Attribute Noise, Classification Technique, and Classification Accuracy

R. Indika P. Wickramasinghe

Abstract Binary data classification is an integral part in cyber-security, as most of the response variables follow a binary nature. The accuracy of data classification depends on various aspects. Though the data classification technique has a major impact on classification accuracy, the nature of the data also matters lot. One of the main concerns that can hinder the classification accuracy is the availability of noise. Therefore, both choosing the appropriate data classification technique and the identification of noise in the data are equally important. The aim of this study is bidirectional. At first, we aim to study the influence of noise on the accurate data classification. Secondly, we strive to improve the classification accuracy by handling the noise. To this end, we compare several classification techniques and propose a novel noise removal algorithm. Our study is based on the collected data about online credit-card transactions. According to the empirical outcomes, we find that the noise hinders the classification accuracy significantly. In addition, the results indicate that the accuracy of data classification depends on the quality of the data and the used classification technique. Out of the selected classification techniques, Random Forest performs better than its counterparts. Furthermore, experimental evidence suggests that the classification accuracy of noised data can be improved by the appropriate selection of the sizes of training and testing data samples. Our proposed simple noise-removal algorithm shows higher performance and the percentage of noise removal significantly depends on the selected bin size.

1 Introduction

Quality of the data takes an upmost importance in data analytics, irrespective of the type of application. In data classification, high quality data are further important as the accuracy of the classification correlates with the quality of data.

Real-world datasets can contain noise, which is one of the reasons that makes the quality of data low [16, 38]. Shahi et al. [39] consider outliers and noise as the

R. Indika P. Wickramasinghe (✉)
Department of Mathematics, Prairie View A&M University, 100 University Dr, Prairie View, TX 77446, USA
e-mail: iprathnathungalage@pvamu.edu

I. Palomares et al. (eds.), *Data Analytics and Decision Support for Cybersecurity*, Data Analytics, DOI 10.1007/978-3-319-59439-2_7

uncertainty of data. Handling data that are mixed with noise brings a hard time for the analysis. Some organizations allocate millions of dollars per year on detecting errors with data [32]. When the noise is mixed with cyber-security related data, one needs to give a serious attention to handle them due to the sensitive nature of the data.

The attention for cyber-security and protective measures grew faster with the expansion of cyberattacks. Stealing intellectual and financial resources are the main reasons behind the deliberate breaching of computer systems. When the mobile commerce expanded revolutionary, a series of thefts started to creep up in which majority of them are related to credit-cards transactions. In a study, Hwang et al. [19] states that approximately 60% of the American adults avoid doing business online due their concern about misuse of the personal information. Riem [33] reports a precarious incident of shutting down a credit-card site of a British bank called, Halifax due to the exposure of consumers' details. In this regard, measurements to minimize credit-card related frauds are in great demand at present than ever before.

Noise hinders the classification [35, 46, 49] by decreasing the performance accuracy, in terms of time in constructing the classifier, and in the size of the classifier. Thist is why the identification of noise is an integral part in data classification. Noise can be introduced in many ways into online transactions of credit card data. Besides the conventional ways that introduce noise, research indicates that magnetic card chips can be vulnerable at times and can introduce noise to the transaction. Therefore, identification and isolation of noise from the data before analyzing is very important.

Apart from the noise, the shape of the attributes' distributions can make an impact on the quality of data classification. It is a fact that most of the natural-continuous random variables adhere some sort of Gaussian distribution. When the data show a departure from the normality, it is considered as skewed. Osborne [31] points out that mistake in data entry, missing data values, presence of outliers, and nature of the variable itself are some of the reasons for the skewness of the data. Furthermore, Osborne [31] makes use of several data transformation techniques such as square root, natural log, and inverse transformation to convert the non-normal data into normal. There is a strong association between the existence of noise and the skewness of the distribution. It is apparent that skewness of data directly influence outliers. Hence, there should be an important connection between the nature of the skewness of data and the classification accuracy.

Even if someone removes the noise and the skewness of the data, it would not completely reach the maximum data accuracy level in the classification. Selection of the correct classification technique based on the available data, and the use of appropriate sample sizes for training and test data are two of options one can consider for improvement of classification. Though there are several literature findings about the identification of appropriate classification technique, only a handful of findings exists in connection with the selection of suitable sample ratios. Even within the available findings, none of them are related to both cyber-security related data that are mixed with noise. In this chapter we have two broad aims. At first, we study the impact of noise in effective data classification. Secondly, we aim

to improve the classification of noisy data. To this end, we consider how skewness, appropriate ratios of the samples, and the classification technique impact on the classification. This study brings the novelty in two ways. According to the author's knowledge, it is rare to find a study aiming to investigate the relationship between above selected features and the classification accuracy. Furthermore, we propose a novel, simple and most importantly an effective noise detection algorithm to improve the classification accuracy. The rest of the chapter is organized as follows: Next Sect. 2 discusses the related work, and Sect. 3 provides the background to classification techniques. In Sect. 4, the dataset is described. The Sect. 5 aims to discuss the issue of attribute noise on classification accuracy. Section 6 investigates how skewness of the data influences the classification accuracy. Section 7 attempts to improve the classification accuracy of the data, which is mixed with noise. This is achieved by using the noise removal and the selection of appropriate sample sizes for the training and testing samples. Then Sect. 8 discusses the results of the study and Sect. 9 concludes the chapter.

2 Related Work

Use of SVM in data classification is not novel and has mixture of opinions regarding the accuracy of it. Sahin and Duman [36] incorporated the knowledge of decision trees and SVM in credit card fraud analysis. Though they found the model based on decision trees outperformed SVM, the difference of performances between both methods became less with the increment of the size of the training datasets. In another study, Wei and Yuan [45] used an optimized SVM model to detect online fraudulent credit card and found their proposed non-linear SVM model performed better than the others. Colas and Brazdil [8] and Fabrice and Villa [13] compared SVM with K-nearest neighbor (kNN) and naive Bayes in text classification. Though they expected that SVM would outperform its counterparts, authors couldn't find SVM as the clear winner. In addition, they pointed out that the performance of kNN continues to improve with the use of suitable preprocessing technique. Scholkopf and Smola [37] and Mennatallah et al. [28] utilized SVM in anomaly detection. In addition to SVM based techniques, other alternative techniques can be found in the literature.

