ISSN: 2456-9348
Vol(01) _Issue (03)

iJETRM

International Journal of Engineering Technology Research & Management

Automobile Insurance Fraud Detection System

Kadiri Guru Preetam
Nalini Nagendran
VIT University, Tamil NaduVellore-632014, India

ABSTRACT

This paper exhibits a approach for identifying frauds in automobile insurance claims by applying Genetic Algorithm (GA) based Fuzzy C-Means (FCM) grouping and different supervised classifier models. At first, a test set is taken from the available insurance dataset. The rest of the train set is subjected to the clustering technique for under sampling in the wake of creating some meaningful clusters. The test cases are then isolated into genuine, malicious or suspicious classes in the wake of subjecting to the clusters. The genuine and fraudulent records are disposed of, while the suspicious cases are additionally broke down by four classifiers – Decision Tree (DT), Support Vector Machine (SVM), Group Method of Data Handling (GMDH) and Multi-Layer Perceptron (MLP) exclusively. The 10-fold cross validation technique is utilized all through the work for training and validation of the models. The adequacy of the proposed framework is outlined by directing a few tests on a certifiable automobile insurance dataset.

KEYWORDS:

Fuzzy C-Means Clustering, Support Vector Machine, Clustering, Data Centres

INTRODUCTION

An automobile insurance contract between a client (customer) and an insurer (company) gives money related help if there should be an occurrence of vehicular harm or robbery. Automobile insurance fraud emerges after submitting counterfeit records in regards to setbacks in an arranged mishap or cases for past misfortunes keeping in mind the end goal to acquire money related benefits (Ngai et al., 2011). In addition, this sort of extortion can be done by anybody like, insured's, chiropractors, mechanics, legal counsellors, cops, insurance employees and others (Šubelj et al., 2011). An investigation done by Insurance Fraud Authority of Australia in 2013 mirrors the rising pattern in ill-conceived assert costs, which is $2 billion more than in 2012 (Australia: Insurance, 2016). In 2014, the Association of British Insurers (ABI) researches the expansion in number of false claims, which is 18% more than the earlier year (Cutting corners, 2015). These insights obviously shows the seriousness of the issue and henceforth, should be tended to immovably to lessen the misfortunes acquired by such vindictive endeavours.

Moreover, the auto insurance fraud can be arranged into a simpler way (documenting a forged application) or a more tricky way for example, creating a mishap or burglaries (Abdallah et al., 2016). Moreover, the despicable portrayal of information concerning asserts makes the misrepresentation recognition to a great degree troublesome (Šubelj et al., 2011). Also, it is watched that exclusive a small portion of mishap claims are ill-conceived prompting the nearness of a skewed class distribution in the dataset. This makes the identification even tougher (Jensen, 1997). Consequently, exact grouping of these deceitful cases are basic for any Automobile Insurance Fraud Detection System (AIFDS). The iterative calculation required for isolating the genuine occurrences may require high calculation time after being subjected to an AIFDS (Panigrahi et al., 2013). Accordingly, there is a need to build up

iJETRM

International Journal of Engineering Technology Research & Management

a powerful AIFDS that can segregate the noxious examples from the normal insurance claims productively while limiting the misclassification rate.

This paper proposes an AIFDS approach that applies the Genetic Algorithm (GA) for optimizing the clusters created from Fuzzy C-Means grouping (FCM) as an under sampling method. This is done to expel the noisy points from the majority class tests of the unbalanced dataset prompting a reduced adjusted dataset. Another protection guarantee is then characterized as honest to goodness, noxious or suspicious in light of its separation measure registered from the optimized cluster centres. The claim set apart as certifiable is permitted to go through for instalment preparing, while the claim observed to be fraudulent is blocked. In the event that the claim is a suspicious one, at that point advance confirmation and grouping is made by going it through four diverse prepared managed learning models independently. The preparation of the classifiers is completed apriori by applying the adjusted preparing dataset. In this work, Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), Decision Tree (DT) and Group Method of Data Handling (GMDH) classifiers are utilized for deciding the best performing classifier.

