IJE TRANSACTIONS A: Basics Vol. 33, No. 7, (July 2020) 1240-1248

International Journal of Engineering

Journal Homepage: www.ije.ir

Fast Unsupervised Automobile Insurance Fraud Detection Based on Spectral Ranking of Anomalies

Z. Shaeiri[1], S. J. Kazemitabar*[2]

[1] *Son Corporate Group, Tehran, Iran*

[2] *Department of Electrical and Computer Engineering, Babol Noshirvani University of Technology, Babol, Iran*

PAPER INFO

Paper history:
Received 10 November 2019
Received in revised form 23 December 2019
Accepted 04 January 2020

Keywords:
Spectral Ranking of Anomalies
Auto-insurance Fraud Detection
Random Forest
Deep Learning

ABSTRACT

Collecting insurance fraud samples is costly and if performed manually is very time consuming. This issue suggests the usage of unsupervised models for fraud data collection. One of the accurate methods in this regards is Spectral Ranking of Anomalies (SRA) that is shown to work better than other methods for auto-insurance fraud detection, specifically. However, this approach is not scalable to large samples and is not appropriate for online fraud detection. This is while, real-time fraud management systems are necessary to prevent huge losses. In this study, we propose an implementation methodology which makes it possible to apply the SRA to big data senarios. We exploit the power of spectral ranking of anomalies to create an estimated target variable from the unlabeled dataset. We then use two robust models, namely, random forest and deep neural networks to fit a model based on the estimated labeled training set. Next, the incoming live data are fed to the mentioned trained models for predicting the target variable. Simulation results confirm that the proposed approach has higher speed and acceptable false alarm rate compared to existing related methods.

***doi**: 10.5829/ije.2020.33.07a.10*

NOMENCLATURE

$k(s_1, s_2)$	Kernel function	$\delta(s_1, s_2)$	Delta Kronickel function
σ	Kernel width parameter	f	Ranking vector
W	Adjacency matrix	L	Laplacian matrix

1. INTRODUCTION

A major concern among insurance companies is the issue of claims fraud. According to ABI (Association of British Insurers) statistics, in 2014, 3.6 million pounds was lost due to claims fraud on a daily basis. According to Tennyson and Salsas-Forn-2002 [1] approximately 21% to 36% of auto-insurance claims are fraudulent but only less than 3% of suspected cases are detected and persecuted. Besides putting the advantages of the insurer at risk, fraud is also harmful for its value chain. It increases the cost of insurance which will be tolerated by the insured parties in the form of increased premium rates. Thus, fraud is detrimental to the very basic dependencies that keep the concept of insurance alive. Fraud detection is essential to prevent huge losses which insurance systems are encountered with. Traditionally, insurance fraud detection was performed by insurance investigators and claim adjusters. Since manually detecting suspicious claims from huge insurance fraud datasets is inefficient, data mining and machine learning methods are used extensively. These methods help the insurers detect fraud prior to reimbursing the customer. This is an essential requirement that machine learning and data mining approaches can handle appropriately. In the literature, auto-insurance fraud detection was most often performed via some supervised models and scarcely with unsupervised methods. Supervised models need to have access to the claims of both classes; fraudulent and legitimate. However, two issues exist that make supervised models difficult to use. First, there are not many labeled data samples available to work with.

*Corresponding Author Institutional Email: *j.kazemitabar@nit.ac.ir* (S. J. Kazemitabar)

Please cite this article as: Z. Shaeiri, S. J. Kazemitabar, Fast Unsupervised Automobile Insurance Fraud Detection Based on Spectral Ranking of Anomalies, International Journal of Engineering (IJE), IJE TRANSACTIONS A: Basics Vol. 33, No. 7, (July 2020) 1240-1248

Moreover, since the fraud datasets are imbalanced most supervised methods cannot perform very well. More precisely, their classification performance is not equal across both fraud and non-fraud samples.

