Journal of King Saud University – Computer and Information Sciences 32 (2020) 568–575

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Use of optimized Fuzzy C-Means clustering and supervised classifiers for automobile insurance fraud detection

Sharmila Subudhi, Suvasini Panigrahi *

Department of Computer Science and Engineering & IT, Veer Surendra Sai University of Technology, Burla, Odisha 768018, India

ARTICLE INFO

Article history:
Received 26 May 2017
Revised 18 August 2017
Accepted 27 September 2017
Available online 28 September 2017

Keywords:
Fraud detection
Insurance claims
Genetic Algorithm
Fuzzy C-Means clustering
Supervised classifiers

ABSTRACT

This paper presents a novel hybrid approach for detecting frauds in automobile insurance claims by applying Genetic Algorithm (GA) based Fuzzy C-Means (FCM) clustering and various supervised classifier models. Initially, a test set is extracted from the original insurance dataset. The remaining train set is subjected to the clustering technique for undersampling after generating some meaningful clusters. The test instances are then segregated into genuine, malicious or suspicious classes after subjecting to the clusters. The genuine and fraudulent records are discarded, while the suspicious cases are further analyzed by four classifiers – Decision Tree (DT), Support Vector Machine (SVM), Group Method of Data Handling (GMDH) and Multi-Layer Perceptron (MLP) individually. The 10-fold cross validation method is used throughout the work for training and validation of the models. The efficacy of the proposed system is illustrated by conducting several experiments on a real world automobile insurance dataset.

1. Introduction

An automobile insurance contract signed between an insured (customer) and an insurer (service provider company) provides monetary support in case of vehicular damage or theft. Automobile insurance fraud arises upon submitting fake documents regarding casualties in a staged accident or claims for past losses in order to obtain financial profit (Ngai et al., 2011). Moreover, this type of fraud can be accomplished by anyone like, insureds, chiropractors, garage mechanics, lawyers, police officers, insurance workers and others (Šubelj et al., 2011). A study done by Insurance Fraud Bureau of Australia in 2013 reflects the rising trend in illegitimate claim costs, which is $2 billion more than in 2012 (Australia: Insurance, 2016). In 2014, the Association of British Insurers (ABI) investigates the increase in number of false claims, which is 18% more than the previous year (Cutting corners, 2015). These statistics clearly illustrates the severity of the problem and hence, needs to be addressed firmly for lessening the losses incurred by such malicious attempts.

Furthermore, the auto insurance fraud can be classified into an easier way (filing a forged application) or a more deceitful way such as fabricating an accident or thefts (Abdallah et al., 2016). Besides, the improper representation of data with respect to a claim makes the fraud detection extremely difficult (Šubelj et al., 2011). Moreover, it is observed that only a small fraction of accident claims are illegitimate leading to the presence of a skewed class distribution in the dataset. This makes the detection even more challenging (Jensen, 1997). Hence, accurate classification of these fraudulent instances are essential for any Automobile Insurance Fraud Detection System (AIFDS). The iterative computation needed for segregating the genuine instances may require high computation time upon being subjected to an AIFDS (Panigrahi et al., 2013). Therefore, there is a need to develop a robust AIFDS that is able to discriminate the malicious samples from the normal insurance claims efficiently while minimizing the misclassification rate.

This paper proposes a novel hybrid AIFDS that applies the Genetic Algorithm (GA) for optimizing the cluster centers generated from Fuzzy C-Means clustering (FCM) as an undersampling approach. This is done to remove the noisy points from the majority class samples of the original unbalanced dataset leading to a reduced balanced dataset. A new insurance claim is then classified as genuine, malicious or suspicious based on its distance measure computed from the optimized cluster centers. The claim marked as

* Corresponding author.
E-mail addresses: sharmilasubudhi1@gmail.com, spanigrahi_cse@vssut.ac.in (S. Panigrahi).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2017.09.010
1319-1578/

genuine is allowed to pass through for payment processing, while the claim found to be fraudulent is blocked. In case the claim is a suspicious one, then further verification and classification is made by passing it through four different trained supervised learning models individually. The training of the classifiers is carried out apriori by applying the balanced training dataset. In this work, Support Vector Machine (SVM), Multi Layer Perceptron (MLP), Decision Tree (DT) and Group Method of Data Handling (GMDH) classifiers are used for determining the best performing classifier.

The rest of the paper is organized as follows: Section 2 briefly introduces the related research carried out in this field and Section 3 sheds some light into the background study of the techniques used in the current work. Section 4 focuses on the proposed fraud detection model. The experimentation and comparative performance analysis are provided in Section 5 to demonstrate the effectiveness of the proposed approach. Finally, Section 6 concludes the paper by providing a brief summary of the contributions made in this area.

2. Related work

In this section, the research work carried out with relevance to automobile insurance fraud detection is reviewed. An hybridization of stacking and bagging meta classifiers has been proposed in Phua et al. (2004). The stacked ensemble initially chooses the best classifier model from a pool of base learners. Then, the bagging technique is used on the selected classifier for predictive analysis of an oversampled real world labeled dataset. Another approach suggests the use of fuzzy logic concept for finding the illegitimate claims from a bunch of settled insurance claims (Pathak et al., 2005). In order to analyze and identify the suspicious claims from the insurance records, the authors have developed a statistical bivariate probit model as an auditing strategy used on a Spanish automobile insurance dataset (Pinquet et al., 2007). A skewed Bayesian dichotomous logit model has been suggested for identifying malicious insurance claims found in a Spanish automobile market (Bermúdez et al., 2008).

