first stage, IPSO has been used to get a global solution in order to get optimal cluster centres. Then, the crossover steps of GA are used to improve the quality of particles and mutation is used for diversification of solution space in order to avoid premature convergence.

Elimination of the noisy instances from the original imbalanced dataset is important for an AIFDS. In this proposed work the SSA based FCM (SSA-FCM) clustering is used for undersampling of data which eliminates the outliers. FCM is used to manage the overlapping of cluster boundaries. The Salp Swarm Algorithm is applied to optimize the cluster centres obtained from FCM. For the designing of AIFDS, the proposed SSA-FCM is used and the claims which are found to be suspicious are verified by three different classifiers.

3. Materials and methods

This section gives a detailed overview of fuzzy C-Means clustering, proposed fuzzy clustering based on Salp Swarm Algorithm and how it is used for the outlier detection. Also, the classifiers used for classification are described.

3.1. Fuzzy C-means clustering

Fuzzy C-Means clustering (FCM) method developed by Dunn [23] and improved by Bezdek [24] permits a single piece of data to belong to more than one cluster by assigning a degree of membership to each data point. The FCM method minimizes the objective function J_m given in Equation (1).

$$J_m = \sum\nolimits_{i=1}^{N} \sum\nolimits_{j=1}^{c} u_{ij}^m ||x_i - c_j||^2 \qquad (1)$$

Where, m is any real number ($m > 1$), x_i is the ith data point, u_{ij} is degree of membership of x_i in cluster j, c_j is the jth cluster centroid. The FCM algorithm consists of the following steps.

Fuzzy C-Means Clustering Algorithm

1. Initialize the data matrix $U = [u_{ij}]$
2. Calculate the centroids vector for kth iteration $C^{(k)} = [c_j]$ using Equation (2).

$$C_j = \frac{\sum_{i=1}^{N} u_{ij}^m x_i}{\sum_{i=1}^{N} u_{ij}^m} \qquad (2)$$

3. $U^{(k)}$ and $U^{(k+1)}$ are updated by calculating $u_i j$ using Equation (3).

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{2/m-1}} \qquad (3)$$

4. If $\|U^{(k+1)} - U^{(k)}\| < \varepsilon$, then stop Else go to step 2.

3.2. Proposed fuzzy clustering using salp swarm algorithm

3.2.1. Overview of Salp swarm algorithm

Salp swarm algorithm (SSA) is inspired from the swarming behaviour of salps based on their foraging and navigation in oceans [25]. Salps belong to the family salpidae of order salpida. They are one of the most successful colonizing marine animals. Salp is a planktonic tunicate which has a barrel shaped, transparent, gelatinous body. It contracts itself and pumps water through its body to move forward. During the life cycle of an adult salp, it has two phases. In the first phase a salp swims alone and lives in solitary. In the second phase the salps join together to form colonies. The colonies of salps form long chains for better foraging and improved locomotion. The chains may be in the shape of a wheel or a line or any other architectural design.

The swarming behaviour of salps and the swarm chain is mathematically designed to solve optimization problems. For better modeling the population in the salp chain is categorized into leader and followers. The salp at the front of the chain is considered as leader and the rest of the salps in the chain are categorized as followers. The leader guides the chain and the followers follow the leader and also each other.

For mathematical modeling the position of all salps are taken in a two dimensional matrix x and there is another variable F, which is considered as the food source in the search space. The food source is the swarm's target. The position of the leader is updated using Equation (4).

$$x_j^1 = \begin{cases} F_j + C_1((ub_j - lb_j)C_2 + lb_j), C_3 \geqslant 0 \\ F_j - C_1((ub_j - lb_j)C_2 + lb_j), C_3 < 0 \end{cases} \tag{4}$$

x_j^1 is the position of the leader in jth dimension, F_j is the Position of the food source in jth dimension, ub_j is the Upper bound of jth dimension, lb_j is the Lower bound of jth dimension and c_1 is a factor which is responsible for exploration and exploitation. It can be formulated as given in Equation (5).

$$C_1 = 2e^{-\left(\frac{4l}{L}\right)^2} \tag{5}$$

Here l is the current iteration and L is the maximum number of iteration. c_2, c_3 are random number in the interval [0,1].

The position of followers is updated using Equation (6).

$$x_j^i = \frac{1}{2}\left(x_j^i + x_j^{i-1}\right) \tag{6}$$

Where, $i \geqslant 2$ and x_j^i is the Position of the ith follower salp in jth dimension and x_j^{i-1} is the position of the $(i-1)^{th}$ follower salp in jth dimension.