Abu-Nimeh et al. [1] compared six classifiers on phishing data. In their study, authors used Logistic Regression, Classification and Regression Trees, Bayesian Additive Regression Trees, Support Vector Machines, Random Forests, and Neural Networks. According to the outcomes, authors claimed that the performance of Logistic Regression was better than the rest.

As the previous research findings indicate, the nature of the data is imperative for the accuracy of classification. Bragging and boosting are considered as popular classifying trees and random forest was proposed by adding an extra layer to bragging [25]. Díaz-Uriarte and Andres [12] incorporated random forest in gene

classification and stated that this technique showed an excellent performance in classification even with the presence of noise in data. Miranda et al. [29] compared three noise elimination algorithms for Bioinformatics datasets and Machine Learning classifiers. In another study, Jayavelu and Bar [20] used Fourier series approach to construct a noise removal algorithm. An effective two-phased algorithm for noise-detection is proposed by Zhu et al. [50]. A novel noise-detection algorithm, which is based on fast search-and-find density peaks is proposed in [48]. The authors clustered the original data before removing the outliers. Another outlier detection algorithm on uncertain data is proposed by Wang et al. [44]. This algorithm is based on the Dynamic Programming Approach (DPA). Cao et al. [6] and Kathiresan and Vasanthi [22] conducted their research about handling noise in credit card data. Lee et al. [24] explored the patterns of credit card transactions in order to classify fraudulent transactions.

When the quality of the data is considered, the shape of the distribution of data is very important as the majority of data analysis techniques assume the symmetric nature of the distribution. Akbani et al. [2] suggested that SVMs do not perform too badly with moderately skewed data, compared to the other available machine learning algorithms. Further studies [27, 34, 40, 41] can be seen in the literature. In addition to the nature of the dataset, the ratio of trainee and test data can impact on the quality of classification. Guyon [14] proposed an algorithm regarding the sizes of training and validation datasets, but the author stated that this framework is not perfect due to the use of simplifying assumptions. Beleites et al. [3] studied about the above issue using learning curves for small sample size situations.

3 Background on Classification Techniques and Measurement Indicators

The aim of this section is to describe the theoretical foundation that is used in this study. First of all, the section starts describing the four types of classification techniques that are used in this study. Secondly, quantitative measurements that are used to compare the four classification techniques are discussed.

3.1 Classification Techniques

Here we consider four types of popular classification techniques, namely Support Vector Machines (SVM), Principal Component Analysis (PCA), Robust Principal Component Analysis (RPCA), and Random Forest. These techniques attempt to categorize the class membership based on the attributes of the given dataset by capturing the hidden patterns of the existing dataset. Ultimately they predict the group membership of novel data.

3.1.1 Support Vector Machines (SVM)

SVM is considered as one of the most popular classification techniques [7], which was introduced by Vapnik [43]. This is based on statistical learning theory and it attempts to separate two types of data (Class A and Class B) using a hyper-plane by maximizing the boundary of the separation, as shown in the Fig. 1.

Consider the n-tuple training dataset, S.

$$S = \{(x_1, y_1), (x_2, y_2), \ldots., (x_i, y_i), \ldots., (x_n, y_n)\}. \tag{1}$$

Here the set of feature vector space is M-dimensional and the class variable is 2-demesioinal. i.e., $\boldsymbol{x}_i \in R^M$ and $\boldsymbol{y}_i \in \{-1, +1\}$. As mentioned before, the ultimate aim of the SVM is to find the optimal hyper-plane that split the dataset into two categories.

If the feature space is completely linearly separable, then the optimal separating hyper-plane can be found by solving the following Linear Programming (LP) problem:

$$\begin{gathered} Min\ ||w||^2 \\ s.t. y_i\left((x_i.w) + b\right) \geq 1, \\ i = 1, 2, \ldots, n \end{gathered} \tag{2}$$

Unfortunately not all the datasets can be linearly separable. Therefore, SVM uses a mapping called the *Kernel*. The purpose of the Kernel, Φ is to project the linearly inseparable feature space into a higher dimension so that it can be linearly separated in the higher space. This feature space transformation is taken place according to the following.

$$w.\Phi(x) + b = 0 \tag{3}$$

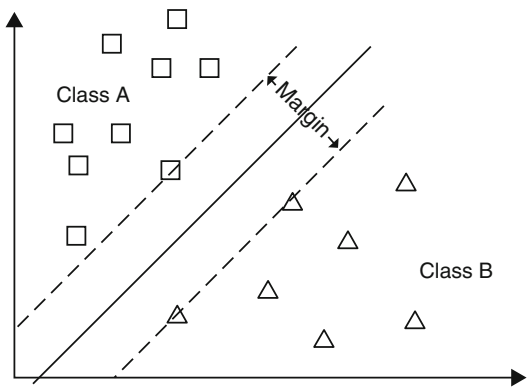

Fig. 1 Support vector machine

where w represents the weight vector, which is normal to the hyperplane. Following are the most popular existing kernel functions that have been extensively studied in the literature.

- Linear Kernel, $\Phi\left(x_i, x_j\right) = x_i^T x_j$
- Polynomial Kernel, $\Phi\left(x_i, x_j\right) = \left(\gamma x_i^T x_j + 1\right)^d$, for $\gamma > 0$
- Radial Kernel, $\Phi(x_i, x_j) = \exp(-\gamma ||x_i - x_j||^2)$, for $\gamma > 0$
- Sigmoid Kernel, $\Phi\left(x_i, x_j\right) = Tanh\left(\alpha x_i^T x_j + r\right)$.

3.1.2 Principal Component Analysis

PCA is an unsupervised and data compression technique used to extract relevant information from a complexed data. The main purpose of PCA is to use in dimension reduction and feature selection [11, 30] by preserving the majority of the characteristics of the initial data, which has been used in various fields. PCA attempts to transform correlated variables into a set of linearly uncorrelated variables in an optimal way. These new set of variables, which is low in dimension is called principal components. This is accomplished via a series of vector transformations as explained in the following algorithm.