In [2], the authors show the performance of binary classifiers (logistic) for auto-insurance fraud detection and obtain models for demonstrating misclassification in the target variable which is caused by auditors' mistakes. Performance of various classification techniques for auto-insurance fraud detection is taken into account in [3]. Performances of logistic, k-nearest neighbor, Bayesian learning, multilayer perceptron neural network, support vector machine, naive Bayes, and tree-augmented naive Bayes classification were compared and contrasted in that paper. In [11], the most relevant attributes are chosen from the original dataset by using an evolutionary algorithm based feature selection method. A test set is then extracted from the selected attribute set and the remaining dataset is subjected to the Possibilistic Fuzzy C-Means (PFCM) clustering technique for the under-sampling approach. In [12], suspicious groups have been detected by applying cycle detection algorithms (using both DFS, BFS trees). Afterwards, the probability of being fraudulent for suspicious components was investigated to reveal fraudulent groups with maximum likelihood, and their reviews were prioritized. In [13], a multiple classifier system based on Random Forest, Principle Component Analysis and Potential Nearest Neighbor is proposed. The authors ameliorate the classification accuracy of the ensemble classifier by improving the difference of the base classifiers. The proposed method is then applied to detect automobile insurance fraud and the fraud rules are obtained. Proposing supervised learning models seems inefficient as collecting target variables for auto-insurance fraud datasets -like other insurance fraud datasets- is very costly and time consuming. Moreover, the investigators act significantly different from one another in their fraud diagnosis. In addition, risk assessment is more appealing to insurers than simply accessing binary fraud/non-fraud classification of claims. Considering these facts, unsupervised methods and specifically unsupervised anomaly ranking seems more appropriate for fraud detection problems. In the literature, few unsupervised auto-insurance fraud detection models are proposed. One of the earlier unsupervised approaches in this context is proposed by Brockett et al., in 1998 [4]. In this paper, the authors propose to apply a self-organizing neural network for classification of fraud datasets. In another work, Brockett et al., proposed the PRIDIT methodology (Principal component analysis and RIDIT scoring method) for auto-insurance fraud detection [5]. In these works, the authors have examined their studies over the PIP (Personal Injury Protection) dataset which is provided by AIB (Automobile Insurance Bureau). It is worth noting that in this dataset suspicion of fraud among samples is ranked somehow by the auditors and experts. Thus, it is fair if one considers these works as semi-supervised methods. Fraud datasets often contain categorical and ordinal variables. These variables require pre-processing prior to being used, a process which itself requires expert knowledgedataset. Another point that needs to be emphasized is that since we are attempting to classify or rank the fraud dataset it seems that the dataset should have one major class of normal samples and one small pattern of anomalies.

However, this will not be the case when the dataset contains many categorical variables. Therefore, methods that rely on the implicit assumption that the dataset contains one major pattern may not be successful. In another work [6], the authors proposed an unsupervised method which is based on spectral ranking of anomalies (SRA). Their work is motivated by the observation that there is a connection between unsupervised Support Vector Machine (SVM) optimization formulation and the spectral optimization. They derived a ranking vector

(a) Visualization of data instances (this dataset is a 2-Dimensional synthetic dataset

(b) The second non-principal Eigen-vector of the Laplacian versus the first one

Figure 1. Visualizing the information contained in the synthetic data

which provides the degree of relative abnormality of samples in the dataset. To the best of our knowledge, among different methods in the literature, SRA provides significant results in ranking anomalies for auto-insurance fraud datasets [15-22]. This observation motivates us to focus more on the SRA and to try to implement and apply it on big datasets. As shown by the authors in [6], the performance of this unsupervised method is very close to the supervised techniques such as SVM which serves as an upper bound for unsupervised methods. While it is considerably accurate and effective in auto-insurance fraud detection, this method suffers from high computational complexity and cannot be applied for online fraud detection. In this paper, we propose an online unsupervised methodology for auto-insurance fraud detection. Here, we want to explain the important drawback of the theoretically accurate SRA method, which makes it impossible to apply to big datasets. Implementation of SRA involves large scale matrix multiplication and Eigen decomposition. For real world datasets the similarity/distance matrix is a huge dense square matrix which is always too large to fit in memory. This makes it impossible to implement SRA on big datasets.And even if it were computationally feasible to do, so, its memory requirements are unusually high. To the best of our knowledge, today a handful of technologies provide a solution for applying linear algebra operations on large dense square matrices which also encompass high memory requirements. The main contribution of this paper is thus to facilitate the implementation of SRA for big datasets with low memory resources. First, we transform the unsupervised problem to a supervised one which means to provide labels for a fraction of the data. We apply SRA on this fraction and obtain the anomaly ranking vector. Then, we use this vector as a guide for estimating the target variable for the mentioned small set of samples. The SRA method is very accurate in ranking the data points. We exploit the ranking vector derived from SRA and apply a threshold on it for labeling the points as normal and fraudulent. The concept of SRA is such that it is less affected by imbalanced data. The details of how SRA handles this issue can be found in [6]. In short, it defines a parameter which is the ratio of the two labels and uses that for applying a threshold. Fraud detection is an interactive task which means that expert knowledge is exploited in various steps of the design and implementation. For example, the mentioned threshold value is given by the experts and is fixed.

The organization of the paper is as follows. In Section 2 background of the spectral anomaly ranking method is provided. An overview of the similarity measures used in auto-insurance fraud detection is provided in Section 3. The distance measure and the kernel function that are used in this paper are discussed in subsections 3.1 and 3.2 respectively. In Section 4 the proposed methodology is introduced. Finally, simulation results are provided in Section 5.