3.2.2. *Proposed hybrid fuzzy clustering using SSA*

Wehave integrated SSA with FCM to obtain better clusters with reduced total cost. This is a meta- heuristic approach which tries to find the optimal solution. FCM is applied to the data points to obtain the initial clusters and SSA is used on the obtained clusters to optimize the cluster centres' positions. Each cluster for which the optimization is being carried out is initialized as the salp population. The cost function of the population is calculated and the best fitness value is considered as the food position F. The leader and the follower positions are updated using the Equations (4 and 6) respectively. The same procedure is continued until the end criterion is satisfied. The final food cost and food position gives us the optimum cost and optimum cluster positions respectively. Again, FCM is performed using the obtained optimized centres which results in optimal clustering. The leader first exploits the solution and then explores around it. During the location update, some salps may deviate out of the search space. They are amended and brought back inside the search space. The pseudocode for the fuzzy clustering algorithm using Salp Swarm Algorithm (SSA-FCM) is given below:

```
1. Perform FCM on the original dataset
2. Segregate the data in terms of their respective clusters $c_1, c_2, \ldots c_k$
3. For each cluster 1 to k
    3.1. Randomly initialize the salp population with the data points of cluster k
    3.2. Calculate the cost function for each salp
    3.3. Food Position = Positionof the salp with best fitness
    3.4. $t = 1$
    3.5. While $t <$ max _iter
        3.5.1. Calculate $c_1$
        3.5.2. Update position of leader using Equation (4)
        3.5.3. Update positions of followers using Equation (6)
    3.6. Amend the salps based on upper and lower bounds of the respective dimensions
    3.7. Calculate Fitness of each salp
    3.8. Update the Food position if there is a better solution
    3.9. $t = t + 1$
4. Return the Food position as the best solution
5. Perform FCM again with the obtained cluster centres
```

The proposed fuzzy clustering algorithm using Salp Swarm Algorithm (SSA-FCM) has been applied on seven benchmark datasets. The datasets are Iris, Wine, Seed, Breast cancer, Glass, E-coli, and CMC [26]. The performance of the SSA-FCM algorithm is compared against the performance of PSO-FCM, FCM and K-Means based on intracluster distance and F-measure as performance parameters. Intracluster distance and F-measure are the factors responsible for the quality of clustering [27]. Better quality of clusters is obtained if the intracluster distance is minimum and the F-measure is maximum. Theresults obtained from SSA-FCM and its comparison with PSO-FCM, FCM and K-Means for the seven datasets are given in Table 1.

From Table 1 it is clear that the proposed fuzzy clustering based on Salp Swarm Algorithm (SSA-FCM) performs better than PSO-FCM, FCM and K-Means in terms of intracluster distance and F-measure. Furthermore from the Friedman's test [28, 29] performed for the four algorithms and 7 datasets gives the value of Friedman's statistic (X_{F^2}) as 21 and the Iman and Davenport statistics (F_F) as infinite. The rank test for the Friedman's test based on intracluster distance is given in Table 2. The critical value for the degree of freedom F(3,18) as calculated from the number of algorithms and number of datasets is 3.16 for level of confidence $\alpha = 0$.

The Friedman's test proofs the existence of difference in the performance of the algorithms as the critical value attained is less than the F_F value. However, the superiority of the proposed SSA-FCM algorithm is proved from the Holm test [30]. Table 3 provides the results of the Holm test for the algorithms with SSA-FCM as the control group. In is clear from Table 3 that the SSA-FCM performs better than all compared methods as the hypothesis is rejected for all cases.

3.3. Overview of classification algorithms

In this work three different classifiers are used for the training of the balanced and reduced dataset. Those are Random Forests Classifier (RF), Logistic Regression (LR) Classifier and XGBoost Classifier.

3.3.1. Random forests classifier

Random forest (RF) is a very powerful machine learning model used for the purpose of classification and regression [32]. A random forest is a collection of a given number of decision trees. This number is specified by the user as a parameter. After a number of decision trees are created from the training data, each testing data sample is allowed to pass through each tree. Therefore, each constituent tree predicts a class label for that test sample. A majority voting is performed in order to obtain the final class label. It is important for the decision trees involved to bedecorrelated. The algorithm works in the following way.

Entropy is a split criterion for decision trees. This is the opposite of Information Gain and is the measure of impurity. For a binary classification problem Entropy is given by Equation (7).

$$E(S) = -p_1 \log p_1 - p_2 \log p_2 \tag{7}$$

Where, p_1 is the Probability of samples of class 1 and p_2 is the probability of samples of class 2. So, for a given feature the entropy values are calculated before and after the cut. If the value decreases the split is done otherwise another feature is chosen. This difference is termed as the Information gain. The attribute having the highest information gain for that given sample is used as the root node. Other nodes are chosen in terms of decreasing order of information gains.

The next parameter is the number of trees. As a general rule the number of trees is taken as the double of the number of features for better accuracy. In this work, we have generated 25 features. The number of features has increased because of dummy encoding. As a result, the number of decision trees is taken as 50. It is found that the model gives best performance when M is 50.

3.3.2. Logistic regression

Logistic Regression (LR) is an excellent binary classification technique which finds its application in fraud detection systems. It is used in datasets where two or more independent variables are used to decide an outcome which is a dichotomous variable. It can be considered as an extended version of linear regression. Instead of dealing with real values, it deals with binary outcomes. A probability score is assigned to each outcome. Based on that score the class of the sample is decided. In case of linear regression, we are concerned with the best fit line. However, in this case the main concern is the best fit sigmoid curve. The sigmoid curve provides an excellent way for studying probability distribution owing to its shape.