Step 1: Begin with the starting variable (the set of features), $X = (X_1, X_2, \ldots., X_n)'$

Step 2: Rotation of the above variables into a new set of variables called, $Y = (Y_1, Y_2, \ldots., Y_n)'$ so that $Y_i's$ are uncorrelated and $\text{var}(Y_1) \geq \text{var}(Y_2) \geq \ldots$ $\text{var}(Y_n) \geq$.

Step 3: Y_i and Y_{i+1} that are constructed so that $\alpha_i' = \left(\alpha_{1j}, \alpha_{2j}, \ldots., \alpha_{pj}\right)$, $Y_i = \alpha_i' X$ and $\sum \alpha_i^2 = 1$.

3.1.3 Robust Principal Component Analysis (RPCA)

PCA is considered as not robust enough to work with data when outliers are present. Halouska and Powers [15] investigated the impact of PCA on noise data related to Nuclear Magnetic Resonance (NMR). They concluded that a very small oscillation in the noise of the NMR spectrum can cause a large variation of PCA, which results in the form of irrelevant clustering. Robust Principal Component Analysis, a modified version of popular PCA method attempts to recover the low-rank matrix from the corrupted measurements. Though there are numerous versions of RPCA, Hubert et al. [18] proposed a method that can be described briefly as follows.

- At first, use singular value decomposition is used to deduce the data space.
- Next, for each data point a measurement of outlyingness is calculated. Out of all the n data points (suppose there are m smallest measurements), the covariance matrix, Σ_m is calculated. In addition, k number of principal components are chosen to retain.

- Finally, a subspace spanned by largest k number of eigenvalues corresponding to the k number of eigenvalues of the covariance matrix, Σ_m. Then all the data are projected onto the above sub space and the robust principal components are computed based on the location of the projected data points.

3.1.4 Random Forest

Random Forest is considered as one of the most popular and widely applied machine learning algorithms in data classification. Though the main use of Random Forest is for classification, it can be used as a useful regression model. Random Forest technique is an ensemble learning technique, proposed by Breiman [4]. Ensemble methods are considered as learning algorithms that builds a series of classifiers to classify a novel data point. This technique has found an answer to the overfitting problem available in individual decision trees (Fig. 2).

The entire classification process of Random Forest is achieved in a series of steps as described below.

- Suppose the number of training dataset contains N cases. A sub set from the above N is taken at random with replacement. These will be used as the training set to grow the tree.
- Assume there are M number of input variables. Then a number m, which is lower than M is selected and m number of variables from the collection of M is selected randomly. After that the best split of the selected m variables is selected to split the node. Throughout this process, m is kept as a constant.
- Without pruning, each tree is grown to the largest extent. The prediction of a new data point is by aggregating predictions.

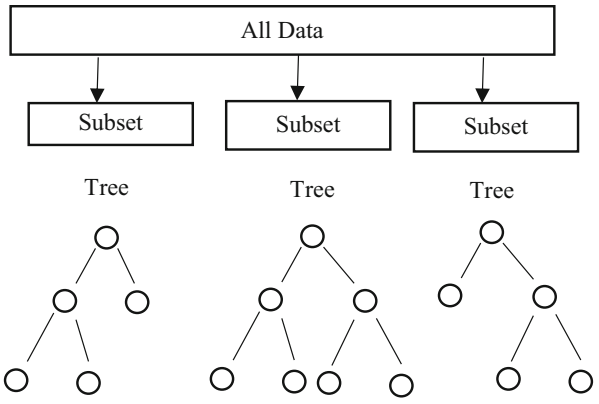

Fig. 2 Random forest

Table 1 True positive and false positive

Label	Meaning
TP-true positive	Positive items that are correctly classified as positives
TN-true negatives	Negative items that are correctly identified as negatives
FP-false positives	Negative items that are wrongly classified as positives
FN-false negatives	Positive items that are wrongly classified as negatives

3.2 Performance Indicators

As a means of quantifying the clarity of classification, following quantitative measurements are discussed. For this purpose, we consider several frequently used performance indicators, namely precision, specificity, sensitivity (recall), and F-Measure.

We adhere the following naming convention that is used to summarize the standard confusion matrix used in data classification (Table 1).

Let's assume P to be the total number of positive instances and N be the total number of negative instances. Then the performance indicators ca be defined as follows.

$$\text{Sensitivity} = \frac{FP}{P}$$

$$\text{Specificity} = \frac{FP}{N}$$

$$\text{Precision} = \frac{TP}{FP + TP}; \tag{4}$$

where P represents total number of positive (Class A), instances while N represents the total number of class B instances.

$$\text{F} - \text{Measure} = 2\left[\frac{Precision^{*}Sensitivity}{Precision + Sensitivity}\right]$$

4 Dataset

In this study we utilize a dataset about online transactions using credit cards. This secondary dataset has been modified from the initial dataset, which contains credit cards' transactions by European credit cards holders within two days in September 2013. This dataset includes 29 features including time, amount, and the time duration of the transaction in seconds. Though it is interesting to know all the included attribute names in the dataset, due to the confidentiality issues the data do not disclose all the background information.

The response variable indicates whether there was an involvement of a fraud with the credit card transaction or not. Therefore, the feature (Class) takes value 1 to represent an occurrence of a fraud in the credit card transaction, while 0 represents the opposite. Due to the highly unbalanced nature of the dataset, the initial dataset was modified to prepare dataset of 1980 instances comprising of a nearly balanced dataset (997 fraud instances and 983 non-fraud instances).

5 Examining the Impact of Attribute Noise, and the Classification Technique on Classification Accuracy

5.1 The Noise

Data can be classified as attributes and class labels [49]. When noise is considered, it may either relate to independent variable (attribute) or to the dependent variable (class). Hence, if the noise exists in the data it can be classified as either attribute noise or class noise [23, 26, 42, 47, 49]. According to the literature findings, class noise brings more adverse effects than the attribute noise for classification. Despite the fact that class noise creates more damages than the attribute noise, it is believed that the latter is more complicate to handle. Therefore, we focus on studying attribute noise and the impact of it on data classification.