2. BACKGROUND OF SPECTRAL RANKING OF ANOMALIES

In [6], an unsupervised method is proposed which uses spectral ranking of anomalies for fraud detection. Motivated by the analogy between unsupervised SVM optimization and the spectral optimization formulation, the authors indicated that spectral optimization can be treated as a relaxation of the unsupervised SVM optimization.They derived a similarity matrix using Hamming distance measure which is appropriate for datasets consisting of categorical and ordinal variables. They demonstrated that the absolute value of the first non-principal Eigen-vector of Laplacian of this similarity matrix provides a measure of anomaly ranking in a bi-class clustering problem. Magnitudes of entries of this non-principal Eigen-vector contain valuable information about the degree by which the corresponding samples (samples in the same positions) are anomalous. An observation is more likely to be an anomaly if the magnitude of its corresponding entry in the non-principal Eigen-vector is larger. It is worth noting that based on the structure of the underlying dataset some possible scenarios may arise. In the SRA a choice of reference is allowed in the anomaly detection process, such that the mass of the minority cluster determines how to generate the ranking. For example, in one of the probable scenarios, the minority cluster does not have a sufficient mass. In this case the anomaly likelihood can be assessed with respect to a single majority class. In another scenario when the minority cluster has sufficient mass, anomaly can be assessed with two main clusters. We refer the readers to [6] for more details about these different scenarios. More details of this method is presented in Figure 1.

2. 1. Overview of Similarity Measures In Auto-insurance Fraud Detection Traditionally, binary predictor variables were being used in the problem of auto-insurance fraud detection. Examples of such binary variables are coverage (third party liability equals 1 and extended coverage equals 0), deductible (existence of a deductible equals 1, otherwise equals 0), witness (existence of witness equals 1, otherwise equals 0), and so on. However, many of the important categorical predictor variables in the fraud detection problem have more than two categories, e.g., age of the driver. Age can be expressed as a categorical variable with for example 5 number of ordered categories. Some works use natural integer scoring for these variables [5]. In natural integer scoring, one simply assigns for instance the numbers 1, 2, 3, 4, and 5 to the five different categories of a variable.

It is worth noting that this kind of variable transformation can impose unwanted scaling and order, and unintended distribution to the original predictor variable which finally will result in weak and incorrect outcomes. The fact is that many of the predictor variables in a real world dataset are categorical or ordinal. Most of the machine learning approaches are sensitive to the above mentioned pre-processing steps which are applied on the datasets. These pre-processing steps sometimes dramatically change the predictor variables meanings, sometimes cause information loss, and ultimately result in unreliable outcomes. Historically, working on similarity measures dates back to the past century [7]. One of the seemingly appropriate methods in the current problem is similarity measures based on match and mismatch of the nominal values of the variables. This similarity measure is known as nominal value definition derived in [6]. While it is simple, this similarity measure preserves the meaning of the predictor variables. This way, one does not require the pre-processing step for transforming the predictor variables. In this paper, we use this similarity measure for obtaining a dissimilarity matrix.

2. 2. Distance Measure A distance measure is a real-valued function which shows the extent of dissimilarity between two samples in the data space. For each pair of N-dimensional samples ($\boldsymbol{s}_1$,$\boldsymbol{s}_2$) Hamming distance is defined as the number of mismatch between the variables divided by total number of variables:

$$d^H(\boldsymbol{s}_1, \boldsymbol{s}_2) = \frac{\sum_{i=1}^{N} \delta(s_1^i, s_2^i)}{\boldsymbol{n}} \quad (1)$$

In this equation $\boldsymbol{s}_1^i$ and $\boldsymbol{s}_2^i$ are the *i*th element of $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ respectively and $\delta(\boldsymbol{s}_1^i, \boldsymbol{s}_2^i)$ is defined as:

$$\delta(\boldsymbol{s}_1^i, \boldsymbol{s}_2^i) = \begin{cases} 1, & \boldsymbol{s}_1^i \neq \boldsymbol{s}_2^i \\ 0, & \boldsymbol{s}_1^i = \boldsymbol{s}_2^i \end{cases} \quad (2)$$

2. 3. Kernel Function To handle complicated relationships among attributes, it is common to transform the data into a usually high dimensional feature space via various kinds of kernel methods [8]. Datasets that are produced based on human activities and behaviors, e.g. insurance claim datasets, naturally contain categorical and ordinal features. As stated above, it is shown that a meaningful treatment for capturing relationships among different features in the datasets containing categorical features is exploiting an appropriate similarity measure. In [9], Couto introduced the Hamming distance kernel for datasets containing categorical features. For each pair of data instances this similarity measure is achieved by considering match and mismatch between the categorical features. In this paper, we use the Gaussian kernel derived from the Hamming distance which is obtained by replacing the Euclidean distance in the standard Gaussian kernel with the Hamming distance. More precisely, the kernel function

$$k(\boldsymbol{s}_1, \boldsymbol{s}_2) = \exp\left(-\frac{d^{\{H\}}(\boldsymbol{s}_1, \boldsymbol{s}_2)}{2\sigma^2}\right) \quad (3)$$

is considered in which $\sigma > 0$ is a constant kernel width parameter.