The mathematical equation for logistic regression is given by Equation (8).

$$logit(p) = a_0 + a_1 X_1 + a_2 X_2 + a_{3X_3} \ldots\ldots + a_n X_n \tag{8}$$

Here $a_0, a_1, \ldots a_n$ are the coefficients and $X_1, X_2, \ldots X_n$ are the independent variables. p is the probability of occurrence of a characteristic.

And,

$$logit(p) = In \frac{p}{1-p} \tag{9}$$

The term $\frac{p}{1-p}$ represents the odds of occurrence of the characteristic.

Table 1
Performance comparison of SSA-FCM, PSO-FCM, FCM and K-Means

Datasets	Methods	Best value of intracluster distance	Average value intracluster distance	Worst value of intracluster distance	F-measure	Standard deviation
Iris	SSA-FCM	**34.9527**	34.9537	34.9550	**0.8926**	**0.0012**
	PSO-FCM	35.8254	35.8756	35.9143	0.8923	0.0338
	FCM	35.8925	36.1973	36.5471	0.8413	0.2905
	K-Means	36.7846	37.0204	38.2134	0.8853	0.3621
Wine	SSA-FCM	**4.981**	4.9812	4.982	**0.7043**	**0.0004**
	PSO-FCM	4.981	4.9826	4.985	0.6986	0.0018
	FCM	5.112	5.334	5.408	0.6855	0.1643
	K-Means	5.70	6.0104	6.36	0.6521	0.2716
Seed	SSA-FCM	**43.2988**	43.3589	43.4690	0.8841	**0.0755**
	PSO-FCM	53.4781	53.6870	53.8981	**0.8949**	0.1459
	FCM	53.5376	53.8990	54.2127	0.8741	0.2721
	K-Means	57.5150	58.1464	58.9015	0.8291	0.5885
Breast cancer	SSA-FCM	**103.5274**	103.655	103.8251	**0.5675**	**0.1157**
	PSO-FCM	105.3188	105.6545	105.9166	0.5555	0.2488
	FCM	106.7207	107.4491	108.9219	0.5217	0.8763
	K-Means	111.3157	113.6715	115.0018	0.4988	1.8143
Glass	SSA-FCM	**2.3406**	2.3413	2.3422	**0.8443**	**0.0006**
	PSO-FCM	2.3721	2.3761	2.3796	0.8322	0.0031
	FCM	2.3755	2.7090	2.9654	0.8318	0.2538
	K-Means	2.8005	3.2797	3.7105	0.8258	0.3506
E-Coli	SSA-FCM	**2.6909**	2.7372	2.8166	**0.6174**	**0.0476**
	PSO-FCM	2.7521	2.8595	2.9921	0.5994	0.1078
	FCM	2.8169	2.9722	3.1016	0.5780	0.1099
	K-Means	3.0015	3.2150	3.4317	0.5126	0.1688
CMC	SSA-FCM	**718.3588**	719.3903	720.8819	**0.4024**	**0.9349**
	PSO-FCM	722.3191	723.8458	725.5053	0.4005	1.2329
	FCM	735.6474	736.9925	738.8018	0.3975	1.3525
	K-Means	739.0016	742.0634	745.8018	0.3711	2.8252

Table 2
Rank test of the clustering algorithms obtained from Friedman's test

Data set	K-Means	FCM	PSO-FCM	SSA-FCM
Iris	37.0204 **(4)**	36.1973 **(3)**	35.8756 **(2)**	34.9537 **(1)**
Wine	6.0104 **(4)**	5.334 **(3)**	4.9826 **(2)**	4.9812 **(1)**
Seed	58.1464 **(4)**	53.8990 **(3)**	53.6870 **(2)**	43.3589 **(1)**
Breast Cancer	113.6715**(4)**	107.4491**(3)**	105.6545**(2)**	103.655 **(1)**
Glass	3.2797 **(4)**	2.7090 **(3)**	2.7090 **(2)**	2.3413 **(1)**
E-Coli	3.2150 **(4)**	2.9722 **(3)**	2.9722 **(2)**	2.7372 **(1)**
CMC	742.0634**(4)**	736.9925**(3)**	736.9925**(2)**	719.3903**(1)**
Average rank (Rj)	**4**	**3**	**2**	**1**

Table 3
Results obtained from Holm test

i	Algorithms	*z*-value	*p*-value	*α/(k-i)*	Hypothesis
1	K-Means	–4.34	<0.00001	0.16	Rejected
2	FCM	–2.89	0.00193	0.025	Rejected
3	PSO-FCM	–1.45	0.00735	0.05	Rejected

Regularization is used in Logistic Regression to simplify the hypothesis and to prevent over fitting. It is basically a penalty term which is added to the cost function. The two most commonly used regularization methods are L_1 and L_2. For the project L_2 regularization is used. The details are given below.