The usage of graph based social network model has been presented in Šubelj et al. (2011), which needs only unlabeled data for processing. An Iterative Assessment Algorithm (IAA) has been developed by the authors to identify the suspicious claims. Initially, a suspicion score is allocated to each point present in the graph and then the determination of suspicious entities is done by analyzing the edges present within their neighboring nodes. The rough set based neural network ensemble technique has been proposed in Xu et al. (2011). In this paper, initially the whole dataset space is segmented into various subspaces with the help of rough set data space reduction method. Then the neural network is applied on these subspaces individually to construct trained models. Later, the results of each model are combined using a voting ensemble technique for final decision making. The concept of fuzzy support vector machine for identification of suspicious (overlapped) insurance cases has been suggested in Tao et al. (2012). The fraud detection model initially calculates a distance value of each fraud instance with respect to two categories of sample mean vector and allocates a dual membership value to them. This helps each malicious sample to be assigned with a probability value used for classifying in two classes (genuine or fraud).

The identification of fraudulent cases by employing an quantitative approach has been proposed in Bernard and Vanduffel (2014). In this work, a Sharpe ratio and its limit values are estimated by increasing and decreasing the mean and variance values of the claim payments respectively. The detection of fraud samples in the insurance claims are then performed on the basis of these boundary values. A fraud detection model has been developed in Sundarkumar and Ravi (2015), which detects and removes outliers for reducing the class imbalance effect present in the automobile insurance dataset. Two unsupervised techniques: *k*-Reverse Nearest Neighborhood (*k*-RNN) and One Class Support Vector Machine (OCSVM) are employed in tandem for solving the skewed class distribution in the original dataset. Further, six different supervised classifiers have been applied independently on the balanced dataset for classification and comparison purposes. In paper (Nian et al., 2016), the authors have suggested the use of an unsupervised anomaly detection model, known as Spectral Ranking Anomaly (SRA) system, for detection of forged instances. This model assigns a degree of anomaly value to each claim after estimating the first non-principal eigenvector from a Laplacian matrix of the claim records. If the rank is less than a preset threshold, then the corresponding point is marked as fraudulent.

Table 1
Existing approaches along with their performance comparison.

Related Work	Techniques Used	Performance Metrics	Values (in %)
Phua et al. (2004)	Stacking-Bagging Ensemble	Accuracy	60.00
Bermúdez et al. (2008)	Bayesian Dichotomous Logit Model	Accuracy Sensitivity Specificity	99.53 99.85 72.88
Šubelj et al. (2011)	Iterative Assessment Algorithm	Accuracy Sensitivity Specificity	87.20 89.13 86.67
Xu et al. (2011)	Rough Subspace based Neural Network ensemble	Accuracy	88.70
Tao et al. (2012)	Fuzzy Support Vector Machine	Sensitivity	91.31
Sundarkumar and Ravi (2015)	*k*-RNN and OCSVM	Accuracy Sensitivity Specificity	60.61 90.74 58.69
Nian et al. (2016)	Spectral Ranking Anomaly	Sensitivity Specificity	91.00 52.00

Table 1 presents a brief summary regarding the performance of some of the techniques described in this literature in terms of *Sensitivity*, *Specificity* and *Accuracy*. The definition of the metrics have discussed in Section 5.

Despite of several AIFDSs developed to handle the fraud detection efficiently, there exist some irrelevant data points in the dataset that can reduce the efficiency of a classifier (Lee et al., 2013). Hence, the removal of these noisy instances from the original imbalanced dataset is required to be done first. In this current work, the GA based FCM (GAFCM) clustering is initially used on the dataset for removing the outliers, thus facilitating the data undersampling. The FCM clustering has been used due to its ability of handling the overlapping cluster boundaries. However, the main challenge of FCM lies in the random initialization of cluster centers in its local optima (Bezdek et al., 1984). Therefore, the GA based optimization technique has been applied on the FCM to make the clustering more robust by searching for the cluster centers in global optima. The suspicious claims are then identified among the insurance records and their behavior is further verified by four different supervised classifiers.

3. Algorithms used in the proposed approach

In order to understand the training and the fraud detection process of the proposed AIFDS, the working principle of the algorithms used in the current work has been briefly summarized. Since the techniques like, GA (Eiben et al., 1994), SVM (Cortes and Vapnik, 1995), MLP (Rosenblatt, 1961) and DT (Quinlan, 1987) are too well-known for an introduction, some basics regarding FCM and GMDH has been presented in the following subsection.

3.1. Fuzzy C-Means clustering

Fuzzy C-Means (FCM) clustering technique tries to find meaningful clusters present in a dataset by assigning some membership values in the range of [0, 1]. The objective function of FCM can be represented as follows (Bezdek et al., 1984):

$$J_m(U,V;D) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik}^m)B_{ik}(v_i,d_k) \quad (1)$$

subject to $\sum_{i=1}^{c}u_{ik} = 1\ \forall\ k$, and $0 \leqslant u_{ik} \leqslant 1$. J_m is the objective function and the weighting exponent $m > 1$ is responsible for the fuzzy overlap between clusters. $U = [u_{ik}]$ is presented as the membership matrix, $D = \{d_1, d_2 d_n\}$ refers to the dataset with n points on which clustering is to be performed. $V = \{v_1, v_2 v_c\}$ denotes a vector of c cluster centroids, while $B_{ik}(v_i, d_k)$ signifies the distance between d_k and v_i. Whenever FCM is employed on a dataset with the required number of clusters (c) as input, it generates a fuzzy membership matrix (U) and a cluster center set (V). Besides, a low membership value is assigned to outliers that are distant from the cluster centers. The FCM clustering algorithm has been successfully applied in different applications such as signal analysis (Łęski and Owczarek, 2005), image segmentation (Park, 2010), gene expression (Mukhopadhyay and Maulik, 2009), fraud detection (Xue et al., 2010; Wang et al., 2010; Zhang and Gu, 2016) and many more.