3. OVERVIEW OF THE PROPOSED METHOD

Spectral clustering is one of the recent clustering techniques that proved its superiority among traditional clustering methods. The spectral ranking of anomalies was applied for fraud detection on insurance claim datasets. It shows considerable improvements in anomaly detection compared with other unsupervised methods such as one-class support vector machine (OC-SVM) and local outlier factor (LOF) when applied on automobile fraud detection dataset as well as various kinds of synthetic datasets [6]. However, SRA is not applicable and appropriate for handling big datasets. The structure of this algorithm makes it difficult to work on big datasets. The SRA requires using all the samples from the beginning. There is no mechanism to exploit it for live data. Like many other precise anomaly detection methods, SRA requires the formation of the similarity/distance matrix. As one knows, this matrix is a dense square matrix of order *M* (the number of records or samples in the dataset). Practically, creating this matrix for real world datasets is resource intensive. When the dense square matrix is big, one cannot possibly afford storing it in a dense way. It probably would not even fit into the memory. On the other hand, implementation of SRA involves large scale matrix multiplication and Eigen decomposition. First, we should find an appropriate way to create the required matrices. It is only at that stage that we can hope to proceed through the rest of the stages. More precisely, just then we may implement the matrix multiplications and the Eigen decomposition steps of the algorithm. We provide a methodology for facilitating SRA for handling big datasets with low memory resources and high performance as well as acceptable false alarm rate. We first use a fraction of the dataset to be processed with the spectral ranking method. Output of the spectral ranking is a ranking vector denoted here as f. Using this ranking vector and a threshold value, the portion of the data with higher risk is labeled. In the SRA, the ratio of the number of fraud cases to the number of normal records is used as the threshold value. Generally, the threshold is chosen experimentally by referring to the domain experts. We have used the same threshold value that is exploited in the SRA method. The result is a two-class labeled dataset in which the samples are tagged as normal and fraudulent. We then use the tagged data and fit a model via supervised learning methods such as random forest or deep learning. This model can then be used to process any incoming data to determine whether it is fraudulent or not.

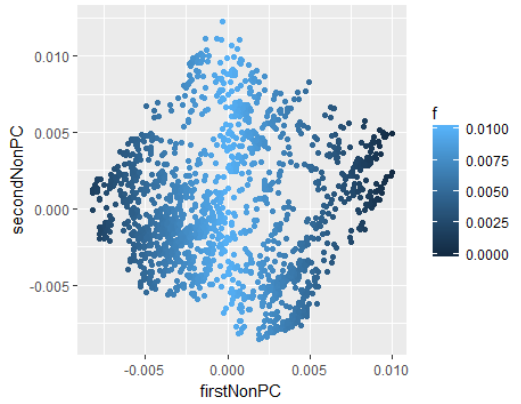
3. 1. Details Let D be the data space consisting of data instances $\boldsymbol{s}_i$ with $i = 1,\ldots,M$ where M is the number of data points in D. The main objective is to cluster the dataset into a number of clusters such that distances are minimized within each cluster. This similarity based clustering can be viewed as a graph cut problem in which each data point is a vertex and each pair of vertices are connected together by an edge with a weight equal to the similarity of the corresponding data points. An adjacency matrix W is derived which summarizes the similarity between each pair of data instances in D. The distance measure which we have used in this paper can handle nominal and ordinal variables as well as numerical variables. The handling of nominal and ordinal variables is achieved by using the general dissimilarity coefficient of Gower [10] in which match and mismatch of the variables entries are considered for deriving the distance measure. The Euclidean distance is used for numeric variables. The numeric variables are rescaled before applying the Euclidean distance. Each rescaled (numeric) variable has range [0,1] exactly. The mentioned distance measure is exerted into the standard Gaussian kernel to obtain the adjacency matrix W. More precisely:

$$W = \exp\left(-\frac{R}{2\sigma^2}\right), \quad (4)$$

where R is the distance matrix with its (i,j)th entry R_{ij} being the distance between data points i and j. Next, we compute the degree matrix D of the vertices which is a diagonal matrix with diagonal entries $d_i = P_j W_{ij}$. From the adjacency matrix W and the degree matrix D the Laplacian matrix L is derived which is a fundamental quantity in the spectral anomaly ranking part of our methodology. We use the following definition of the Laplacian matrix in our analysis:

$$L = I - D^{-0.5} W D^{-0.5}, \quad (5)$$

in which I is the identity matrix. SRA introduces a technique for ordering the data instances based on their anomalous behavior. The main idea of this method is that entries of a non-principal Eigen-vector of the Laplacian matrix provide valuable information for anomaly detection. Let $\lambda_0 < \ldots < \lambda_{M-1}$ be the M Eigen-values of Laplacian matrix L. Associated with each Eigen-value λ_i there is an Eigen-vector v_i . Based on the first non-principal Eigen-vectors of L a ranking vector is derived which provides meaningful distinguishability among normal data instances and abnormal ones. Figure 1 provides an illustrative example of the spectral ranking method in which the distinguishability or clustering strength of the first non-principal Eigen-vector of L is demonstrated for a balanced two-class dataset. The first and the second non-principal Eigen-vectors of the Laplacian, corresponding to the Gaussian kernel are depicted. In Figure 1.a, true output class labels are specified by the color, where red points indicate the normal data instances, while blue points show anomaly cases. Each point in the 2-dimensional Eigen-space corresponds to one point in the original data space. To each point in the Eigen-vectors space a value is assigned, the magnitude of which shows the level of abnormality of the corresponding point. In Figure 1.b, points with larger f (ligther color) are associated with data instances that are more abnormal. Figure 2 shows similar results for the automobile fraud dataset. As can be inferred from this figure, vector f provides a good measure for anomaly since it properly distinguishes the normal data points from the abnormal ones. Despite its precision in ranking the data instances, SRA is a huge resource consumer. This is mainly due to the matrix multiplication operation and the Singular Value Decomposition (SVD) calculation in this method. Thus, SRA is impractical for large datasets and it cannot be applied for online anomaly/fraud detection. In this study, we propose a methodology in which the power of SRA -its accuracy- is exploited in designing an online auto-insurance fraud detection technique. Based on the ranking vector (f) derived from the SRA and by applying a wisely chosen



(a) The second non-principal Eigen-vector of the Laplacian versus the first one.

(b) Visualization of data instances (The auto-insurance fraud dataset)

Figure 2. Visualizing the information contained in the auto-insurance dataset

threshold upon it we estimate the target variables/labels for a small fraction of the dataset. This fraction is selected using random sampling technique. Random sampling is as fair and unbiased as possible since it makes units equally likely to be chosen. It ensures independent selection by gathering as much independent information as possible. Thus, the sample is fair and representative. In the SRA the threshold is chosen approximately based on the prior knowledge, i.e., approximate percentage of normal and anomaly cases. Generally, one can approximately achieve an acceptable threshold by referring to the domain experts. We treat the labeled dataset as a training set. In other words, relying on correctness of the ranking vector f, we transform the unsupervised problem to a supervised one. Data points with f greater than the threshold value are treated as anomaly and the remaining as normal cases. Next, two powerful supervised methods, namely the Random Forrest (RF) and the Deep Learning (DL) classification models are trained using the generated training set and their speeds and accuracies are analyzed. The results show that while the proposed method is simple and straightforward, it is considerably accurate as well as fast.

3. 2. Remarks For computing the similarity/distance matrix, we used the R "daisy" function from "cluster" package which computes all the pairwise dissimilarities between observations in the dataset. This function can handle datasets with mixed type variables (nominal, ordinal and numeric). Table 1 shows the amount of the allocated memory for different sample sizes.

It shows that we face serious memory issues for large sample sizes. The main contribution of this paper is to facilitate implementation of SRA for big datasets with low memory resources. We used random sampling technique for the selection of the small dataset. Results show significant robustness with standard deviation of percentage of accuracy 0.35 for 10 runs. We exchanged the accuracy by speed.

4. SIMULATION RESULTS

Several experiments are conducted on an auto-insurance claim dataset as well as two synthetic datasets. Description of the datasets are given in Table 1. The

TABLE 1. Memory usage of creating the distance matria

# of records of the sample data	Memory allocated
1542	185 MB
3084	731 MB
4626	1.64 GB
6168	2.9 GB

synthetic datasets contain two normal clusters and anomaly points. Each of the normal clusters consists of 2000 Gaussian distributed data instances. The synthetic dataset 1 contains 200 uniformly distributed point anomalies and synthetic dataset 2 contains two small Gaussian distributed anomaly clusters together with 200 uniformly distributed point anomalies. The auto-insurance claim dataset is collected by Angoss KnowledgeSeeker software from January 1994 to December 1996. This dataset contains 15420 observations where each claim is assigned a label indicating if that claim is a normal case or it is an anomaly. Totally, it contains 14497 normal and 923 fraudulent cases. To the best of our knowledge this dataset is the only labeled auto-insurance fraud dataset available in the academic literature. The predictor variables contained in this dataset are categorical and ordinal including base policy, day of week, number of cars, witness present, and the past number of claims. First, to obtain the target variable, spectral ranking of anomalies is performed by running the SRA on the unlabeled training set. For categorical features, the Hamming distance and for numerical features the Euclidean distance is used as the dissimilarity measure. In the SRA based estimation of the target variable the Gaussian similarity kernel is considered. Using a wisely chosen threshold, we labeled the training dataset by the derived target variable. The threshold is chosen by referring to the prior knowledge and the domain experts. In the training phase, the mentioned labeled dataset which is the output of the previous stage, is used to train the RF and DL models. In the RF model the number of trees are set to 50. The algorithm converges when the 2-tree average is within 0.001 of the prior two 2-tree averages. The DL model is a 5-layer neural network. We used deep learning implementation in R language's H2O package. Hyperbolic tangent is used as the activation function and the number of epochs are set to 1000. First, we will create two independent splits for train (80 percent) and test (20 percent) sets. The models are trained on the train set. The test set is used for ensuring that the model can predict properly on the new datasets. The results are compared with the original SRA model. We have used the ROC curves to compare our results with the SRA. Tracing the ROC curves is a commonly used way for visualizing the performance of binary classifiers. The ROC curves show the trade-off between true-positive and false-positive quantities for different choices of threshold. Therefore, it does not depend on prior knowledge to combine true-positive and false-positive quantities into a unique object. From the ROC plot, one can distinguish the dominant algorithm, i.e. the one that provides a better solution at any false-positive value. We have also calculated the Area Under Curve (AUC) for these methods as another measure of the quality. Simply speaking, AUC summarizes the performance of a binary classifier in a single number. Table 3 contains the