Let C_{old} represent the cost function of the logistic regression. After adding the penalty term, the new cost function is given by Equation (10).

$$C_{new} = C_{old} + \lambda \sum\nolimits_{i=1}^{m} (a_i)^2 \tag{10}$$

Here λ is the regularization parameter, m is the number of independent variables and a_i refers to the ith coefficient.

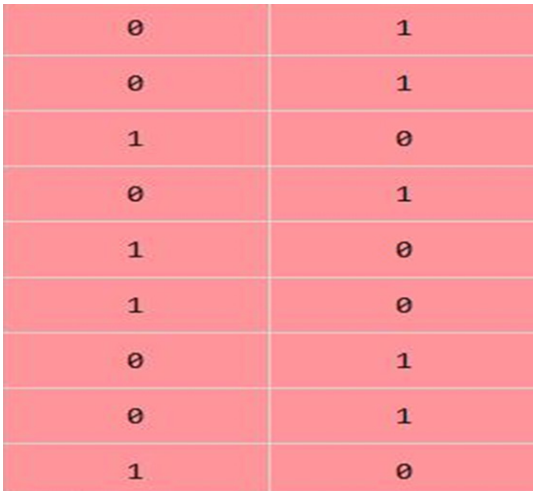
In case of L_1 regularization, instead of squaring each coefficient, we take the absolute values. Regularization aims to decrease the magnitude of the coefficients, which helps in preventing overfitting.

3.3.3. XGBoost

XGBoost is developed by Tong He and Tianqi Chen [33] and it stands for eXtreme Gradient BOOSTing. It involves boosting techniques which aims to obtain a stronger classifier by combining the predictive power of the weaker constituent classifiers. It uses Decision trees as its constituent classifiers. It is a highly scalable and a computationally faster model and is used for classification, ranking and regression. It is an ensemble method just like random forests. However, it uses an iterative learning algorithm. In this method new tries are grown using the information from the previously grown trees. At each iteration the misclassification points of the current tree are taken into consideration, and these points are assigned weights. The next iteration aims to produce a tree which helps to classify the above-mentioned points properly. Therefore, a new tree is obtained at the end of every iteration. The final model combines all the trees, thereby decreasing the misclassification error. Thus, the whole process is iterative and additive in nature. XGBOOST has dedicated libraries in python and R. Therefore, it can be implemented easily. Here also, just like logistic regression L_1 or L_2 regularization is used.XGBOOST is a very powerful model and as expected produces better results than the previously used models. In this work default parameters were used for creating the model.

3.4. Details and preprocessing of the dataset

The dataset “carclaims.txt” [4] has been used in this work which is the only publicly available dataset for automobile fraud detection. However, the data has been cleaned resulting in no missing or out of bounds value in the whole data. For verification the above the dataset has been examined properly using Weka Data Mining software to get a basic idea about the range of values for each feature. Here, the problem is the binary classification, i.e. there are two classes namely Fraud and No Fraud. The dataset is completely numeric and the categorical variables present in the dataset are encoded to numbers. The dataset has 21 features including the class label (Fraud or No Fraud) and has 15420 rows. The data distribution is however heavily imbalanced. The number of fraud samples is 923 and the number of No fraud cases account to 14496. The minority class accounts to just 5.985% of the total dataset. Some important features of the dataset are described in the Table 4.



0	1
0	1
1	0
0	1
1	0
1	0
0	1
0	1
1	0

Fig. 1. Encoded dummy variables from a feature which has 3 categories.

The concept of dummy variables is used for dealing with categorical data. For example, for the class marital status there are 4 categories. They are encoded from 0–3. They are converted to Dummy variables by creating 4 new rows (The number of rows to be created is always equal to the number of categories present). Thus 0 leads to row vector – [1, 0,0,0], 1 leads to – [0, 1, 0, 0] and so on. Figure 1 shows the encoded dummy variables of a feature having three categories. In order to prevent the dummy variable, the first newly generated row is deleted. Here there is no loss of information.

For tackling the class imbalance problem, we have used 3 methods and compare among them. Those are: (1) Performing Random Undersampling over the majority class, (2) Performing FCM over the majority class and remove the outliers and (3) Performing SSA optimized FCM over the majority class and remove the outliers.

4. Application of proposed SSA-FCM method for outlier detection

In this work, we have proposed an AIFDS to reduce misclassification error in turn to increase the accuracy

Table 4
Important features of the dataset

Name	Description	Type
Week_past	No. of weeks between the accident and claim	Numeric
Is_holiday	The day of the accident falls in an American Holiday week or not	Boolean
Age_price_wsum	Combination of age of vehicle attribute and vehicle price attribute	Numeric
Marital Status	Self - explanatory	Numeric categorical (Single, Married, Widow, Divorced)
Fault	Entity responsible for the accident	Numeric Categorical (Policy Holder or Third Party)
Vehicle Category	Type of vehicle	Numeric Categorical (Sports, sedan,etc)
Deductible	The amount paid by the policy holder to the company.	Numeric
Driver Rating	Self – explanatory	Numeric (1–5)
Past Number of claims	Self – explanatory	Numeric
Police Report Filed	Self – explanatory	Boolean
Witness Present	Self – explanatory	Boolean
Age of Policy Holder	Self – explanatory	Categorical (Sequential)
Base Policy	Type of policy applied for	Numeric Categorical (Liability, Collision, All perils)
Week_past	No. of weeks between the accident and claim	Numeric