3.2. Group method for data handling

The Group Method for Data Handling (GMDH) classifier is a self-organized inductive supervised learning algorithm used for modeling complex nonlinear systems (Ivakhnenko, 1968). This algorithm tries to establish a quadratic polynomial relationship between input and output variables in the training dataset iteratively in order to minimize the error generated during prediction (difference between predicted value and actual value). The quadratic representation of GMDH model can be given as follows:

$$y = t_0 + t_1 d_1 + t_2 d_2 + t_3 d_1^2 + t_4 d_2^2 + t_5 d_1 d_2 \quad (2)$$

where, y represents the output node, $t = \{t_0 \cdots t_5\}$ is a coefficient vector and d_1 and d_2 are input points. The GMDH takes the dataset through an input layer, while the second layer elements are generated from the first layer by initially estimating the regressions of the inputs and then selecting the optimal ones. Likewise, the next layer is designed from the elements of previous layer and so on, thus selecting the best values for processing in subsequent layer. Finally, the generated output of the GMDH model (y) contains only the optimum value which has the minimum prediction error.

4. Proposed approach

In this work, a novel hybrid AIFDS has been developed that effectively handles the class imbalance problem and also reduces the misclassification error. Initially, a test set is extracted from the original unbalanced insurance dataset. The proposed system then applies an undersampling approach on the imbalanced train data points by eliminating the outliers present in the train set after applying the GA based FCM (GAFCM) clustering. During the fraud detection process, the test set is subjected to the GAFCM clustering module which marks the points as genuine, suspicious or malicious. The legitimate and fraudulent points are discarded and the suspecting instances are further analyzed by some supervised classifiers individually for accurate classification. The procedure involved in the training and fraud detection phase have been elaborated in following subsections.

4.1. Training phase

As discussed earlier in Section 1, reducing the skewed class distribution present in the dataset is essential as it affects the efficiency of an AIFDS. In the current work, the FCM clustering technique has been employed on the majority class (genuine) samples in the original unbalanced train set as an undersampling approach. This is achieved by removing the noisy points after generating some meaningful clusters. But since the performance of FCM is influenced by the arbitrary initialization of cluster centers, the GA is used on the cluster centers of FCM for enhancing its search space globally, thus helping FCM to overcome its vulnerability. Fig. 1 presents the work flow of the proposed undersampling method.

Initially, the 10-fold cross validation method (Refaeilzadeh et al., 2009) is used on the major class samples of the imbalanced train set for identifying and removing noisy points from the set. This method randomly divides the original train set into 10 subsamples, out of which 9 subsets are combinedly used for training and the remaining subset is considered for validation. The results from each fold are then averaged to yield the final result.

In order to facilitate the optimization operation, some parameters needed for GA are initially set. The length of genomes (l) is selected to be the number of features in the training set, while the cluster center matrix (V) is chosen as size $c \times l$ with c rows and l columns respectively denoting c as the number of clusters. Each point of V matrix is mapped into strings of 0's and 1's of length l and the center (v) is being updated iteratively as follows (Bezdek and Hathaway, 1994):

$$v_j = \frac{\sum w_{i=1}^{n} u_{ij}^m . d_i}{\sum_{i=1}^{n} u_{ij}^m} \quad (3)$$

where, n signifies the number of data points present in the dataset, m measures the fuzzifier exponent assigned to each point d_i and u_{ij} denotes the elements of fuzzy membership matrix (U). Likewise, the U matrix is also updated in each iteration as follows:

$$u_{ik} = 1 \Big/ \sum_{j=1}^{c}\left[\frac{B_{ik}(v_i,d_k)}{B_{jk}(v_j,d_k)}\right]^{1/(m-1)} \quad for\ 1 \leqslant i \leqslant c\ \ and\ \ 1 \leqslant k \leqslant n \quad (4)$$

where, B_{ik} represents any distance measure between cluster center v_j and data instance d_k. The GAFCM update the cluster centers and

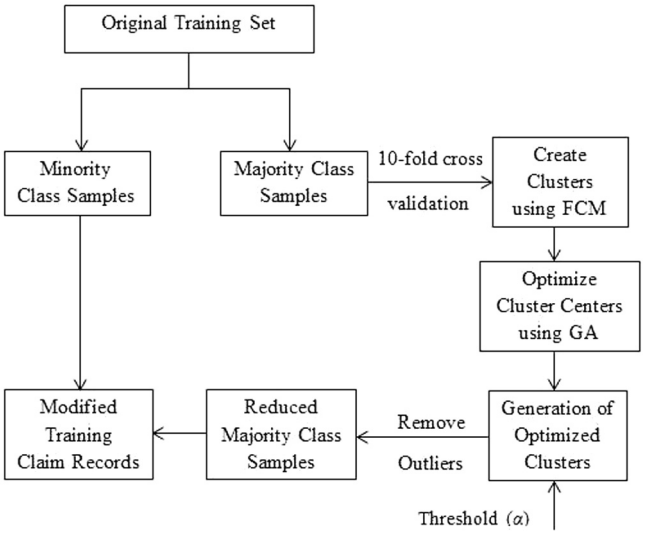

Fig. 1. Proposed Undersampling Approach using GAFCM.

membership values of each data point iteratively according to Eq. (3) and Eq. (4) respectively so that the cost of the fitness function (Eq. (1)) can be minimized. The *Euclidean* distance measure (e) is used for computing the distance measure B_{ik} between a cluster center (v_i) and a data point (d_i) with n instances, which can be calculated as follows:

$$e_i = \sqrt{\sum_{i=1}^{n} v_i - d_i} \tag{5}$$

Initially, the FCM tries to place the data in a cluster and allocate a fuzzy membership value (m) to it where, $m \to 1$ denotes higher affinity towards a cluster, while $m \to 0$ indicates lesser similarity. The AIFDS estimates the euclidean distance (e) of the instance from the cluster centers by using Eq. (5). The computed distance is compared with respect to a threshold value (α), which has been determined by the Tukey method for threshold detection (Tukey, 1977). Initially, this technique sorts the distance values in ascending order and then segregates into four quarters defined by Q_1 (1st quartile), Q_2 (2nd quartile) and Q_3 (3rd quartile). The threshold value is computed by using these quartiles presented as below:

$$\alpha = Q_3 + 3\|Q_3 - Q_1\| \tag{6}$$

The corresponding data point is marked as an outlier, if $e > \alpha$ holds true. Subsequently, the outliers are discarded from the majority class samples of the original imbalanced train set resulting in generation of a reduced train set. The modified major class instances are then combined with the minority class points to produce balanced train claim records.