5.2 Attribute Noise, Classification Technique, and Classification Accuracy

Impact of attribute noise, and the classification technique on classification accuracy is tested in several phases. Four previously stated (Sect. 3.1) classification techniques were selected for this. As the first approach, credit card data was classified using the conventional SVM. In the second approach, PCA was used for data reduction before applying SVM. Third technique is exactly similar to the second, in which replacement of PCA by RPCA was the only difference. In the final technique, Random Forest is directly applied on the dataset similar to the first approach. After the application of each technique, sensitivity and the F-measure were calculated for the comparison.

Table 2 Average F-measure, classification methods, and noise

Method	5%	10%	15%	20%
SVM	0.48	0.92	0.96	1.82
PCA	1.56	0.33	0.62	1.20
RPCA	1.08	0.59	0.14	0.40
RForest	0.04	0.03	0.07	0.07

In the next phase random noise was introduced to the dataset in 5%, 10%, 15%, and 20% levels and F-measure was measured accordingly. Using the F-measures in each case, the percent change of F-measure was calculated. Table 2 summarizes this findings. In the implementation of algorithm, training and testing samples were generated according to the 70:30 ratio and performance indicators were calculated based on 100 randomly selected samples.

6 Skewness of the Data and Classification Accuracy

6.1 Skewness

Skewness provides a measure about the symmetry of the distribution. This measurement can be either positive or negative, in either case, skewness makes the symmetric distribution asymmetric. Transformation of skewed data into symmetric is often seen in data analysis. Though there is no strict set of guidelines regarding the type of transformation to use, following are some of the popular transformations. As Howell [17] suggested, if data are positively skewed *sqrt*(X) is used. If the skewness is moderate, then the log(X) can be used. Brown [5] stated the identification of skewness based on the Joanes and Gill [21] and Cramer [10] formulas as described below.

Consider a sample of n data points, $x_1, x_2, \ldots, x_n$. Then the method of moment coefficient of skewness is computed according the Eq. (5).

$$g = \frac{m_3}{m_2^{\frac{3}{2}}}; \text{where } m_3 = \frac{\sum \left(x - \bar{x}\right)^3}{n}, m_2 = \frac{\sum \left(x - \bar{x}\right)^2}{n} \tag{5}$$

Using the above g, the sample skewness (G) is calculated according to the Eq. (6).

$$G = g\frac{\sqrt{n\,(n-1)}}{(n-2)} \tag{6}$$

Finally, the declaration of the skewness of the data is decided based on the value of Z_g. Here,

$$Z_g = \frac{G}{SES}; \text{where SES, Standard Error of Skewness}$$

$$SES = \sqrt{\frac{6n\,(n-1)}{(n-2)\,(n+1)\,(n+3)}}$$

Therefore, we can classify the data as a negatively skewed if $Z_g < -2$. If $|Z_g| < 2$, then the distribution is either symmetric, negatively skewed or positively skewed. Finally if $|Z_g| > 2$, we can classify that the distribution is positively skewed.

6.2 Impact of Skewness on Classification

With the idea of quantifying the impact of skewness on classification accuracy, we generate data samples with noise levels 5%, 10%, 15%, and 20%. Each attribute in each sample is tested for the skewness as explained in Sect. 6.1. If the skewness is present, it is removed using an appropriate transformation. SVM, PCA, RPCA and RForest methods are applied afterwards. At the end, percent change of sensitivity and percent change of F-measure are calculated and these results can be seen in Table 3. In addition, Figs. 3 and 4 displays the observed quantitative measures graphically.

Table 3 Noise level, classification method, and the change of measures

Noise level	Method	% Change of sensitivity	% Change F-measure
5%	SVM	4.24	1.71
	PCA	1.20	0.44
	RPCA	0.20	0.32
	R FOREST	0.06	0.06
10%	SVM	1.68	1.32
	PCA	1.00	0.20
	RPCA	1.36	0.90
	R FOREST	0.08	0.08
15%	SVM	2.33	0.73
	PCA	0.14	0.55
	RPCA	2.40	0.87
	R FOREST	0.11	0.11
20%	SVM	2.23	0.50
	PCA	0.78	0.27
	RPCA	0.47	0.84
	R FOREST	0.05	0.05