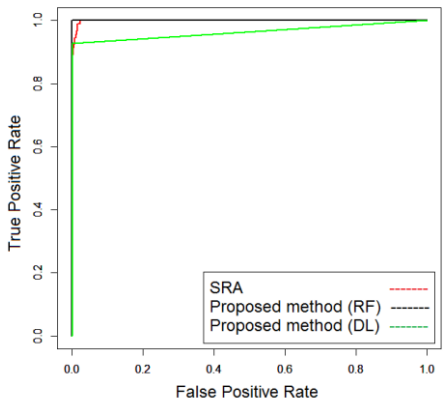
execution time of the methods for the datasets. For the auto-insurance fraud dataset the execution time for the proposed method does not include the time it takes to create the labeled training set. The fact is that by increasing the volume of dataset the gap between the execution time of the proposed method and the SRA will increase dramatically. It can be seen that the execution time of the proposed methodology is significantly less than the execution time of the original SRA method. Figures 3 and 4 show the ROC curves of the supervised SRA and our methodology on the insurance claim dataset and on the two synthetic datasets. From Tables 2 and 3 and Figure 4 it can be inferred that while comparable with the SRA in terms of the accuracy the proposed methodology has higher speed. Table 4 contains comparisons between SRA and two non-spectral ranking methods such as LOF and OC-SVM.

TABLE 2. Description of the datasets

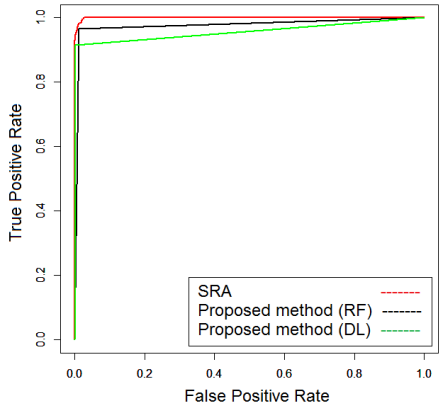
Dataset	# of Normal	# of Anomaly	Description
Synthetic data 1	2000	200	The dataset consists of 2 large normal clusters and 200 point anomalies.
Synthetic data 2	2000	287	The dataset consists of 2 large normal clusters together with 200 point anomalies and 2 small clusters.
Insurance data [1]	14497	923	The dataset is provided by Angoss KnowledgeSeeker software. It consists of 31 categorical features.

TABLE 3. Execution time (s)

Dataset	SRA	Proposed method (RF)	Propose method (DL)
Synthetic data 1	18.43	3.08	2.84
Synthetic data 2	18.86	2.22	3.14
Insurance data [1]	17.31	2.45	3.91



(a) Synthetic dataset 1



(b) *Synthetic dataset 2*

Figure 3. ROC Curves of methods for two synthetic datasets

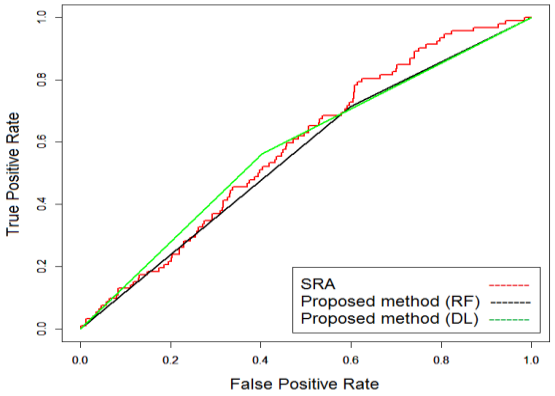


Figure 4. ROC curves of methods for insurance fraud dataset; AUC of SRA, RF, and DL are respectively, 0.6, 0.52, and 0.57

5. CONCLUSION

In this paper, a fast online unsupervised methodology was proposed for the challenging auto-insurance fraud detection problem. The proposition was based on the spectral ranking of anomalies [6]. This method is one of the recently published unsupervised anomaly detection methods which is used for anomaly detection for auto-insurance claim dataset. Despite its high precision among previous unsupervised methods, for very large datasets, a number of large matrix multiplication and Eigen value decomposition stages make this method useless. To tackle this problem, in this study, we proposed to exploit the power of the spectral anomaly ranking approach for generating a training set. The SRA was applied on a small fraction of the unlabeled dataset. The ranking vector generated by the SRA was used for creating the training set. The generated training set was used for training two models, namely, random forest and deep learning models. These trained models were used for estimating the target variable from the unlabeled dataset.