4.2. Fraud detection phase

Once the class imbalance problem is resolved, the proposed AIFDS detects the fraudulent claims in two stages. The steps involved in the identification of fraudulent claims have been depicted in Fig. 2.

When a test claim record is given to the AIFDS, it computes the euclidean distance (e) from the cluster centers (using Eq. (5)). A first stage decision is made on the record according to the result of comparison of the distance value against two threshold values β_L and β_U. These two limits are determined by the Tukey method (Tukey, 1977). The upper threshold (β_U) is estimated by using Eq. (6), while the lower threshold (β_L) is determined as follows:

$$\beta_L = Q_1 - 3\|Q_3 - Q_1\| \tag{7}$$

The discrimination of new insurance record is done as follows:

1. If $e < \beta_L$, the claim is labeled as genuine.
2. If $e > \beta_U$, the instance is marked as fraudulent.
3. If $\beta_L \leqslant e \leqslant \beta_U$, the record is identified as suspicious.

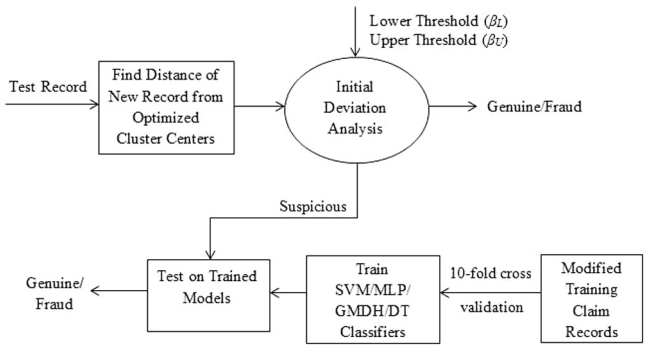

Fig. 2. Proposed Two-stage Fraud Detection Approach.

The claims labeled as genuine are notified to the company employer for payment clearance, while necessary precautionary steps are taken for the illegitimate cases. The suspicious instances are passed to the four different trained supervised classifier models individually for further verification.

Second stage decision making is done by analyzing the behavior of the suspicious insurance claims by the trained supervised models. In this work, four different classifiers – DT, SVM, GMDH and MLP have been used. Initially, the balanced train set is given to each classifier for learning and building corresponding trained model. The 10-fold cross validation method is applied during the training and validation of the classifier models. Upon submitting the suspicious samples to the validated models individually, a final decision (genuine/malicious) regarding each suspicious instance is made. Further, the performance of all the supervised learners are analyzed and compared in order to obtain the best classification accuracy and minimizing the misclassification error.

5. Experimental results and analysis

The proposed system has been implemented in MATLAB 8.3 on a 2.40 GHz i5 CPU system. Extensive experimentation are done for determining the optimal cluster centers for GAFCM clustering as well as for showing effectiveness of four classifiers. The efficacy of the proposed AIFDS has been demonstrated by testing with a real world automobile insurance labeled dataset (Phua et al., 2004).

The following standard performance metrics – *Sensitivity*, *Specificity* and *Accuracy* are used to measure the effectiveness of the proposed system. *Sensitivity* denotes the ratio of truly positive samples that are correctly classified by the classifier. *Specificity* presents the fraction of correctly detected true positive samples and true negative samples, while *Accuracy* estimates the correctness of a classifier. The model with the highest *Sensitivity* value has been chosen as the optimal one, since *Sensitivity* measures the efficiency of a classifier by recognizing more number of fraudulent samples.

$$Sensitivity = \frac{TP}{TP + FN} \tag{8}$$

$$Specificity = \frac{TN}{TN + FP} \tag{9}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

where, *TP* denotes true positive, *FN* stands for false negative, *FP* indicates false positive and *TN* refers to true negative.

5.1. Dataset description and preprocessing

To evaluate the efficiency of the proposed system, we have applied a labeled automobile insurance dataset popularly known as "carclaims.txt". It is found from study of the literature that this is the only publicly available fraud dataset in this domain. Since the proposed AIFDS is based on a supervised fraud detection model, the labels in the dataset are useful for performance comparison. The dataset comprises of various insurance cases filed during 1994–1996 in the United States have been used (Phua et al., 2004). The dataset contains 15,420 records having 14,497 genuine samples (94%) and 923 fraud instances (6%). For experimentation, the data of the year 1996 are considered as the test set with 4,082 instances, while the claims of 1994–95 are taken as the training set consisting of 11,337 data points (Phua et al., 2004).

Due to the publicly availability of the dataset, various researchers have successfully applied it for exhibiting their system's performance (Xu et al., 2011; Sundarkumar and Ravi, 2015; Sundarkumar

Table 2
Comparative Performance Analysis of AIFDSs using *carclaims.txt*.