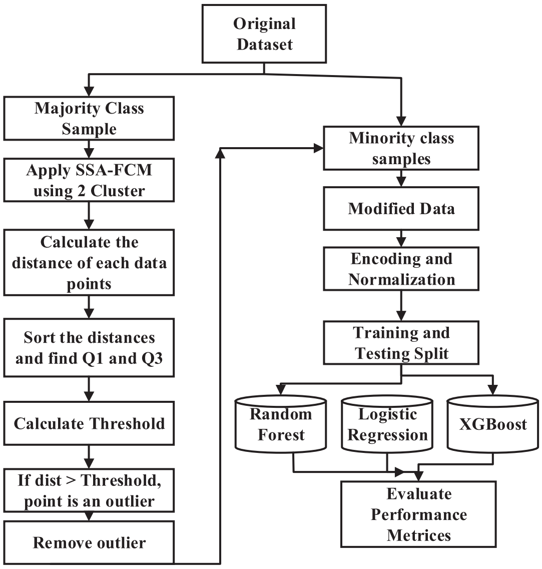

Original Dataset
Majority Class Sample
Apply SSA-FCM using 2 Cluster
Calculate the distance of each data points
Sort the distances and find Q1 and Q3
Calculate Threshold
If dist > Threshold, point is an outlier
Remove outlier
Minority class samples
Modified Data
Encoding and Normalization
Training and Testing Split
Random Forest
Logistic Regression
XGBoost
Evaluate Performance Metrices

Fig. 2. Outlier detection using proposed fuzzy clustering.

for the considered insurance data set. The proposed fuzzy clustering technique based on SSA-FCM is applied over the majority class to detect the outliers present in the samples. The outliers are removed and outlier free data set is fed to the classifiers. Figure 2 shows the process of outlier detection and classification using the classifiers. Outlier detection is performed by the following methodology:

Step 1. After clustering is performed, the distance of each data point from their respective cluster centres are calculated using the distance formula is given by Equation (11).

$$dist_i = \sqrt{\sum\nolimits_{i=1}^{N} |c_i - d_i|} \tag{11}$$

Where, $dist_i$ is the distance between the cluster centre and ith data point, N is the number of dimensions, c_i is the cluster centre of the ith data point and d_i is the ith data point.

Step 2. The respective distance values obtained are sorted. From the sorted values the first quartile (Q_1) and the third quartile(Q_3) are calculated.

Step 3. A threshold value α is calculated by Turkey's method [31]. The formula used for calculation of α is given in Equation (12).

$$a = Q3 + 3 * |Q3 - Q1| \tag{12}$$

If $dist_i > \alpha$ the data point is an outlier. The outliers are marked accordingly.

Step 4. The outliers are deleted and the pruned samples are combined with the minority class samples to form the modified data set.

After removing the outliers from the data set, a balanced data set is prepared by considering the outlier free major class points and minor class instances. Once the balanced data set is ready, the proposed AIFDS evaluates the claim as genuine, fraudulent or suspicious based on comparison of cluster centers distance value with the lower limit (L_l) and upper limit (U_l) as given in Equations (12 and 13). These upper and lower boundary values are calculated using Tukey method [31].

$$U_l = Q1 - 3 * |Q3 - Q1| \tag{13}$$

If the cluster center distances are less than the lower boundary then the claims are considered as genuine. If the distance is greater than the upper boundary then claims are taken as fraudulent and distance is in between lower and upper limits, then it is considered as suspicious. The clearance should be made for genuine cases where as precautions will be taken for the fraudulent case and suspicious cases are further analyzed by supervised classifiers. Classifiers such as random forest, logistic regression and XGBoost are considered in our work. The classifiers will take the input suspicious data sets and based on the classifiers output a decision will be made regarding genuine or fraudulent. Next, the performance of the classifiers is evaluated and analyzed.

5. Results and analysis

The proposed Automobile Insurance Fraud Detection System based on SSA optimized fuzzy clustering method is implemented in MATLAB 2016 on a 2.4 Ghz core i5 CPU system. The majority class of the dataset has 14496 samples, which accounts for almost 94% of the total data set. Hence by performing clustering over the majority class the outliers present in the samples are detected and removed before the dataset being fed to the classifier. The clustering parameters are specified in Section 5.1 and the performance metrics such as sensitivity, accuracy and specificity are given in Section 5.2. These measures are used to evaluate the performance of the proposed system.

5.1. Clustering parameters determination

In FCM the number of clusters required to group the data is determined by two parameters namely Partition Co-efficient (PC) and Partition Entropy (PE). PC is the average of membership value shared in between each fuzzy subset pair inside themembership matrix U. The Partition Co-efficient (PC) is given by Equation (14).

$$PC = \frac{1}{n \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{2}} \tag{14}$$

Where, n is the number of data points and c is the number of clusters.