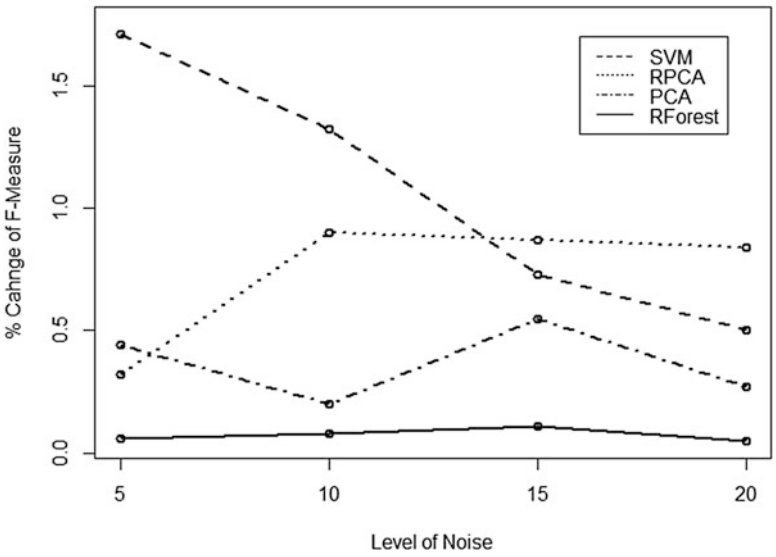

Fig. 3 Impact of skewness on F-measure

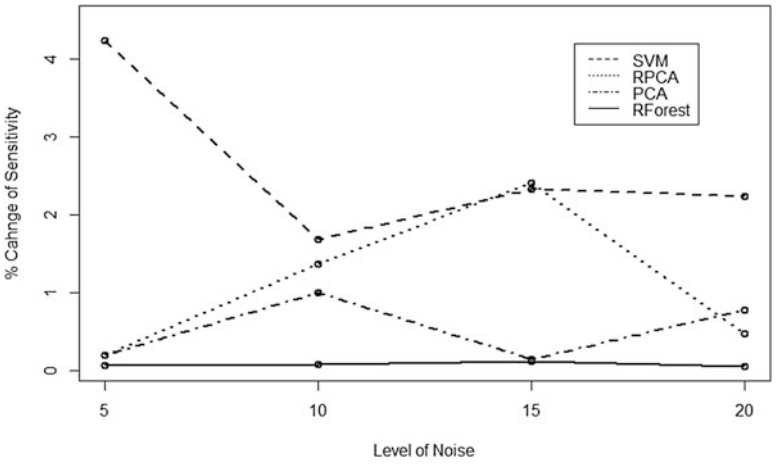

Fig. 4 Impact of skewness on sensitivity

7 Improving the Classification Accuracy When Noise Is Present

Improving the classification accuracy can be achieved in many ways. Appropriate model selection, purity of the data, and the use of efficient algorithms are some of the ways. Though much attention has not given, selection of appropriate sample sizes (training and testing samples) plays a major role in classification accuracy. With the

Table 4 Pseudo code of the algorithm for noise removal

```
Input : Data file with noise
Output: New data file after removing noise
   1.  For i=1 to (# of attributes)
   2.   For j=1 to (# observations)
   3.     For k=1 to (# bins)
   4.       bin the i^th attribute with selected bin
                 sizes of 5, 10,15,25, and 50)
   5.       Calculate the mean and standard
                 deviation of k^th  bin
   6.       Convert each data points on bin k
                 and attribute i into their
                 corresponding Z-scores
   7.    End k
   8.   End j
   9.  End i
  10. Summarizing the z-scores of each instance to
      a number and test for outliers using the
      method in section 5.1
  11. Eliminate Outliers
```

presence of noise of the data, noise removal algorithm may be critically important for the classification accuracy. In this section, we aim to improve the classification accuracy by proposing a novel, but simple noise removal algorithm. This algorithm is tested using the credit card data and compare with one of the standard outlier detection method, based on the Cook's [9] distance.

7.1 A Simple Noise Removal Algorithm

As can be seen in the algorithm, the algorithm is executed in several steps. At first, the data in each attribute is binned according the selected bin sizes (5, 10, 15, 25, and 50). In the next step, data in each bin is converted into their corresponding sample z-scores by treating the entire bin as the sample. This process continues until the algorithm covers the entire dataset by taking each attribute at a time. After completing the z-score calculation for each attribute, standard outlier detection algorithm, as explained in Sect. 6.1 is applied. In the last step, outliers are removed from the dataset. The pseudo code of this proposed algorithm is displayed in Table 4.

This algorithm is implemented under each classification method on data with 5%, 10%, 15%, and 20% noise levels. Performance indicators are recorded before and after the implementation of the algorithm. Obtained results are shown in Figs. 5 and 6.

The effectiveness of this algorithm is compared with the Cook's distance approach. Logistic regression is fitted on the above data and Cook's distance is calculated on each data point. Then the data points are declared as unusual if

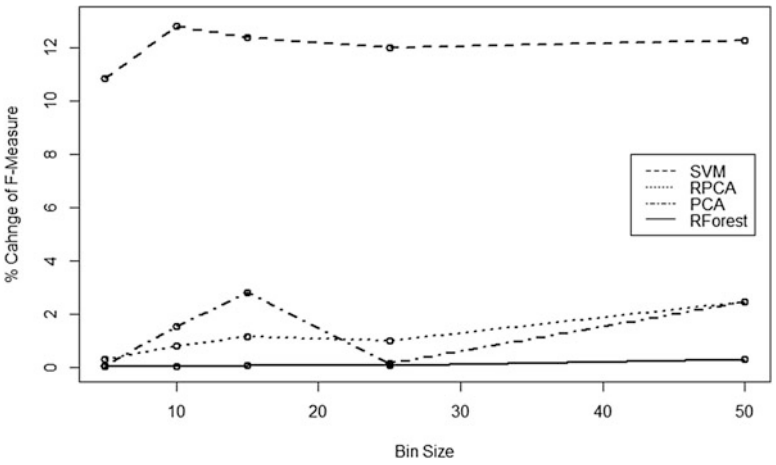

Fig. 5 Change (%) of F-measure, Bin Size, and Classification Method for Noise Level of 5%

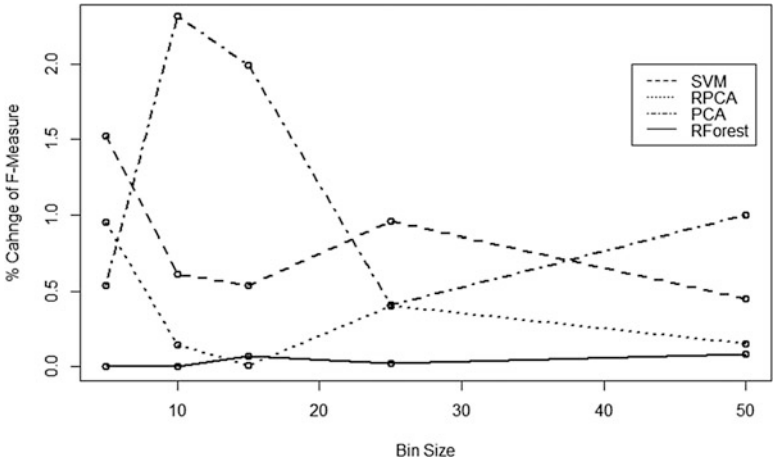

Fig. 6 Change (%) of F-measure, bin size, and classificationmethod for noise level of 15%

the Cook's distance is more than the ratio of 4*/ (total number of instances in the dataset)*. This is implemented for all the datasets of 5%, 10%, 15%, and 20% noise levels.