Research Articles	Performance Metrics (in %)		
	Accuracy	Sensitivity	Specificity
Xue et al. (2010)	88.70	–	–
Sundarkumar et al. (2015)	58.92	95.52	56.58
Sundarkumar and Ravi (2015)	60.31	90.79	58.69
Nian et al. (2016)	–	91.00	52.00

et al., 2015; Nian et al., 2016). Table 2 presents a brief comparative performance analysis of these research articles that have used the same dataset in terms of the performance metrics – *Sensitivity*, *Specificity* and *Accuracy*.

Before subjecting the raw dataset to the AIFDS, some data preprocessing steps are taken by following the procedures suggested in Phua et al. (2004). After the data cleaning is over, the newly modified features are mapped to numerical values, since the fraud detection model requires integers for analysis. Then the data normalization procedure is employed on the dataset for normalizing them in [0, 1] range. This is done to ensure that every data point will get equal opportunity rather than the high-valued attributes while subjecting to the AIFDS. Initially, the original train set contains 10,627 legitimate and 710 fraud points leading to a class imbalance ratio of 0.94: 0.06. An undersampling approach is used on the genuine samples with the help of GAFCM clustering technique for reducing the data imbalance issue. The determination of essential GAFCM parameters has been shown in following subsections.

5.2. Determination of FCM parameters

Initially, experiments are carried out for determining the correct number of clusters (c) for FCM clustering as mentioned in Section 3.1. In order to find the required cluster number, two fuzzy validity indices – Partition Coefficient (PC) and Partition Entropy (PE) are used (Pal and Bezdek, 1995). The PC computes an average of membership value shared in between each fuzzy subset pair inside the membership matrix (U).

$$PC = 1/n \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^2 \tag{11}$$

where, clustering is done on n points with c number of clusters and u_{ij} is the membership value assigned to each instance. The optimal cluster center ($c^{\dagger}$) can be found by solving $max_{2 \leqslant c \leqslant n-1} PC$. Similarly, PE measures the quantity of fuzziness present in the U matrix, which can be described as:

$$PE = -1/n \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij} log\, u_{ij} \tag{12}$$

The optimal cluster number ($c^{\dagger}$) can be derived as well using $min_{2 \leqslant c \leqslant n-1} PE$. Bezdek's suggestion for selecting the best range of cluster number (Bezdek et al., 1984) has been considered. The optimal cluster number $c^{\dagger}$ has been presented in bold in Table 3 for better visualization.

For $c = 2$, both PC = 0.5 (maximum) and PE = 0.6931 (minimum) exhibit the best result. Hence, we have chosen $c = 2$ for clustering in rest of the experiment. Moreover, another two cluster validity measures – Fuzziness Performance Index (FPI) and Normalized Classification Entropy (NCE) has been used (Odeh et al., 1992). The FPI measures the degree of membership shared between each class, which can be calculated as follows:

$$FPI = 1 - \frac{(c * PC - 1)}{c - 1} \tag{13}$$

where, PC is the partition coefficient estimated from Eq. (11). Likewise, the NCE value decides how many clusters are suitable for an effective grouping, which can be computed as follows:

$$NCE = \frac{PE}{log\, n} \tag{14}$$

where, PE indicates the partition entropy as shown in Eq. (12). The more distinct partition of clusters can be found for smaller values of FPI and NCE.

Once the determination of correct cluster number required for FCM is over, experiments regarding cluster center optimization by applying GA are carried out. The functional parameters – *population size* = 50, mutation rate = 0.02, crossover rate = 0.8 and maximum iteration = 100 has been set for GA. Table 4 presents the results of optimized cluster centers (v_1 and v_2) with respect to FPI and NCE values. The best result is obtained at Run 5 with minimum FPI and NCE, which has been highlighted in bold. Hence, the cluster centers v_1 and v_2 of Run 5 has been chosen as the optimized one.

For finding out the optimal cluster center, the performance of the objective function of GA over 100 iterations has been presented in Fig. 3. It is clear from the figure that, the value remains constant after around 92^{th} iteration.

After the optimization process is over, the euclidean distance (e) between the majority class points of the original imbalanced train set and the optimized cluster centers are computed (using Eq. (5)). In order to find and remove the noisy instances from the train set, a threshold value (α) has been set by following the Tukey method as discussed in Section 4.1. The quartile (Q_1 and Q_3) values needed for threshold computation has been found as: $Q_1 = 0.5433$ and $Q_3 = 0.6217$. Finally, the value of $\alpha = 0.8569$ has been estimated by using Eq. (6). The point corresponding to the distance value is marked as noise for $e > \alpha$.

Initially, the amount of genuine samples in the train set was 10,627. But after employing GAFCM, 4,773 instances are detected as noisy and removed in the form of outliers, thus reducing to

Table 4
Results Obtained From GA based FCM.

Run	v_1	v_2	FPI	NCE
1	0.5239	0.5476	−4.7995 e+3	4.7679 e+3
2	0.4269	0.6178	−4.8521 e+3	4.7297 e+3
3	0.3843	0.4742	−4.9345 e+3	4.6694 e+3
4	0.6267	0.3656	−4.8389 e+3	4.7393 e+3
5	**0.4973**	**0.4329**	**−4.9775 e+3**	**4.6373 e+3**
6	0.5384	0.4256	−4.8778 e+3	4.7110 e+3
7	0.4261	0.6057	−4.8208 e+3	4.7524 e+3
8	0.5731	0.5716	−4.8467 e+3	4.7336 e+3
9	0.5306	0.4545	−4.8563 e+3	4.7266 e+3
10	0.5004	0.3723	−4.8433 e+3	4.7360 e+3

Table 3
Determination of Number of Cluster.

c	2	3	4	5	6	7	8	9	10
PC	**0.5**	0.3333	0.2500	0.2000	0.1667	0.1429	0.1250	0.1111	0.1000
PE	**0.6931**	1.0986	1.3863	1.6094	1.7918	1.9459	2.0794	2.1972	2.3026