RELATED WORK

In this section, the research work completed with importance to automobile insurance fraud identification is looked into. A hybridization of stacking and packing Meta classifiers has been submitted in Phua et al. (2004). The stacked troupe at first picks the best classifier from a group of base learners. At that point, the bagging system is utilized on the chosen classifier for prescient examination of an oversampled true marked dataset. Another approach recommends the utilization of fluffy logic idea for finding the ill-conceived claims from a cluster of settled protection claims (Pathak et al., 2005). With a specific end goal to investigate and distinguish the suspicious claims from the protection records, the creators have built up a factual bivariate probit display as an examining methodology utilized on a Spanish automobile insurance dataset (Pinquet et al., 2007). A skewed Bayesian dichotomous legit display has been proposed for distinguishing malevolent protection claims found in a Spanish car market (Bermúdez et al., 2008).

The use of graph based social network model has been displayed in Šubelj et al. (2011), which needs unlabelled information for handling. An Iterative Assessment Algorithm (IAA) has been created by the developers to recognize the suspicious cases. At first, a suspicion score is assigned to each point shown in the graph and after that the assurance of suspicious elements is finished by dissecting the edges which are shown inside their neighbouring nodes. The rough set based neural system group strategy has been proposed in Xu et al. (2011). In this paper, at first the entire dataset space is divided into different subspaces with the assistance of rough set data space reduction strategy. At that point the neural system is connected on these subspaces independently to build trained models. Afterward, the consequences of each model are consolidated utilizing a voting system for official decision making. The idea of fuzzy support vector machine for recognizable proof of suspicious (covered) insurance cases has been proposed in Tao et al. (2012). The fraud detection model at first ascertains a separation estimation of every fraud case concerning two classes of sample mean vector and allots a dual membership participation to them. This makes a difference each malicious sample to be allocated with a probability value utilized for arranging in two classes (genuine or fraud).

The identification of fraudulent cases by utilizing a quantitative approach has been proposed in Bernard and Vanduffel (2014). In this work, a Sharpe ratio and its breaking point values are evaluated by decreasing and increasing the variance and mean values of the claim instalments individually. The identification of fraud tests in the insurance claims are then performed based on these limit values. A fraud detection model has been created in

iJETRM

International Journal of Engineering Technology Research & Management

Sundarkumar and Ravi (2015), which detects and removes exceptions for reducing the class irregularity effect introduce in the automobile insurance dataset. Two unsupervised systems: k-Reverse Nearest Neighbourhood (k-RNN) and One Class Support Vector Machine (OCSVM) are utilized pair for solving the skewed class distribution in the available dataset. Further, six diverse administered classifiers have been connected autonomously on the balanced dataset for order and correlation purposes. In paper (Nian et al., 2016), the authors have recommended the utilization of an unsupervised anomaly recognition model, known as Spectral Ranking Anomaly (SRA) framework, for detection of forged occurrences. This model doles out a level of peculiarity incentive to each claim in the wake of evaluating the first non-principal eigenvector from a Laplacian matrix of the claim records. In the event that the rank is not as much as a pre-set edge, at that point the comparing point is set apart as fraudulent.

In spite of a various AIFDSs created to deal with the fraud detection effectively, there exist some irrelevant data points in the dataset that can diminish the proficiency of a classifier (Lee et al., 2013). Subsequently, the expulsion of these noisy cases from the first imbalanced dataset is required to be done first. In this current work, the GA based FCM (GAFCM) clustering is at first utilized on the dataset for removing the outliers, consequently encouraging the information under sampling. The FCM clustering has been utilized because of its capacity of dealing with the overlapping cluster boundaries. Be that as it may, the fundamental test of FCM lies in the irregular initialization of cluster centres in its nearby optima (Bezdek et al., 1984). Along these lines, the GA based optimization system has been applied on the FCM to make the clustering more robust via searching for the cluster centres in worldwide optima. The suspicious cases are then distinguished among the insurance records and their conduct is additionally confirmed by four unique supervised classifiers.