TABLE 1. Summary of the AUC for the auto-insurance fraud detection dataset For entries marked by *, AUC reported is one minus the actual AUC, OS: Overlapping Similarity, AGK: Adaptive Gaussian Kernel, HDK: Hamming Distance Kernel, DISC: DISK similarity

Automobile Fraud Detection Dataset										
Method			OS	AGK				HDK		Disk
				β				λ		
				10	100	1000	3000	0.5	0.8	
LOF	k_{svm}	10	0.53	0.5	0.52	0.58	0.64	0.52	0.53	0.55
		100	0.51	0.51*	0.54	0.58	0.67	0.51	0.52	0.57
		500	0.53	0.52*	0.55	0.59	0.68	0.51	0.51	0.57
		1000	0.53	0.52*	0.53	0.59	0.69	0.5	0.5	0.56
		3000	0.5	0.58*	0.55	0.58	0.69	0.54*	0.55*	0.53
OC-SVM	v_{svm}	0.01	0.51*	0.53*	0.51*	0.54	0.59	0.51*	0.52*	0.53*
		0.05	0.51*	0.53*	0.51*	0.55	0.59	0.52*	0.53*	0.52*
		0.1	0.51*	0.54*	0.51*	0.55	0.59	0.53*	0.54*	0.56*
SRA	mFlag	1	0.73	0.74	0.74	0.66	0.74	0.74	0.74	0.66

The approach is tested on a real auto-insurance claim dataset as well as two synthetic datasets. Results confirm the superiority of the proposed method in terms of accuracy as well as speed and performance.

Modern datasets are rapidly growing in size. Today, a handful of technologies provide solutions for handling large matrix operations. Apache Spark has emerged as a widely used open-source engine which is a fault-tolerant and general-purpose cluster computing framework. It provides APIs in Python, R, Java, and Scala. It also provides an optimized engine that supports general execution graphs. Recently, distributed linear algebra and some new optimization libraries have been developed in Spark. The linalg library consists of fast and scalable implementations of standard matrix computations. Common linear algebra operations such as multiplication, and more advanced operations such as factorizations are implemented in this library [14]. Using these technologies, one can proceed in implementing the SRA for the entire dataset. We suggest using these libraries to implement the SRA algorithm on large datasets.

6. REFERENCES

1. Tennyson. S, and Salsas-Forn. P, "Claims Auditing in Automobile Insurance: Fraud Detection and Deterrence Objectives", ***Journal of Risk and Insurance***, Vol. 69, No. 3, (2002), 289-308, doi: 10.1111/1539-6975.00024.
2. Artis. M, Ayuso. M, and Guillén. M, "Detection of Automobile Insurance Fraud with Discrete Choice Models and Misclassified Claims", ***Journal of Risk and Insurance***, Vol. 69, No. 3, (2002), 325-340, doi: 10.1111/1539-6975.00022.
3. Viaene. S, Derrig. R. A, Baesens. B, and Dedene. G, "A Comparison of State-Of-The-Art Classification Techniques For Expert Automobile Insurance Claim Fraud Detection", ***Journal of Risk and Insurance***, Vol. 69, No. 3, (2002), 373-421, doi: 10.1111/1539-6975.00023.
4. Brockett. P. L, Xia. X, and Derrig. R. A, "Using Khonen's Self Organizing Feature Map to Uncover Automobile Bodily Injury Claim Fraud", ***Journal of Risk and Insurance***, Vol. 65, No. 2, (1998), 245-274, doi: 10.2307/253535.
5. Brockett . P. L, Derrig. R. A, Golden. L. L, Levine. A, and Alpert. M "Fraud Classification Using Principal Component Analysis of RIDITs", ***Journal of Risk and Insurance***, Vol. 69, No. 3, (2002) 341-371, doi: 10.1111/1539-6975.00027.
6. Nian. K, Zhang. H, Tayal. A, Coleman. Th, and Li. Y, "Unsupervised Spectral Ranking For Anomaly and Application to Auto Insurance Fraud Detection", ***Journal of Finance and Data Science***, Vol. 2, No. 1, (2016), 58-75, doi: 10.1016/j.jfds.2016.03.001.
7. Boriah. S, Chandola. V, and Kumar. V, "Similarity measures for categorical data: A comparative evaluation***", In Proceedings of the eighth SIAM International Conference on Data Mining***, 243-254, (2008).
8. Gartner. T, Le. Q. V, and Smola. A. J, "A short tour of kernel methods for graphs", ***Technical report, NICTA***, Australia, Canberra, (2006).
9. Couto. J, "Kernel k-means for categorical data", ***Lecture Notes in Computer Science, Springer***, (2005), 46-56, doi: 10.1007/11552253_5.
10. Gower. J. C, "A General Coefficient of Similarity and Some of Its Properties", ***Biometrics***, Vol. 27, No. 4, (1971), 857-871, doi: 10.2307/2528823.
11. Subudhi. Sh, and Panigrahi. S, "Detection of Automobile Insurance Fraud Using Feature Selection and Data Mining Techniques", ***International Journal of Rough Sets and Data Analysis***, Vol. 5, No. 3, (2018), 1-20, doi: 10.4018/IJRSDA.2018070101.
12. Wang. Y, and Xu. W, "Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud",

Decision Support Systems, Elsevier, Vol. 105, (2018), 87-95, doi: 10.1016/j.dss.2017.11.001.