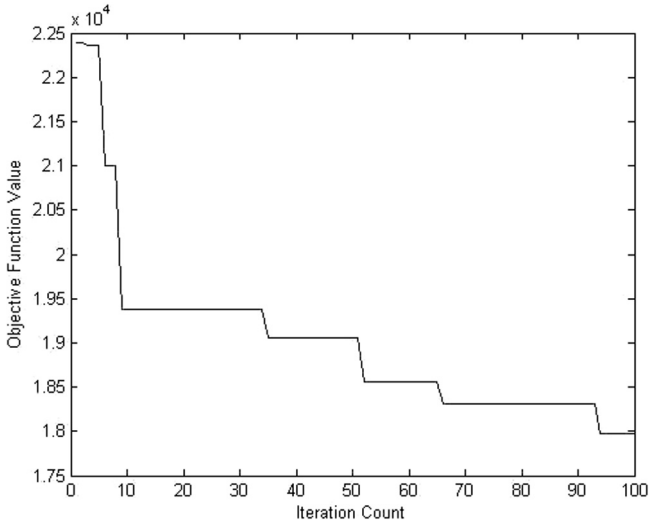

Fig. 3. Optimization of Fitness Function of Genetic Algorithm over 100 Iterations.

5,854 claims. After combining the 710 points belonging to minority (fraud) class samples with the reduced major class, the modified train set of size 6,564 has been generated.

During the fraud detection phase, when the test set of size 4,082 is given to the clustering module, initially the distances are estimated from the cluster centers (using Eq. (5)). These distance values are then compared with the thresholds β_L and β_U for segregating the test records into genuine, malicious or suspicious classes in the first stage. The quartile values needed for calculating the limits are found to be $Q_1 = 0.3783$ and $Q_3 = 0.4597$. This leads to $\beta_U = 0.7039$ and $\beta_L = 0.1341$ estimated by Eq. (6) and Eq. (7) respectively. The test set has been discriminated into 113 fraudulent samples, 2,028 genuine instances and 1,941 suspicious records in the first stage. The experimental results based on the specified parameters are described in the following subsection.

5.3. Performance analysis of the proposed system

Table 5 presents the efficacy of FCM and GAFCM without the use of supervised learners in both the balanced and imbalanced dataset. It is evident from the table that the performance of both the clustering techniques has been improved in case of balanced dataset, while GAFCM outperforms FCM in terms of all performance metrics in both type of dataset.

After identifying the suspicious instances in the first stage is over, further verification and classification of these points are done by employing four supervised classifiers – SVM, MLP, DT and GMDH individually. Essential parameters needed for tuning the performance of these classifiers are set. For SVM, the *kernel type* = *rbf*, kernel scale = 1, *iteration* = 1000 and regularization parameter = 1 has been chosen. The parameters – minimum number of leaf size = 1, split criterion = *gdi* (Gini's diversity index) and minimum split size = 10 have been selected for functioning of DT. The relevant parameters for MLP are hidden layer size = 3, nodes per hidden layer = 8, training function = *trainlm*, performance function = *crossentropy*, activation function = *tansig* for hidden layer and *softmax* for output layer and maximum iteration = 1000. The GMDH parameters have been set as: maximum number of hidden layer = 3 and maximum number of neurons in a layer = 8.

Once the functional parameters are set for each learner, the classification of suspicious instances is done on previously trained classifier models individually. A comparative performance analysis of each classifier with and without clustering on the unbalanced insurance dataset has been presented in Table 6. The output of best performing classifier has been highlighted in bold for better visualization. The results in the table clearly show the effectiveness of using GAFCM over normal FCM for classification. The MLP and GMDH yielded 0% *Sensitivity* without using clustering due to the skewed class distribution present in the insurance dataset. The MLP gives the maximum *Sensitivity* = 73.35% after using FCM clustering on the imbalanced dataset, while SVM produces the best efficiency in terms of highest *Sensitivity* = 69.70% and *Specificity* = 84.71%.

Similarly, the performance analysis of each classifier with respect to balanced dataset has been presented in Table 7. The output of the best performing classifier has been highlighted in bold. The MLP produces the highest *Sensitivity* = 75.75% when using FCM as data balancing technique, while DT gives the maximum *Accuracy* = 71.79% and *Specificity* = 73.19%. When using the GAFCM as an undersampling approach, the SVM outperforms all other classifiers in terms of all performance metrics.

5.4. Comparative analysis

In this section, a comparative analysis of the proposed system has been done with another automobile insurance fraud detection approach suggested by considering the same insurance dataset

Table 5
Performance Analysis of FCM and GAFCM.

Performance Metrics (in %)	Unbalanced Dataset	Balanced Dataset		
	FCM	GAFCM	FCM	GAFCM
Sensitivity	56.06	**61.54**	59.22	**66.67**
Specificity	76.85	**84.79**	84.49	**86.95**
Accuracy	72.56	**83.22**	81.97	**84.34**

Table 6
Performance of Supervised Classifiers on Imbalanced Dataset.

Techniques Used	Performance Metrics (in %)								
	Without Clustering			Using FCM			Using GAFCM		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
DT	6.94	99.83	94.39	60.23	65.35	64.28	66.25	87.65	86.99
SVM	**70.76**	63.20	63.65	55.79	68.12	65.51	**69.70**	84.71	83.16
MLP	0	100	94.03	**73.35**	63.12	65.02	61.07	84.00	81.45
GMDH	0	100	94.03	52.98	66.07	63.02	57.27	79.76	77.10

Table 7
Performance of Supervised Classifiers on Balanced Dataset.