PE gives the amount of fuzziness present in U and it is given by Equation (15).

$$PE = \frac{-1}{n \sum_{i=1}^{c} u_{ij} \log(u_{ij})} \tag{15}$$

These values are calculated for different values of c, where c is the no. of cluster centres.

5.2. Performance metrics

The performance of the proposed method is measured based on different performance metrics such as Sensitivity, Specificity and Accuracy.

Sensitivity is the ratio of the number of truly classified samples to the total number of true samples. Sensitivity is considered as the main metric since it gives the fraud detection rate of the model.Sensitivity is calculated using Equation (16).

$$SENSITIVITY = \frac{TP}{TP + FN} \tag{16}$$

Where TP is True Positive and FN is False Negative.

Specificity refers to the proportion of the truly negative samples that were classified as negative by the classifier. Equation (17) calculates the Specificity.

$$SPECIFICITY = \frac{TN}{TN + FP} \tag{17}$$

Where TN is True Negative and FP is False Positive.

Accuracy is an overall metric which estimates the correctness of the classifier. Equation (18) is used to calculate Accuracy.

$$ACCURACY = \frac{TP + TN}{TP + TN + FP + FN} \tag{18}$$

Where TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative.

5.3. Analysis

By considering, the number of clusters $c = 2$, Partition Co-efficient (PC) is found to be maximum and Partition Entropy (PE) isfound to be minimum using Equations (14 and 15). Therefore, the number of clusters for the given dataset is considered as 2. The distance of each data point with their respective cluster centres are calculated using the distance formula given in Equation (11) and the obtained distance values are sorted. From the sorted values the first quartile (Q1) and the third quartile(Q3) are calculated. For the given dataset, the value of first quartile is found to

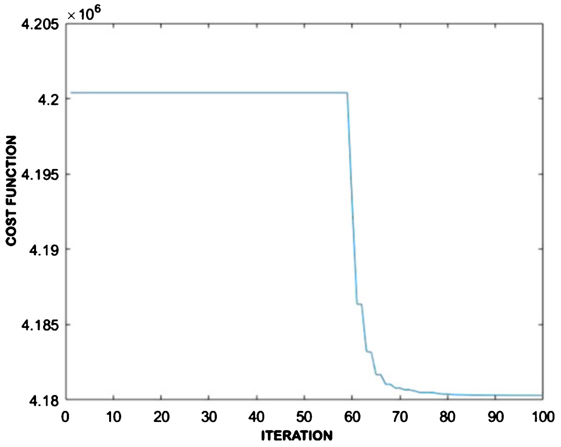


Fig. 3. Convergence curve for SSA-FCM.

be Q1 = 4.0065 and value of third quartile is found to be Q3 = 5.0045. The threshold value α is calculated using the values of Q1 and Q3 and for the given dataset the value of α is calculated as 8.0029. If the distance of the data point is greater than the threshold value then the point is detected as an outlier and is removed. The pruned samples are combined with the minority class samples to form the modified data set.

FCM as well as proposed fuzzy clustering using salp swarm algorithmare applied over the majority class samples. Considering the above given value of the threshold, 7267 data points has been detected as outliers and removed. The modified class now has 7229 genuine samples and 923 fraud samples. This trimmed data is fed to the classification models described in Section 3.3. For proposed fuzzy clustering, the number of iterations performed was 100, and from the convergence curve it is clear that the value remains constant after the 95th iteration. The convergence curve is given in Fig. 3.

The performances of the classifiers are based on sensitivity, specificity and accuracy. The efficiency of a classifier model is calculated using the sensitivity factor which recognizes the maximum number of falsified samples. Therefore the model having the highest sensitivity factor is considered as the optimal classifier. The performance of the classifiers on the balanced dataset is given in Table 5.

From Table 5, it can be observed that XGBoost classifier gives high sensitivity factor of 85.66% and 97.47% for random under sampling and SSA-FCM outlier detection method respectively. For FCM based outlier detection method Logistic regression classifier gives higher sensitivity of 93.86%.

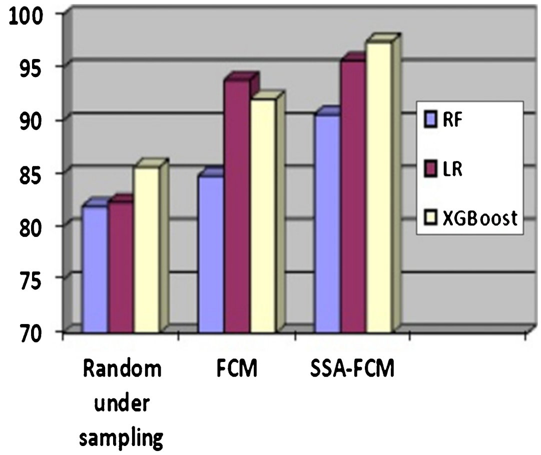


Fig. 4. Performance Comparison of the Classifiers based on Sensitivity.