ALOGITHMS USED IN PROPOSED APPRAOCH

Fuzzy C-Means Clustering

Fluffy C-Means (FCM) clustering system tries to discover significant clusters found in a dataset by giving some membership values in the scope of [0, 1]. The objective function of FCM can be represented as follows (Bezdek et al., 1984):

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

subjected to $\sum_{i=1}^{c} u_{ik} = 1\forall k$ and $0 \leq u_{ik} \leq 1$. J_m is the target function and the weighting example m > 1 is in charge of the fuzzy overlap clusters. U = [u_{ik}] is introduced as the membership matrix, D = {d_1, $d_2 d_n$} alludes to the dataset with n points on which clustering is to be done. V = {v_1, $v_2 v_c$} indicates a vector of c cluster centroids, while $B_{ik}(v_i,d_k)$ means the distance between v_i and d_k. At whatever point FCM is utilized on a dataset with the required number of clusters (c) as input, it creates a fuzzy membership matrix (U) and a cluster centre set (V). Additionally, a low membership value is assigned to outliers that are far off from the cluster centre. The FCM clustering algorithm has been effectively utilized in various applications, for example, image segmentation (Park, 2010), fraud detection (Xue et al., 2010; Wang et al., 2010; Zhang and Gu, 2016), gene expression (Mukhopadhyay and Maulik, 2009), signal analysis (Łe˛ski and Owczarek, 2005).

Group Meta Data Handling

iJETRM

International Journal of Engineering Technology Research & Management

The Group Method for Data Handling (GMDH) classifier is a self-organized inductive supervised learning algorithm utilized for displaying complex nonlinear frameworks (Ivakhnenko, 1968). This algorithm tries to set up a quadratic polynomial relationship between output and input factors in the training dataset iteratively in request to limit the error produced amid prediction (contrast between predicted value and expected value). The quadratic portrayal of GMDH model can be seen as:

$$y = t_0 + t_1d_1 + t_2d_2 + t_3d_1^2 + t_4d_2^2 + t_5d_1d_2$$

where, y speaks to the output node, t = {t0 ... t5} 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 components are produced from the primary layer by at first evaluating the regressions of the data sources and after that choosing the ideal ones. Moreover, the following layer is composed from the components of previous layer and so forth, hence choosing the best value for preparing in resulting layer. At long last, the created yield of the GMDH show (y) contains just the optimum value which has the minimum prediction error.

Support Vector Machines

The support vector network is another learning machine for two-group classification problems. The machine theoretically actualizes the following idea: input vectors are non-linearly mapped to an extremely high dimension feature space. In this element space a linear decision surface is developed. Extraordinary properties of the decision surface guarantees high generalization capacity of the learning machine. The thought behind the support vector organize was already actualized for the confined situation where the training data can be isolated without errors. We here extend this outcome to non-separable training data.

High specialization capacity of support vector networks using polynomial input changes are illustrated. We likewise analyse the execution of the support vector network to different traditional learning algorithms that all participated in a benchmark study of Optical Character Recognition.

A model of two normal distribution population, $N(m_1, \sum_1)$ furthermore, $N(m_2, \sum_2)$ of n dimensional vectors x with mean vectors ml and m2 and co-variance lattices $\sum_1$ and $\sum_2$, and demonstrated that the optimal (Bayesian) solution is a quadratic decision function:

$$F_{sq}(\mathrm{x}) = sign\frac{1}{2}(x-m)^T\sum_1^{-1}(x-m_1) - \frac{1}{2}(x-m_2)^T\sum_2^{-1}(x-m_2) + ln\frac{|\sum 1|}{|\sum 2|}$$

Decision Trees

Numerous frameworks have been created for developing decision trees from accumulations of various examples. In spite of the fact that the choice trees created by these strategies are precise and productive, they frequently endure the inconvenience of over the top unpredictability that can render them boundless to specialists. It is faulty whether misty structures of this kind can be portrayed as learning, regardless of how well they work.