Techniques Used	Performance Metrics (in %)					
	Using FCM			Using GAFCM		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
DT	67.81	73.19	71.79	77.42	81.16	80.16
SVM	64.34	72.23	70.15	**83.21**	88.45	87.02
MLP	**75.75**	66.53	68.81	81.12	82.56	82.38
GMDH	60.17	71.61	68.32	65.04	83.44	77.23

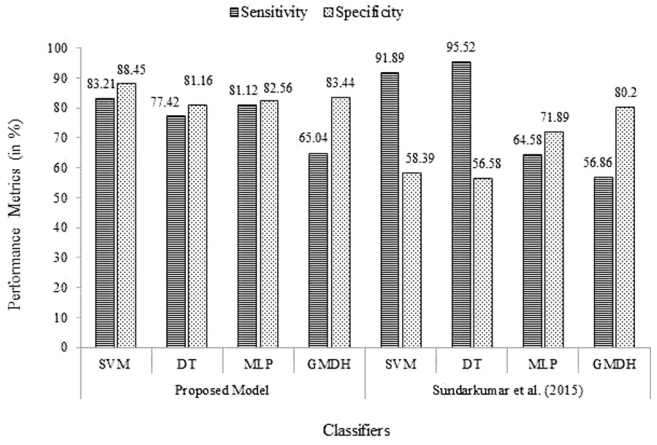

Fig. 4. Comparative Performance Analysis on Insurance Dataset.

"carclaims.txt" (Sundarkumar et al., 2015). The authors in Sundarkumar et al. (2015) have segmented the dataset into train, test and validation set. Initially, a validation set of size 20% of the original dataset has been kept aside and the rest 80% data are subjected on their system. The 10-fold cross validation procedure has been employed on the remaining 80% dataset for segregating into train and test set.

The authors (Sundarkumar et al., 2015) have used One Class Support Vector Machine (OCSVM) for removing the skewed class distribution from the dataset. The OCSVM has been applied on the genuine samples of the train set for extracting the support vectors leading to a reduced size of normal class data. The fraudulent instances (minor) are then combined with the major class set to form a modified training set. Four different supervised classifiers – SVM, MLP, GMDH and DT have been applied on the modified train set independently for generating their corresponding trained model. The test set is employed on the respective trained models for testing the efficacy of the models. Finally, the validation set is used for validating each model.

The performance metrics – *Sensitivity* and *Specificity* are used for comparative performance analysis. A good classifier should necessarily exhibit higher values of *Sensitivity* as well as *Specificity* as they indicate accuracy of classification of forged instances and lowered false alarms respectively. After following the experimental procedure described in Sundarkumar et al. (2015), the results obtained for the proposed model and the approach under comparison has been presented in Fig. 4. It can be inferred from the figure that the proposed model effectively minimizes the misclassification rate by giving the highest *Specificity* in all classifiers as compared to Sundarkumar et al. (2015). Moreover, the proposed model produces the highest *Sensitivity* = 83.21% and *Specificity* = 88.45% upon using SVM as the classifier, whereas, in case of Sundarkumar et al. (2015), DT produces the best performance in terms of *Sensitivity* = 95.52% and *Specificity* = 56.58%.

6. Conclusions

In this research, we have proposed a novel hybrid approach for automobile insurance fraud detection that proceeds in two phases – training and fraud detection. In the training phase, a GA based optimized FCM (GAFCM) clustering has been used for undersampling the majority class samples in the highly skewed train dataset so as to improve the efficiency of the classifiers. Initially, the GAFCM clustering is employed on the majority class instances for generating clusters with optimal cluster centers. The outliers as well as redundant data points present in the majority class are then identified and removed, thus facilitating undersampling. The reduced majority class samples are then combined with the original minority class points to obtain a balanced dataset, which is used for further experimentation.

The fraud detection procedure is carried out in two stages in the proposed system. During the first stage of fraud detection, the GAFCM categorizes the test data points as genuine, malicious and suspicious classes based on their distance measure from the optimized cluster centers. The samples identified as genuine and fraudulent are not further processed, while the suspicious ones are additionally verified in the second stage by four different supervised learners – DT, SVM, MLP and GMDH individually.

A labeled automobile insurance dataset popularly known as "carclaims.txt" has been used for measuring the efficiency of the proposed system. Initially, the dataset contains 15,420 records having 11,337 training samples and 4082 test samples. The 10-fold cross validation technique is used throughout the experiments for training and validation of the clustering as well as the classifier models. Experiments were carried out for selecting the optimal number of clusters required for FCM. Extensive tests were further conducted for the generation of optimized cluster centers by employing GAFCM. Before undersampling, the training sample of size 11,337 comprised of 10,627 genuine samples (major) and 710 fraud samples (minor). Upon applying GAFCM on the majority class samples of training set for undersampling, 4773 were identified as noisy points and removed in the form of outliers, thus reducing to 5854 instances. Finally, these reduced majority class points are combined with the 710 illegitimate minor samples to produce a modified dataset of size 6564 claims.

Extensive tests were further carried out for measuring the effectiveness of the proposed model with respect to the modified train set of 6,564 records and original test set of size 4,082 samples. While making the first stage decision, the GAFCM produces an amount of 113 forged instances, 2,028 genuine samples and 1,941 suspicious claims. In the second stage classification, the 1,941 suspicious records were passed for accurate classification. It is found that SVM outperforms all other classifiers with the highest 88.45% Specificity and 83.21% Sensitivity. Besides, comparative analysis with another recent AIFDS demonstrates the effectiveness of the proposed system in terms of lowered false alarms while simultaneously controlling the imbalanced class distribution and efficient identification of fraudulent cases.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Abdallah, A., Maarof, M.A., Zainal, A., 2016. Fraud detection system: a survey. J. Network Comput. Appl. 68, 90–113.

Australia: Insurance, April 2016. Australia: Insurance fraud costs us 1.5 bln annually.http://www.insurancefraud.org/IFNS-detail.htm?key=22516 (accessed: 9.05.17).