7.2 Impact of the Sample Size Ratio of Training and Testing on Classification Accuracy

It is evident that the classification accuracy would be biased if a large portion of data is selected to train the model by leaving very small portion to test the model. This may inflate standard deviations of the performance indicators. Therefore,

identification of appropriate balance between the two sample sizes is very crucial. Therefore, it would be beneficial to understand how these ratio of samples impact the classification accuracy, when the data are mixed with noise.

We study this using nine sample ratios of training-test datasets. 90:10, 80:20, 70:30, 60:40, 50:50, 4:60, 30:70, 20:80, and 10:90 are the selected sample ratios. For each case, we consider datasets with 5%, 10%, 15%, and 20% noise levels. After simulating 100 instances, the average and the standard deviations of each measurement indicator were calculated at each noise levels. In addition, standard deviation and the mean values of F-measure are calculated.

7.2.1 Coefficient of Variation (CV)

When comparing two statistics, with different distributions for their standard deviations and means, the ratio between standard deviation (σ) and the mean (μ) values is considered as a better measurement. This measurement is considered as the Coefficient of Variation (CV), which can be calculated according to the Eq. 8. CV quantifies the dispersion of the statistics compared to its mean value.

$$CV = 100^* \frac{\sigma}{\mu} \tag{8}$$

After calculating above measurements, Figs. 7 and 8 display the fings.

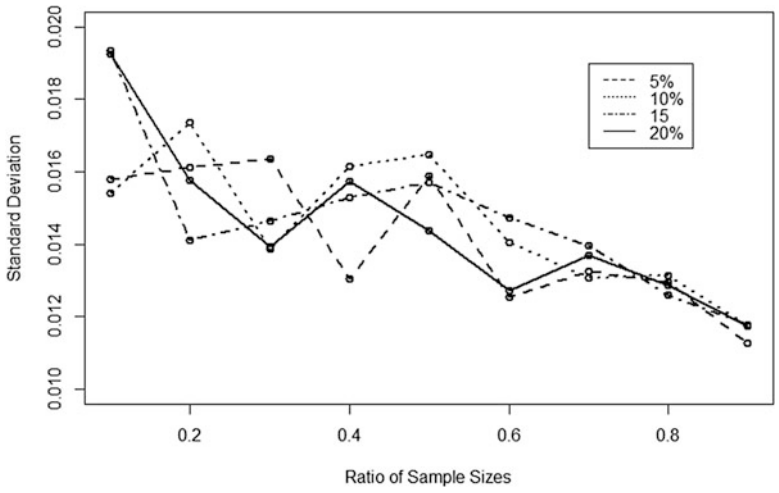

Fig. 7 Training: test ratio, noise levels and standard deviation of the F-measure

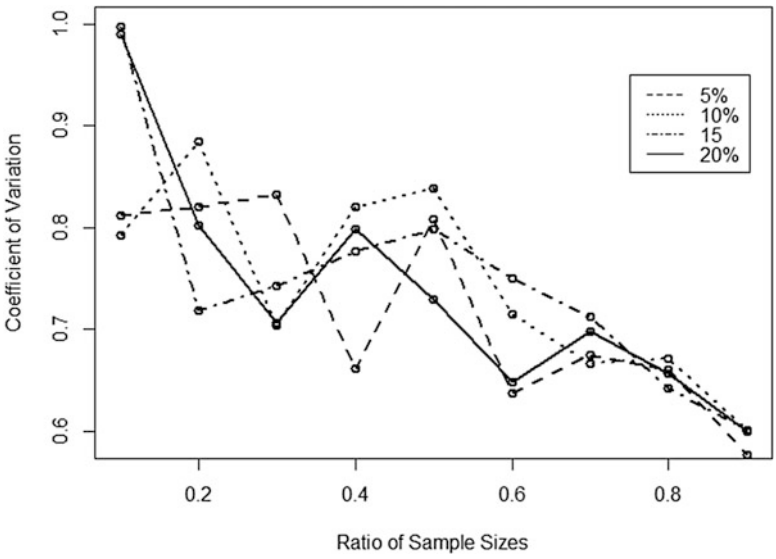

Fig. 8 Training: test ratio, noise levels and standard deviation of the F-measure

8 Results

Table 2 shows the average F-measure for each classification method at each noise levels. Smaller values are indicating robustness of the model to the noise. According to the Table 2 and outcomes of the Analysis of Variance (ANOVA), it is obvious that there is a significant difference in the values of RForest method, compared to the other counterparts $[F(3, 9) = 4.064\,, p = 0.0442]$. This means, Random Forest method outperforms the rest of the alternatives across all the levels of noises. Though SVM works better when there is less noise, with the increment of noise RPCA works better than SVM.

Table 3 summarizes how skewness of the data influences the classification accuracy, in the presence of noise. Even in this case, Random Forest method shows the smallest change in both sensitivity and the F-measure. This indicates that Random Forest method handles skewness of the data well. SVM shows that it is not a robust classification technique for skewed data in comparison to the other counterparts. Most importantly, the performance of SVM gets better with the increment of noise levels. Out of PCA and RPCA, the latter performs better than the former. This suggests that PCA is a better classification technique for noisy and skewed data. This behavior is clearly explained by the Figs. 3 and 4.

Figures 5 and 6 display the relationship among the bin size, classification technique, noise level and the classification accuracy. At 5% noise level, the highest change of F-measure across all bin levels is recorded by SVM. Clearly, this is significantly different than other methods. This indicates that classification accuracy can

be significantly enhanced using the proposed algorithm, with SVM in particularly. As the Fig. 6 and other results indicate, PCA also performs better with this algorithm for higher levels of noise. All the methods, except the Random Forest show significant improvement of F-measures. This indicates that classification accuracy can be improved by implementing this noise removal algorithm. Furthermore, there exists a clear connection between the bin size and the classification accuracy. When the bin size is either 10 or 20, the highest performance can be achieved across all the classification techniques and all noise levels. ANOVA was conducted to evaluate the performance of capturing the noise using our proposed method. According to the findings there is a significant effect of bin size on the percentage of noise capturing at the $p < 0.05$ level for the five bin sizes $[F(4, 12) = 232.18, p = 0]$. When comparing effectiveness of the proposed algorithm and the Cook's distance method, there is significant evidence to claim that the proposed method captures higher percentage of noise than the alternative at $p < 0.05$, $[F(1, 6) = 35.01, p = 0.001]$. Finally, when studying the relationship between ratio of sizes for training-test datasets and the classification accuracy, Figs. 7 and 8 provide interesting findings. When the ratio is small, there is a higher variability of the performance indicator (F-measure) compared to its average. When the sample ratio is increased, the coefficient of variation decreases irrespective of the noise level. All the above stated outcomes were based on 100 iterations in each of the appropriate situation.