PROPOSED APPROACH

iJETRM

International Journal of Engineering Technology Research & Management

In this work, a hybrid AIFDS has been produced that adequately handles the class imbalance issue and furthermore decreases the misclassification error. At first, a test set is removed from the first unbalanced insurance dataset. The proposed framework at that point applies an under sampling approach on the imbalanced train data points by wiping out the outliers present in the train set after applying the GA based FCM (GAFCM) grouping. Amid the fraud identification process, the test set is subjected to the GAFCM clustering module which denotes the points as genuine, malicious and suspicious. The true and false points are disposed of and the suspecting occurrences are additionally broke down by some supervised classifiers separately for precise calculation. The technique engaged with the training and fraud detection phase have been expounded in following subsections:

Training Phase

As talked about before in Section 1, lessening the skewed class distribution shown in the dataset is basic as it influences the efficiency of an AIFDS. In the present work, the FCM clustering technique has been utilized on the majority class (genuine) tests in the original unbalanced train set as an under sampling approach. This is accomplished by expelling the noisy points in the wake of producing some meaningful clusters. However, since the execution of FCM is affected by the subjective introduction of cluster centres, the GA is utilized on the cluster centres of FCM for upgrading its search space, in this way helping FCM to overcome its weakness. Fig. 1 displays the work workflow of the proposed under sampling technique.

At first, the 10-fold cross validation technique (Refaeilzadeh et al., 2009) is utilized on the significant class samples of the imbalanced train set for recognizing and expelling noisy points from the set. This strategy arbitrarily separates the first train set into 10 subsamples, out of which 9 subsets are combined utilized for training furthermore, the rest of the subset is considered for approval. The outcomes from each fold are then averaged of to yield the final outcome.

With a specific end goal to facilitate the optimization algorithm, a few parameters required for GA are at first, set. The length of genomes (l) is chosen to be the quantity of features in the training set, while the cluster centre matrix (V) is picked as size c x l with l columns and c rows, also separately signifying c as the number of clusters. Each purpose of V matrix is mapped into strings of 0's and 1's of length l and the centre (v) is being refreshed iteratively as (Bezdek and Hathaway, 1994):

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

where, n implies the number of data points in the dataset, m measures the fuzzier exponent allotted to each point, d_i and u_{ij} signifies the components of fuzzy membership matrix (U). In like manner, the U matrix is additionally refreshed in every iteration:

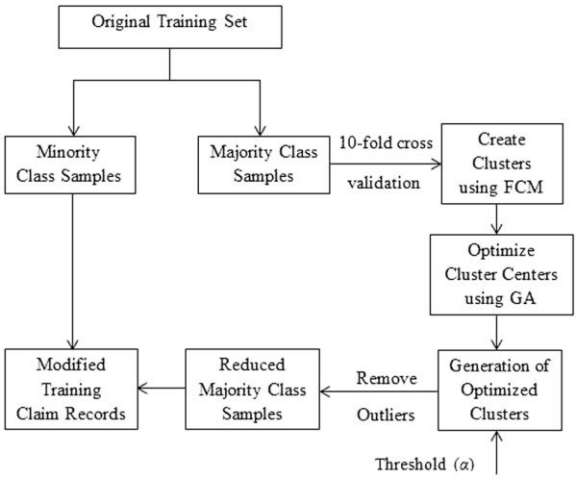
ISSN: 2456-9348
Vol(01) _Issue (03)

iJETRM

International Journal of Engineering Technology Research & Management

$$u_{ik} = 1 \Big/ \sum_{j-1}^{c} \left[\frac{B_{ik}(v_i, d_k)}{B_{jk}(v_j, d_k)}\right]^{1/(m-1)}$$

For 1≤i≤c and 1≤k≤n.