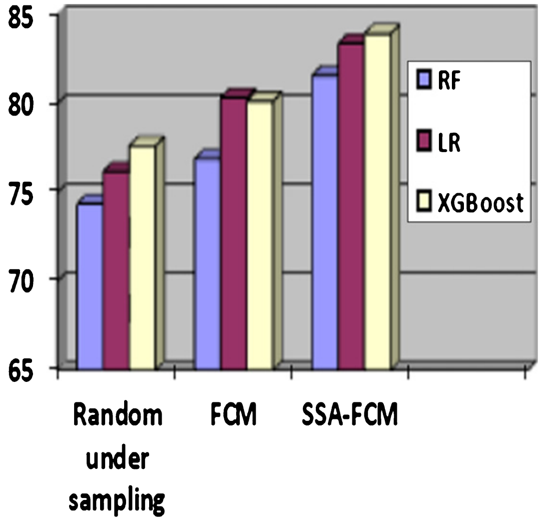


Fig. 5. Performance Comparison of the Classifiers based on Accuracy.

However the highest percentage of accuracy for all the outlier detection methods is obtained by XGBoost classifier. Moreover, the method proposed in this work performs preferably better than the results reported by Sunderkumar et al. [9] (sensitivity = 95.52%) and Subudhi et al. [10] (sensitivity = 83.21%). Figure 4 shows the comparison among the considered classifiers' performance based on the sensitivity measure and Fig. 5 compares the classifiers' performance based on accuracy.

Table 5
Performance of the classifiers on the balanced dataset

Outlier Detection Method	Classifiers used	Performance Metrics (in %)		
		Sensitivity	Specificity	Accuracy
Using Random Under Sampling	RF	81.94%	66.78%	74.36%
	LR	82.33%	61.01%	76.17%
	XGBoost	**85.66%**	68.56%	77.61%
FCM	RF	84.83%	68.95%	76.89%
	LR	**93.86%**	66.78%	80.32%
	XGBoost	92.05%	68.23%	80.14%
SSA-FCM	RF	90.61%	72.56%	81.58%
	LR	95.66%	71.11%	83.39%
	XGBoost	**97.47%**	70.39%	83.93%

6. Conclusion

In this paper, we have proposed an automobile insurance fraud detection system that uses a new hybrid fuzzy clustering based on Salp Swarm Algorithm (SSA-FCM) for outlier detection and removal. The performance of the proposed fuzzy clustering method has been evaluated by considering seven standard data sets. Moreover, the statistical tests reveal that the proposed clustering method gives good accuracy. The proposed fuzzy clustering is applied for undersampling of majority class samples of the automobile insurance data set for enhancing the effectiveness of the classifiers. The salp swarm algorithm helps in obtaining the optimal cluster centre in the SSA-FCM. The SSA-FCM calculates the distance of the datapoints from the cluster centres based on which the suspicious classes are detected. The suspicious classes are again verified using three different classifiers Random forest, Logistic regression and XGBoost. The proposed model is applied on the "carclaims.txt" automobile insurance dataset. The dataset contains 15,420 records. The SSA-FCM detected 7267 data-points as outliers. After the removal of the outliers the modified class now had 7229 genuine samples and 923 fraudulent samples. The trimmed data is fed to the classifiers for correct classification. It is observed that XGBoost classifier outperforms Random forest and Logistic Regression with a high sensitivity of 97.47% and accuracy of 83.93%. Moreover, the proposed AIFDS gives better result as compared to other recently published methods.

References

[1] J.H. Wang, Y.L. Liao, T.M. Tsai and G. Hung, Technology-based financial frauds in Taiwan: Issues and approaches, In *Systems, Man and Cybernetics, 2006 SMC'06 IEEE International Conference on*, Vol. 2, 2006, pp. 1120–1124). IEEE.

[2] E. Ngai, Y. Hu, Y. Wong, Y. Chen and X. Sun, The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature, *InDecis Support Syst* **50**(3) (2011), 559–569.

[3] L. Šubelj, Š. Furlan and M. Bajec, An expert system for detecting automobile insurance fraud using social network analysis, *Expert Systems with Applications* **38**(1) (2011), 1039–1052.

[4] C. Phua, D. Alahakoon and V. Lee, Minority report in fraud detection: Classification of skewed data, *Acmsigkdd explorations newsletter* **6**(1) (2004), 50–59.

[5] L. Bermúdez, J.M. Pérez, M. Ayuso, E. Gómez and F.J. Vázquez, A Bayesian dichotomous model with asymmetric link for fraud in insurance, *Insurance: Mathematics and Economics* **42**(2) (2008), 779–786.

[6] W. Xu, S. Wang, D. Zhang and B. Yang, Random rough subspace based neural network ensemble for insurance fraud detection, In *Computational Sciences and Optimization (CSO), 2011 Fourth International Joint Conference on*, 2011, pp. 1276–1280. IEEE.