9 Conclusion and Future Work

In this empirical study we investigated the impact of attribute noise in the data on the classification accuracy. At first, we studied the influence of attribute noise on the classification accuracy. Secondly, we tried to enhance the power of data classification in the presence of noise. Hence, we collected a dataset about online transactions using credit-card, which is related to cyber-security. With this dataset, we classify whether a transaction has been involved a fraud or not. According to our findings, it is clear that classification accuracy is hindered by the presence of noise in the data. Furthermore, Random Forest method outperforms the other three classification techniques even in the presence of noise. SVM seems better than other two in the absence of Random Forest, but when the level of noise increases, RPCA performs better than the SVM. When testing the influence of skewness on data classification, we found that there is a direct impact from skewness on the classification accuracy. This influence affects differently on each technique. Among the selected techniques, Random Forest is robust even when data are skewed. Though SVM looks vulnerable to classify skewed data, when the noise level of the data is higher, SVM performs better. Further analysis shows that PCA can classify data well when the data are noisy and skewed. As a means of improving the classicization accuracy for noisy data, we proposed a simple noise removal algorithm. According to the obtained results, our algorithm significantly improve the classification accuracy. This was compared with the Cook's distance approach. According to the obtained results,

our proposed algorithm shows better performances compared to the Cooks' distance approach. Further analysis indicates that there is a strong relationship between the selected bin size and the classification accuracy. Though this influence does not impact all the classification techniques uniformly, bin sizes 10 and 20 record higher performances than other bin sizes. At last, we studied about the appropriate ratio of sample sizes for both training and test datasets. According to the outcomes of this study, there is a strong connection between the ratio of datasets and the classification accuracy. When selecting appropriate sample sizes, one needs to pay attention about this ratios. As the results indicate, if the ratio is too small the variability of the performance indicator inflates. In cyber-security related data, enhancing the performance even in small amount will be advantageous. Though we have improved the classification accuracy significantly, this obtained outcomes motivate further research work to explore inclusion of novel classification techniques. In addition, future work will be conducted to see the influence other factors on the accuracy of data classification. This current study was conducted for the balanced data, therefore it would be interesting to extend this study for unbalanced data as well. Furthermore, an extension of the noise removal algorithm to address class noise will be beneficial too.

References

1. Abu-Nimeh, S., Nappa, D., Wang, X., & Nair, S. A comparison of machine learning techniques for phishing detection. In Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit, pp. 60-69, ACM. (2007).
2. Akbani R., Kwek S., and Japkowicz N.: "Applying support vector machines to imbalanced datasets," in Proceedings of the 15th European Conference on Machine Learning, pp. 39–50, (2004).
3. Beleites C., Neugebauer U., Bocklitz T., Krafft C., Popp J.: Sample size planning for classification models. Anal Chim Acta. Vol. (760), pp. 25–33, (2013).
4. Breiman L.: Random forests. Machine Learning, Vol. 45(1), pp. 5–32, (2001).
5. Brown S., Measures of Shape: Skewness and Kurtosis, https://brownmath.com/stat/shape.htm, (2008-2016)
6. Cao Y., Pan X., and Chen Y.: "SafePay: Protecting against Credit Card Forgery with Existing Card Readers", in Proc. IEEE Conference on Communications and Network Security, pp. 164–172, (2015).
7. Carrizosa, E., Martin-Barragan, B., Morales, D. R.: Binarized support vector machines. INFORMS Journal on Computing, Vol. 22(1), pp. 154–167, (2010).
8. Colas F., and Brazdil P.,"Comparison of SVM and Some OlderClassification algorithms in Text Classification Tasks", "IFIP International Federation for Information Processing", Springer Boston Volume 217, Artificial Intelligence in Theory and Practice, pp. 169–178, (2006).
9. Cook, R. D.: "Influential Observations in Linear Regression". Journal of the American Statistical Association. Vol. 74 (365), pp. 169–174, (1979).
10. Cramer, Duncan Basic statistics for social research: step-by-step calculations and computer techniques using Minitab. Routledge, London.; New York, (1997).
11. Cureton, Edward E, and Ralph B. D'Agostino. Factor Analysis, an Applied Approach. Hillsdale, N.J: L. Erlbaum Associates, (1983).