The GAFCM updates the cluster centre and membership value of every data point iteratively concurring in Eq. (3) and Eq. (4) individually so the cost of the fitness function (Eq. (1)) can be limited. The Euclidean distance measure (e) is utilized for registering the distance measure B_{ik} between a cluster centre (v_i) and a data point (d_i) with n occurrences, which can be computed as:

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

At first, the FCM tries to put the information in a cluster and designate a fluffy membership value (m) to it where, m → 1 indicates higher affinity towards a group, while m → 0 demonstrates lesser similarity. The AIFDS estimates the Euclidean distance (e) of the example from the cluster centres by utilizing Eq. (5). The figured separation is analysed as for a threshold value (α), which has been resolved by the Tukey method for threshold detection (Tukey, 1977). At first, this procedure sorts the distance values in ascending order and after that isolates into four quarters characterized by Q_1 (first quartile), Q_2 (second quartile) and Q_3 (third quartile). The edge esteem is figured by utilizing these quartiles displayed as:

$$\alpha = Q_3 + 3||Q_3 - Q_1||$$

The data point is set apart as an outlier, if e > a is true. Therefore, the exceptions are disposed of from the majority class samples of the first imbalanced train set resulting in generation of a reduced train set. The adjusted significant class occurrences are then joined with the minority class points to create balanced train claim records.

ISSN: 2456-9348
Vol(01) _Issue (03)

iJETRM

International Journal of Engineering Technology Research & Management

Fraud Detection Phase

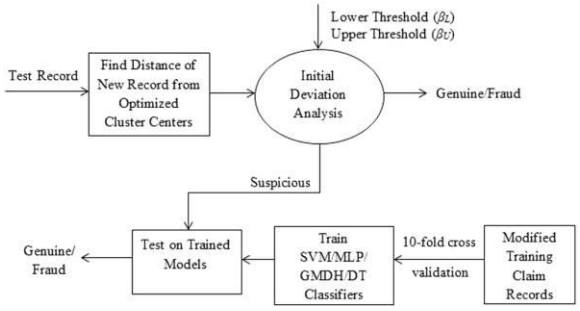
Once the class irregularity issue is settled, the proposed AIFDS identifies the fake claims in two phases. The steps associated with the distinguishing proof of false claims have been portrayed in Fig. 2.

At the point when a test claim record is given to the AIFDS, it processes the Euclidean distance (e) from the cluster centres (utilizing Eq. (5)). A decision is made on the record as indicated by the outcome of correlation of the distance value against two threshold values b_L and b_U. These two limits are controlled by the Tukey strategy (Tukey, 1977). The upper edge (b_U) is assessed by utilizing Eq. (6), while the lower edge (b_U) is resolved as:

$$b_L = Q_1 - 3||Q_3 - Q_1||$$

The segregation of new insurance record is done as takes after:

1. If e $< b_L$, the claim is labeled as genuine.
2. If e $> b_U$, the instance is marked as fraudulent.
3. If $b_L \leq$ e $\leq b_U$, the record is identified as suspicious.



The cases named as certifiable are told to the organization boss for payment clearance, while fundamental preparatory steps are taken for the ill-conceived cases. The suspicious examples are passed to the four different trained supervised classifier models exclusively for further evaluation.

Second stage decision making is finished by examining the conduct of the suspicious insurance claims by the trained supervised models. In this work, four distinct classifiers – DT, SVM, GMDH and MLP have been utilized. At first, the adjusted prepare set is given to every classifier for learning and building relating trained data model. The 10-fold cross approval technique is connected amid the preparing and approval of the classifier models. After submitting the suspicious examples to the approved models independently, a last decision (malicious/genuine) with respect to each suspicious case is made. Further, the execution of all the supervised learners are broken down and contrasted all together with get the best classification accuracy and minimizing the misclassification mistake.

EXPERIMENTAL RESULT AND ANALYSIS

The proposed framework has been actualized in MATLAB 8.3 on a 2.40 GHz i3 CPU framework. Broad experimentation are improved the situation deciding the ideal cluster centres for GAFCM clustering as well with

respect to indicating adequacy of four classifiers. The viability of the proposed AIFDS has been shown by testing with a real world automobile data (Phua et al., 2004).