Bermúdez, L., Pérez, J., Ayuso, M., Gómez, E., Vázquez, F., 2008. A bayesian dichotomous model with asymmetric link for fraud in insurance. Insurance: Math. Econ. 42 (2), 779–786.

Bernard, C., Vanduffel, S., 2014. Mean–variance optimal portfolios in the presence of a benchmark with applications to fraud detection. Eur. J. Oper. Res. 234 (2), 469–480.

Bezdek, J.C., Ehrlich, R., Full, W., 1984. Fcm: The fuzzy c-means clustering algorithm. Comput. Geosci. 10 (2–3), 191–203.

Bezdek, J.C., Hathaway, R.J., 1994. Optimization of fuzzy clustering criteria using genetic algorithms. In: Evolutionary Computation, 1994. IEEE World Congress on Computational Intelligence., Proceedings of the First IEEE Conference on. IEEE, pp. 589–594.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297.

Cutting corners, August 2015. Cutting corners to get cheaper motor insurance backfiring on thousands of motorists warns the abi.https://www.insurancefraudbureau.org/media-centre/news/2015/cutting-corners-to-get-cheaper-motor-insurance-backfiring-on-thousands-of-motorists-warns-the-abi/ (accessed: 9.05.17).

Eiben, A.E., Raue, P.-E., Ruttkay, Z., 1994. Genetic algorithms with multi-parent recombination. In: International Conference on Parallel Problem Solving from Nature. Springer, pp. 78–87.

Ivakhnenko, A.G., 1968. The group method of data handling-a rival of the method of stochastic approximation. Sov. Autom. Control 13 (3), 43–55.

Jensen, D., 1997. Prospective assessment of ai technologies for fraud detection: A case study. In: AAAI Workshop on AI Approaches to Fraud Detection and Risk Management. pp. 34–38.

Lee, Y.-J., Yeh, Y.-R., Wang, Y.-C.F., 2013. Anomaly detection via online oversampling principal component analysis. IEEE Trans. Knowl. Data Eng. 25 (7), 1460–1470.

Łęski, J.M., Owczarek, A.J., 2005. A time-domain-constrained fuzzy clustering method and its application to signal analysis. Fuzzy Sets Syst. 155 (2), 165–190.

Mukhopadhyay, A., Maulik, U., 2009. Towards improving fuzzy clustering using support vector machine: application to gene expression data. Pattern Recogn. 42 (11), 2744–2763.

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

Nian, K., Zhang, H., Tayal, A., Coleman, T., Li, Y., 2016. Auto insurance fraud detection using unsupervised spectral ranking for anomaly. J. Finance Data Sci. 2 (1), 58–75.

Odeh, I., Chittleborough, D., McBratney, A., 1992. Soil pattern recognition with fuzzy-c-means: application to classification and soil-landform interrelationships. Soil Sci. Soc. Am. J. 56 (2), 505–516.

Pal, N.R., Bezdek, J.C., 1995. On cluster validity for the fuzzy c-means model. IEEE Trans. Fuzzy Syst. 3 (3), 370–379.

Panigrahi, S., Sural, S., Majumdar, A.K., 2013. Two-stage database intrusion detection by combining multiple evidence and belief update. Inf. Syst. Front. 15 (1), 35–53.

Park, D.-C., 2010. Intuitive fuzzy c-means algorithm for mri segmentation. In: Information Science and Applications (ICISA), 2010 International Conference on. IEEE, pp. 1–7.

Pathak, J., Vidyarthi, N., Summers, S.L., 2005. A fuzzy-based algorithm for auditors to detect elements of fraud in settled insurance claims. Managerial Auditing J. 20 (6), 632–644.

Phua, C., Alahakoon, D., Lee, V., 2004. Minority report in fraud detection: classification of skewed data. Acm Sigkdd Explor. Newslett. 6 (1), 50–59.

Pinquet, J., Ayuso, M., Guillén, M., 2007. Selection bias and auditing policies for insurance claims. J. Risk Insurance 74 (2), 425–440.

Quinlan, J.R., 1987. Simplifying decision trees. Int. J. Man-mach. Stud. 27 (3), 221–234.

Refaeilzadeh, P., Tang, L., Liu, H., 2009. Cross-validation. In: Encyclopedia of Database Systems. Springer, pp. 532–538.

Rosenblatt, F., 1961. Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Tech. rep., DTIC Document.

Šubelj, L., Furlan, Š., Bajec, M., 2011. An expert system for detecting automobile insurance fraud using social network analysis. Expert Syst. Appl. 38 (1), 1039–1052.

Sundarkumar, G.G., Ravi, V., 2015. A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance. Eng. Appl. Artif. Intell. 37, 368–377.

Sundarkumar, G.G., Ravi, V., Siddeshwar, V., 2015. One-class support vector machine based undersampling: Application to churn prediction and insurance fraud detection. In: Computational Intelligence and Computing Research (ICCIC), 2015 IEEE International Conference on. IEEE, pp. 1–7.

Tao, H., Zhixin, L., Xiaodong, S., 2012. 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. IEEE, pp. 457–460.

Tukey, J.W., 1977. Exploratory data analysis.

Wang, G., Hao, J., Ma, J., Huang, L., 2010. A new approach to intrusion detection using artificial neural networks and fuzzy clustering. Expert Syst. Appl. 37 (9), 6225–6232.

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

Xue, Z., Shang, Y., Feng, A., 2010. Semi-supervised outlier detection based on fuzzy rough c-means clustering. Math. Comput. Simul. 80 (9), 1911–1921.

Zhang, Z., Gu, B., 2016. Intrusion detection network based on fuzzy c-means and particle swarm optimization. In: Proceedings of the 6th International Asia Conference on Industrial Engineering and Management Innovation. Springer, pp. 111–119.