13. Li. Y, Yan. C, Liu. W, and Li. M, "A principle component analysis-based random forest with the potential nearest neighbor method for automobile insurance fraud identification", ***Applied Soft Computing, Elsevier***, Vol. 70, (2018), 1000-1009, doi: 10.1016/j.asoc.2017.07.027.
14. Bosagh-Zadeh. R, Meng. X, Ulanov. A, Yavus. B, Pu. L, Venkataraman. Sh, Sparks. E, Staple. A, Zaharia. M, "Matrix computations and optimization in Apache Spark", ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 31-38, (2016).
15. Tennyson. Sh, and Salsas-Forn. P, "Claims Auditing in Automobile Insurance: Fraud Detection and Deterrence Objectives", ***The Journal of Risk and Insurance***, Vol. 69, No. 3, (2002), 289-308, doi: 10.1111/1539-6975.00024.
16. Itri. B, Mohamed. Y, Mohamed. Q, and Omar. B, "Performance comparative study of machine learning algorithms for automobile insurance fraud detection", Third International Conference on Intelligent Computing in Data Sciences (ICDS), (2019).
17. Roy. R, George. K. Th, "Detecting insurance claims fraud using machine learning techniques", International Conference on Circuit ,Power and Computing Technologies (ICCPCT), (2017).
18. Stephen-Kalwihura. J, and Logesvaran. R, "Auto-Insurance fraud detection: a behavioral feature engineering approach", ***Journal of Critical Reviews***, Vol. 7, No. 3, (2020), 125-129, doi: 10.31838/jcr.07.03.23.
19. Abdallah. A, Aaizaini-Maarof. M, and Zainal. A, "Fraud detection system: A survey", ***Journal of Network and Computer Applications***, Vol. 68, (2016), 90-113, doi: 10.1016/j.jnca.2016.04.007.
20. Phua. C, Lee. V, Smith. K, and Gayler. R, "A Comprehensive Survey of Data Mining-based Fraud Detection Research***", Computers in Human Behavior***, Vol. 28, (2012), 1002–1013, doi: 10.1016/j.chb.2012.01.002.
21. Phua. C, Alahakoon. D, and Lee. V, "Minority report in fraud detection: classification of skewed data", ***ACM SIGKDD Explorations***, Vol. 6, No. 1, (2004), 50-59, 10.1145/1007730.1007738.
22. H. Hassanpour, and A. Darvishi, "A Geometric View of Similarity Measures in Data Mining", ***International Journal of Engineering, Transactions C: Aspects*** Vol. 28, No. 12, (2015), 1728-1737. doi: 10.5829/idosi.ije.2015.28.12c.05

Persian Abstract

چکیده

جمع‌آوری کردن نمونه‌های تقلب بیمه هزینه‌بر است و درصورتیکه به طور دستی انجام شود بسیار زمانبر خواهد بود. این امر استفاده از روش‌های یادگیری ماشین بدون ناظر را می‌طلبد. یکی از روش‌های دقیق در حوزه کشف تقلب بیمه خودرو، روش رتبه‌بندی طیفی آنومالی است که دارای دقت بسیار بالاتری نسبت به روش‌های معروف دیگر بوده است. با این حال، این روش در مواجهه با داده های حجیم، مقیاس‌پذیری کافی ندارد و برای کشف برخط آنومالی مناسب نیست. جهت جلوگیری از خسارت‌های هنگفت، سیستم‌های مدیریت تقلب برخط ضروری هستند. در این مطالعه، ما یک متدلوژی پیاده‌سازی را پیشنهاد می‌کنیم که استفاده از الگوریتم رتبه‌بندی طیفی را برای کلان داده ممکن می‌سازد. ما از توانایی روش رتبه بندی طیفی آنومالی، جهت تولید متغیر هدف تخمینی برای داده بدون برچسب استفاده می‌کنیم. سپس، از این داده دارای برچسب تخمینی برای آموزش دو مدل رگرسیون پایدار جنگل تصادفی و شبکه عصبی عمیق استفاده می‌کنیم. در مرحله بعد داده ورودی بدون برچسب، به مدل‌های آموزش‌دیده اعمال می‌شود و برچسب تخمینی به دست می‌آید. نتایج شبیه‌سازی‌ها نشان می‌دهد که روش پیشنهادی دارای سرعت بسیاری زیادی در کنار نرخ مثبت کاذب قابل قبول می‌باشد.