The standard execution measurements – Sensitivity, Specificity also, Accuracy are utilized to quantify the adequacy of the proposed system. Sensitivity indicates the proportion of truly positive examples that are accurately ordered by the classifier. Specificity introduces the part of accurately distinguished genuine true samples and genuine negative tests, while Accuracy gauges the rightness of a classifier. The model with the most noteworthy Sensitivity esteem has been picked as the ideal one, since Sensitivity measures the proficiency of a classifier by perceiving more number of fraudulent samples.

$$\text{Specifity} = \frac{TN}{FP+TN}$$

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

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

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

Dataset processing and Description

To assess the effectiveness of the proposed framework, we have connected a named automobile insurance dataset famously known as "carclaims.txt". It is found from investigation of the writing that this is the main freely accessible misrepresentation dataset in this area. Since the proposed AIFDS depends on a supervised fraud detection model, the names in the dataset are helpful for performance examination. The dataset contains different insurance cases documented amid 1994– 1996 in the United States have been utilized (Phua et al., 2004). The dataset contains 15,420 records having 14,497 genuine tests (94%) and 923 fraud occurrences (6%). For experimentation, the information of the year 1996 are considered as the test set with 4,082 samples, while the cases of 1994– 95 are taken as the preparation set comprising of 11,337 data points (Phua et al., 2004).
Due to the public accessibility of the dataset, different analysts have effectively connected it for displaying their systems performance (Xu et al., 2011; Sundarkumar and Ravi, 2015; Sundarkumar)

5.2 Performance Analysis of Proposed System

The viability of FCM and GAFCM without the utilization of directed students in both the adjusted and imbalanced dataset. It is obvious from the table that the execution of both the bunching strategies has been enhanced if there should arise an occurrence of adjusted dataset, while GAFCM beats FCM regarding all execution measurements in both kind of dataset.

Subsequent to recognizing the suspicious occurrences in the main stage is over, facilitate check and grouping of these focuses are finished by utilizing four regulated classifiers – SVM, MLP, DT and GMDH independently. Basic parameters required for tuning the execution of these classifiers are set. For SVM, the piece write = rbf, portion scale = 1, emphasis = 1000 and regularization parameter = 1 has been picked. The parameters – least number of leaf measure = 1, split basis = gdi (Gini's decent variety record) and least split size = 10 have been chosen for working of

iJETRM

International Journal of Engineering Technology Research & Management

DT. The applicable parameters for MLP are concealed layer measure = 3, hubs per concealed layer = 8, preparing capacity = trainlm, execution work = crossentropy, actuation work = tansig for shrouded layer what's more, softmax for yield layer and greatest emphasis = 1000. The GMDH parameters have been set as: most extreme number of covered up layer = 3 and most extreme number of neurons in a layer = 8. Once the practical parameters are set for every student, the order of suspicious occurrences is done on beforehand prepared classifier models separately. A relative execution investigation of every classifier with and without grouping on the uneven protection dataset has been introduced in Table 6. The yield of best performing classifier has been featured in strong for better perception. The outcomes in the table obviously demonstrate the viability of utilizing GAFCM over typical FCM for characterization. The MLP and GMDH yielded 0% Sensitivity without utilizing grouping due to the skewed class dissemination display in the protection dataset. The MLP gives the greatest Sensitivity = 73.35% in the wake of utilizing FCM grouping on the imbalanced dataset, while SVM produces the best proficiency as far as most noteworthy Sensitivity = 69.70% and Specificity = 84.71%.

Correspondingly, the execution investigation of every classifier with regard to adjusted dataset has been exhibited in Table 7. The yield of the best performing classifier has been featured in strong. The MLP produces the most astounding Sensitivity = 75.75% when utilizing FCM as information adjusting procedure, while DT gives the most extreme Exactness = 71.79% and Specificity = 73.19%. When utilizing the GAFCM as an undersampling approach, the SVM beats every other classifier regarding all execution measurements.