12. Díaz-Uriarte R., De Andres, S. A.: Gene selection and classification of microarray data using random forest. BMC bioinformatics, 7(1), p. 3, (2006).
13. Fabrice, R, Villa, N.: Support vector machine for functional data classification. Neurocomputing/EEG Neurocomputing, Elsevier, 69 (7–9), pp.730–742, (2006).
14. Guyon I.: A scaling law for the validation-set training-set size ratio, AT & T Bell Laboratories, Berkeley, Calif, USA, (1997).
15. Halouska S., Powers R.: Negative impact of noise on the principal component analysis of NMR data, Journal of Magnetic Resonance Vol. (178) (1), pp. 88–95, (2006).
16. Hickey R. J., "Noise modelling and evaluating learning from examples," Artif. Intell., vol. 82, nos. 1–2, pp. 157–179, (1996).
17. Howell, D. C. Statistical methods for psychology (6th ed.). Belmont, CA: Thomson Wadsworth, (2007).
18. Hubert, M., Rousseeuw, P. J., Branden, K. V.: ROBPCA: a new approach to robust principal components analysis, *Technometrics*, vol. 47, pp. 64–79, (2005).
19. Hwang, J. J., Yeh, T. C., Li, J. B.: Securing on-line credit card payments without disclosing privacy information. Computer Standards & Interfaces, Vol. 25(2), pp. 119-129, (2003).
20. Jayavelu D., Bar N.: A Noise Removal Algorithm for Time Series Microarray Data. In: Correia L, Reis L, Cascalho J, editors. Progress in Artificial Intelligence, vol. 8154. Berlin: Springer, pp. 152–62, (2013).
21. Joanes, D. N., Gill C. A.: "Comparing Measures of Sample Skewness and Kurtosis". The Statistician Vol. 47(1), pp. 183–189, (1998).
22. Kathiresan K., Vasanthi N. A., Outlier Detection on Financial Card or Online Transaction data using Manhattan Distance based Algorithm, International Journal of Contemporary Research in Computer Science and Technology (IJCRCST) Vol. 2(12), (2016).
23. Khoshgoftaar T., Hulse J. V.: Identifying noise in an attribute of interest. In ICMLA '05: Proceedings of the Fourth International Conference on Machine Learning and Applications (ICMLA'05), pp. 55–62, Washington, DC, USA, 2005. IEEE Computer Society. ISBN 0-7695-2495-8. doi: 10.1109/ICMLA.2005.39, (2005).
24. Lee C. C., Yoon J. W.: "A data mining approach using transaction patterns for card fraud detection", Seoul, Republic of Korea, pp. 1-12, (2013).
25. Liaw A., Wiener M.: Classification and Regression by Random Forest, R News, Vol. 2(3), (2002).
26. Liebchen G.: Data Cl eaning Techniques for Software Engineering Data Sets. Doctoral thesis, Brunel University, (2011).
27. Maratea A., Petrosino, A.: Asymmetric kernel scaling for imbalanced data classification, in: Proceedings of the 9th International Conference on Fuzzy Logic and Applications, Trani, Italy, pp. 196–203, (2011).
28. Mennatallah A., Goldstein M., Abdennadher, S.: Enhancing oneclass support vector machines for unsupervised anomaly detection. In Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description pp. 8–15, (2013).
29. Miranda A. L., Garcia L. P., Carvalho A. C., Lorena A. C., "Use of classification algorithms in noise detection and elimination", Proc. 4th Int. Conf. Hybrid Artif. Intell. Syst., pp. 417–424, (2009).
30. Oja, E.: Principal components, minor components, and linear neural networks. Neural Networks, pp. 927–935, (1992).
31. Osborne, J. Notes on the use of data transformations. Practical Assessment, Research & Evaluation, 8(6). http://PAREonline.net/getvn.asp?v=8&n=6, (2002).
32. Redman, T.: Data Quality for the Information Age. Artech House, (1996).
33. Riem, A.: Cybercrimes of the 21st century: crimes against the individual—part 1, Computer Fraud and Security. Vol 6, pp. 13–17, (2001).
34. Rosenberg A.: "Classifying Skewed Data: Importance Weighting to Optimize Average Recall," Proc. Conf. Int'l Speech Comm. Assoc. (InterSpeech '12), (2012).
35. Sáez, J.A., Galar M., Luengo, J. et al. Analyzing the presence of noise in multi-class problems: alleviating its influence with the One-vs-One decomposition. Knowl Inf Syst 38: 179. doi: 10.1007/s10115-012-0570-1, (2014).

36. Sahin Y., Duman E.: "Detecting Credit Card Fraud by Decision Trees and Support Vector Machines", International Multi-conference of Engineers and computer scientists, (2011).
37. Scholkopf, B., Smola A. J.: Support Vector Machines and Kernel Algorithms, The Handbook of Brain Theory and Neural Networks. MIT Press, Cambridge, UK, (2002).
38. Seo S.: Masters thesis. University of Pittsburgh; Pennsylvania: A review and comparison of methods for detecting outliers in univariate data sets, (2006).
39. Shahi A., Atan R. B., Sulaiman M. N.: Detecting effectiveness of outliers and noisy data on fuzzy system using FCM. Eur J Sci Res 36: pp. 627–638, (2009).
40. Siddiqui F., and Ali, Q. M.: Performance of non-parametric classifiers on highly skewed data, Global Journal of Pure and Applied Mathematics. ISSN 0973-1768 Vol. 12(2), pp. 1547–1565, (2016).
41. Tang L., Liu H.: Bias analysis in text classification for highly skewed data. In ICDM '05: Proceedings of the Fifth IEEE International Conference on Data Mining, IEEE Computer Society, pp. 781–784, (2005).
42. Teng M. C.: Combining noise correction with feature selection. pp. 340–349, (2003).
43. Vapnik V., *The Nature of Statistical Learning Theory*. Springer-Verlag, ISBN 0-387-98780-0, (1995).
44. Wang, Bin, et al. "Distance-based outlier detection on uncertain data." Ninth IEEE International Conference on Computer and Information Technology, 2009. CIT'09. Vol. 1. IEEE, (2009).
45. Wei X., and Yuan L.: "An Optimized SVM Model for Detection of Fraudulent Online Credit Card Transactions," International Conference on Management of e-Commerce and e-Government, 2012.
46. Xiong H., Pandey G., Steinbach M, Kumar V.: "Enhancing data analysis with noise removal," IEEE Trans. Knowl. Data Eng., Vol. 18(3), pp. 304–319, (2006).
47. Yoon K., Bae D.: A pattern-based outlier detection method identifying abnormal attributes in software project data. Inf. Softw. Technol., Vol. 52(2), pp. 137–151. ISSN 0950-5849. (2010).
48. Zhou X., Zhang Y., Hao S., Li S., "A new approach for noise data detection based on cluster and information entropy." The 5th Annual IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems, (2015).
49. Zhu X., Wu X., Class noise vs. attribute noise: a quantitative study, Artificial Intelligence Review Vol. 22 (3). pp.177–210, (2004).
50. Zhu, X., Wu X., Yang, Y.: Error Detection and Impact-sensitive Instance Ranking in Noisy Datasets. In Proceedings of 19th National conference on Artificial Intelligence (AAAI-2004), San Jose, CA. (2004).