[7] H. Tao, L. Zhixin and S. Xiaodong, Insurance fraud identification research based on fuzzy support vector machine with dual membership, In *Information Management, Innovation Management and Industrial Engineering (ICIII), 2012 International Conference on*, Vol. 3, 2012, pp. 457–460. IEEE.

[8] R. Pears, J. Finlay and A.M. Connor, Synthetic Minority over-sampling technique (SMOTE) for predicting software build outcomes, arXiv preprint arXiv:1407.2330, 2014.

[9] G.G. Sundarkumar and V. Ravi, A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance, *Engineering Applications of Artificial Intelligence* **37** (2015), 368–377.

[10] S. Subudhi and S. Panigrahi, Use of optimized Fuzzy C-Means clustering and supervised classifiers for automobile insurance fraud detection, *Journal of King Saud University-Computer and Information Sciences* (2017).

[11] Y.J. Lee, Y.R. Yeh and Y.C.F. Wang, Anomaly detection via online oversampling principal component analysis, *IEEE Transactions on Knowledge and Data Engineering* **25**(7) (2013), 1460–1470.

[12] J.C. Bezdek, R. Ehrlich and W. Full, FCM: The fuzzy c-means clustering algorithm, *Computers & Geosciences* **10**(2-3) (1984), 191–203.

[13] M. Taherdangkoo and M.H. Bagheri, A powerful hybrid clustering method based on modified stem cells and Fuzzy

C-means algorithms, *Engineering Applications of Artificial Intelligence* **26**(5-6) (2013), 1493–1502.

[14] A.A. Esmin, R.A. Coelho and S. Matwin, A review on particle swarm optimization algorithm and its variants to clustering high-dimensional data, *Artificial Intelligence Review* **44**(1) (2015), 23–45.

[15] Z. Ji, Y. Xia, Q. Sun and G. Cao, Interval-valued possibilistic fuzzy C-means clustering algorithm, *Fuzzy Sets and Systems* **253** (2014), 138–156.

[16] T. Hassanzadeh and M.R. Meybodi, A new hybrid approach for data clustering using firefly algorithm and K-means, In *Artificial Intelligence and Signal Processing (AISP), 2012 16th CSI International Symposium on*, 2012, pp. 007–011. IEEE.

[17] T. Hassanzadeh and M.R. Meybodi, A new hybrid approach for data clustering using firefly algorithm and K-means, In *Artificial Intelligence and Signal Processing (AISP), 2012 16th CSI International Symposium on*, 2012, pp. 007–011. IEEE.

[18] T. Niknam, B.B. Firouzi and M. Nayeripour, An efficient hybrid evolutionary algorithm for cluster analysis, *In World Applied Sciences Journal* (2008).

[19] Y.T. Kao, E. Zahara and I.W. Kao, A hybridized approach to data clustering, *Expert Systems with Applications* **34**(3) (2008), 1754–1762.

[20] C.A. Murthy and N. Chowdhury, In search of optimal clusters using genetic algorithms, *Pattern Recognition Letters* **17**(8) (1996), 825–832.

[21] S. Bandyopadhyay and U. Maulik, An evolutionary technique based on K-means algorithm for optimal clustering in RN, *Information Sciences* **146**(1-4) (2002), 221–237.

[22] M. Alswaitti, M. Albughdadi and N.A.M. Isa, Density-based particle swarm optimization algorithm for data clustering, *Expert Systems with Applications* **91** (2018), 170–186.

[23] J.C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, 1973.

[24] J.C. Bezdek, Objective Function Clustering, In *Pattern Recognition with Fuzzy Objective Function Algorithms*, Springer, Boston, MA, 1981, pp. 43–93.

[25] S. Mirjalili, A.H. Gandomi, S.Z. Mirjalili, S. Saremi, H. Faris and S.M. Mirjalili, Salp swarm algorithm: A bio-inspired optimizer for engineering design problems, *Advances in Engineering Software* **114** (2017), 163–191.

[26] K. Bache and M. Lichman, UCI machine learning repository, 2013.

[27] S. Chen, Z. Xu and Y. Tang, A hybrid clustering algorithm based on fuzzy C-means and improved particle swarm optimization, *Arabian Journal for Science and Engineering* **39**(12) (2014), 8875–8887.

[28] M. Friedman, The use of ranks to avoid the assumption of normality implicit in the analysis of variance, *Journal of the American Statistical Association* **32**(200) (1937), 675–701.

[29] M. Friedman, A comparison of alternative tests of significance for the problem of m rankings, *The Annals of Mathematical Statistics* **11**(1) (1940), 86–92.

[30] S. Holm, A simple sequentially rejective multiple test procedure, *Scandinavian Journal of Statistics* (1979), 65–70.

[31] J.W. Tukey, Exploratory data analysis (Vol. 2), 1977.

[32] L. Breiman, Random forests, *Machine Learning* **45**(1) (2001), 5–32.

[33] T. Chen and C. Guestrin, Xgboost: A scalable tree boosting system, In *Proceedings of the 22nd Acmsigkdd International Conference on Knowledge Discovery and Data Mining* 2016, pp. 785–794. ACM.