Comparative Analysis

In this area, a near examination of the proposed framework has been finished with another collision protection extortion identification approach proposed by considering a similar protection dataset "carclaims.txt" (Sundarkumar et al., 2015). The creators in Sundarkumar et al. (2015) have portioned the dataset into prepare, test and approval set. At first, an approval set of size 20% of the unique dataset has been kept aside and the rest 80% information are subjected on their framework. The 10-crease cross approval system has been utilized on the staying 80% dataset for isolating into prepare and test set.

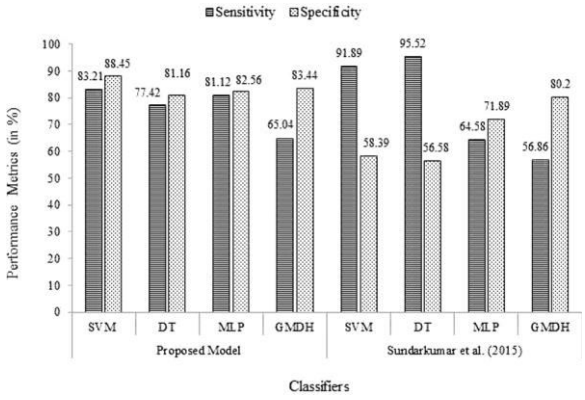
The creators (Sundarkumar et al., 2015) have utilized One Class Bolster Vector Machine (OCSVM) for expelling the skewed class circulation from the dataset. The OCSVM has been connected on the veritable examples of the prepare set for extricating the help vectors prompting a diminished size of ordinary class information. The deceitful occasions (minor) are then joined with the real class set to frame an altered preparing set. Four distinctive managed classifiers – SVM, MLP, GMDH and DT have been connected on the adjusted prepare set autonomously to generate their relating prepared demonstrate. The test set is utilized on the individual prepared models for testing the adequacy of the models. At long last, the approval set is utilized for approving each model.

The execution measurements – Sensitivity and Specificity are utilized for near execution examination. A decent classifier ought to fundamentally show higher estimations of Sensitivity and in addition Specificity as they show precision of arrangement of manufactured occasions and brought down false alerts individually. In the wake of following the test system portrayed in Sundarkumar et al. (2015), the outcomes acquired for the proposed display and the approach under examination has been displayed in Fig. 4. It can be induced from the figure that the proposed display successfully limits the misclassification rate by giving the most noteworthy Specificity in all classifiers as thought about to Sundarkumar et al. (2015). Additionally, the proposed demonstrate produces the most astounding Sensitivity = 83.21% and Specificity = 88.45% after utilizing SVM as the classifier, though, if there should be an occurrence of Sundarkumar et al. (2015), DT produces the best execution in terms of Sensitivity = 95.52% and Specificity = 56.58%.

ISSN: 2456-9348
Vol(01) _Issue (03)

iJETRM

International Journal of Engineering Technology Research & Management



CONCLUSION

In this examination, we have proposed a novel hybrid approach for automobile insurance fraud detection that returns in two stages – training and fraud detection. In training stage, a GA based enhanced FCM (GAFCM) clusters has been utilized for under sampling the majority class tests in the skewed prepare dataset in order to enhance the effectiveness of the classifiers. At first, the GAFCM grouping is utilized on the majority class occurrences for creating clusters with optimal cluster centers. The outliers and in addition repetitive data points exhibit in the majority class are then distinguished and expelled, along these lines encouraging under sampling. The reduced majority part class tests are then joined with the first minority class focuses to get an adjusted dataset, which is utilized for further experimentation.

The fraud detection method is completed in two phases in the proposed system. Amid the first phase of fraud identification, the GAFCM arranges the test data points as malicious, suspicious and genuine classes in light of their distance measure from the optimized cluster centers. The samples recognized as malicious and genuine are not additionally prepared, while the suspicious ones are furthermore checked in the second stage by unique supervised learner – SVM.

REFERENCES

[1]Abdallah, A., Maarof, M.A., Zainal, A., 2016. Fraud detection system: a survey. J.Network Comput. Appl. 68, 90–113. (accessed: 9.05.17).
[2]Bermúdez, L., Pérez, J., Ayuso, M., Gómez, E., Vázquez, F., 2008. A bayesiandichotomous model with asymmetric link for fraud in insurance. Insurance:Math. Econ. 42 (2), 779–786.
[3]Bernard, C., Vanduffel, S., 2014. Mean–variance optimal portfolios in the presence ofa benchmark with applications to fraud detection. Eur. J. Oper. Res. 234 (2),469–480.
[4]Bezdek, J.C., Ehrlich, R., Full, W., 1984. Fcm: The fuzzy c-means clustering algorithm.
[5]Comput. Geosci. 10 (2–3), 191–203.Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297.
[6]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-getcheaper-
motor-insurance-backfiring-on-thousands-of-motorists-warns-the-abi/(accessed: 9.05.17).

iJETRM

International Journal of Engineering Technology Research & Management

[7]Eiben, A.E., Raue, P.-E., Ruttkay, Z., 1994. Genetic algorithms with multi-parentrecombination. In: International Conference on Parallel Problem Solving from Nature. Springer, pp. 78–87.case study. In: AAAI Workshop on AI Approaches to Fraud Detection and RiskManagement. pp. 34–38.
[8]Lee, Y.-J., Yeh, Y.-R., Wang, Y.-C.F., 2013. Anomaly detection via online oversamplingprincipal component analysis. IEEE Trans. Knowl. Data Eng. 25 (7), 1460–1470.
[9]Łe˛ ski, J.M., Owczarek, A.J., 2005. A time-domain-constrained fuzzy clusteringmethod and its application to signal analysis. Fuzzy Sets Syst. 155 (2), 165–190.
[10]Pathak, J., Vidyarthi, N., Summers, S.L., 2005. A fuzzy-based algorithm for auditors todetect elements of fraud in settled insurance claims. Managerial Auditing J

[11]Phua, C., Alahakoon, D., Lee, V., 2004. Minority report in fraud detection:classification of skewed data. Acm Sigkdd Explor. Newslett. 6 (1), 50–59.Quinlan, J.R., 1987. Simplifying decision trees. Int. J. Man-mach. Stud. 27 (3), 221–234.
[12]Rosenblatt, F., 1961. Principles of neurodynamics. perceptrons and the theory ofbrain mechanisms. Tech. rep., DTIC Document.
[13]Šubelj, L., Furlan, Š., Bajec, M., 2011. An expert system for detecting automobileinsurance fraud using social network analysis. Expert Syst. Appl. 38 (1), 1039–1052.
[14]Sundarkumar, G.G., Ravi, V., 2015. A novel hybrid undersampling method formining unbalanced datasets in banking and insurance. Eng. Appl. Artif. Intell.37, 368–377.
[15]Sundarkumar, G.G., Ravi, V., Siddeshwar, V., 2015. One-class support vectormachine based undersampling:
[16]Application to churn prediction and insurance fraud detection. In: Computational Intelligence and Computing Research(ICCIC), 2015 IEEE International Conference on. IEEE, pp. 1–7.
[17]Tao, H., Zhixin, L., Xiaodong, S., 2012. Insurance fraud identification research basedon fuzzy support vector machine with dual membership. In: Information
[18]Management, Innovation Management and Industrial Engineering (ICIII), 2012 International Conference on. Vol. 3. IEEE, pp. 457–460.
[19]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.
[20]Xu, W., Wang, S., Zhang, D., Yang, B., 2011. Random rough subspace based neuralnetwork ensemble for insurance fraud detection. In: Computational Sciencesand Optimization (CSO), 2011 Fourth International Joint Conference on. IEEE, Spp. 1276–1280.