Research Article

A hybrid model for class noise detection using k-means and classification filtering algorithms

Zahra Nematzadeh[1] · Roliana Ibrahim[1] · Ali Selamat[2,3,4]

Received: 31 March 2020 / Accepted: 22 June 2020 / Published online: 29 June 2020

Abstract
Real data may have a considerable amount of noise produced by error in data collection, transmission and storage. The noisy training data set increases the training time and complexity of the induced machine learning model, which led to reduce the overall performance. Identifying noisy instances and then eliminating or correcting them are useful techniques in data mining research. This paper investigates misclassified instances issues and proposes a clustering-based and classification filtering algorithm (CLCF) in noise detection and classification model. It applies the k-means clustering technique for noise detection, and then five different classification filtering algorithms are applied for noise filtering. It also employs two well-known techniques for noise classification, namely, removing and relabeling. To evaluate the performance of the CLCF model, several experiments were conducted on four binary data sets. The proposed technique was found to be successful in classify class noisy instances, which is significantly effective for decision making system in several domains such as medical areas. The results shows that the proposed model led to a significant performance improvement compared with before performing noise filtering.

Keywords Clustering · Classification filtering · Class noise detection · K-means

1 Introduction

One of the main elements for successful learning and knowledge discovery in data mining is data quality. Data cleansing can be done manually which is difficult, time consuming and inclined to errors and noise. So, effective automatic tools are necessary in data cleansing process. Noise refers to the inaccuracies and inconsistencies of data, which reduces the quality of the real data. Besides, noise can affect the quality of information extracted from the data, as well as the models created from the data and the decisions made by the data [1]. Identifying noisy instances and then eliminating or correcting them are useful techniques in data mining research [2]. Eliminating noisy samples from training sets may improve data reliability and quality [3]. Noise detection is the critical part for data understanding and cleaning, as well as semi-supervised outlier detection [4]. Noise filtering is used to eliminate incorrect instances from real-life data. Noise reduction is a difficult and important process in machine learning to achieve precise and high performance models. If the noisy data is not removed, it might yield wrong decision [5]. Effective noise detection process decreases the risk of poor decision making using erroneous data [6]. Noise in data categorized into attribute noise and class noise or combination of both categories. Attribute noise is related to the errors or unusual values and class noise is wrong class label. Several experimental researches shows

✉ Zahra Nematzadeh, zahra_nematzadeh@yahoo.com | [1]School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia (UTM), Johor Baharu, Malaysia. [2]School of Computing, Faculty of Engineering, UTM and Media and Games Center of Excellence (MagicX), Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia. [3]Center for Basic and Applied Research, Faculty of Informatics and Management, University of Hradec Kralove, Rokitanskeho 62, 500 03 Hradec Králové, Czech Republic. [4]Malaysia Japan International Institute of Technology, Universiti Teknologi Malaysia, 54100 Kuala Lumpur, Malaysia.

SN Applied Sciences (2020) 2:1303 | https://doi.org/10.1007/s42452-020-3129-x

that class noise has negative effects on the performance of machine learning classifiers [7]. Class noise is known as the major challenge in data mining research, which has negative effects on the performance of the model. Enhancing classification accuracy of induced models is known as the main issue of noise detection techniques [8]. It is also clear that classification accuracy extremely depends on the quality of the training set [1].

The review of the existing studies shows that many researchers have proposed methods to handle the noise in the data sets using machine learning algorithms [3, 8–10]. Sluban et al. [8] developed new class noise detection algorithms including the high agreement random forest filter on two UCI data sets. Xiong H et al. [11] explored four approaches to increase data analysis through noise removal using unsupervised techniques. Lowongtrakool and Hiransakolwong [5] developed unsupervised clustering intelligence method to reduce the quantity of spam. The outcomes from noise filtering were beneficial for data processing which makes them more precise. Zeidat et al. [12] compared several popular data set editing techniques which are Wilson editing, Citation editing, and multi-edit. They also introduced supervised clustering editing. Smith et al. [13] identified the reasons cause instances to be misclassified. Moreover, Thongkam et al. [14] applied SVM on training set to detect and eliminate all samples which misclassified by the SVM. Jeatrakul et al. [15] also applied same approach using neural networks. The proposed cleaning method enhances the confidence of cleaning noisy training instances. Likewise, it is important to have good classifiers in classification filtering and existence of class noise produce poor classifiers [16]. Since SVM removes the instances that their prediction is not reliable, a local support vector machines (LSVM) noise reduction technique is proposed by Segata et al. [17]. According to Segata et al. [18], a new strategy is proposed to reduce the number of local SVMs for noise reduction process. A neural network based automatic noise reduction (ANR) was presented to clean noisy instances in data sets by Martinez et al. [19]. Sánchez et al. [20] applied K-nearest neighbor classifier (KNN) to predict data sets and then, the misclassified instances are removed. Sabzevari et al. [21] applied randomized ensembles such as bagging and random forest for detecting and handling class noisy instances in training set. The results showed that removing is better than relabeling when the noise levels are low and medium and relabeling is more precise at high noise levels. Also, there are studies which figured out the most important factors that deteriorate the performance of the k-means algorithm. Fränti and Sieranoja [22] found that if the clusters overlap, the choice of initialization technique does not matter much, and repeated k-means is usually good enough for the application. However, if the data has well-separated clusters, the result of k-means depends merely on the initialization algorithm. Since, the performance of evolutionary k-means often decreased by noisy data, a clustering stability-based EKM algorithm (CSEKM) which evolves partitions and the aggregated matrices simultaneously was proposed by He and Yu [23]. The experimental results show that the CSEKM is more robust to noise.

Based on these studies, two main issues are investigated. First, there is a lack of attention to the misclassified instances which has a great impact on the clustering efficiency. Second, removing class noise may affect the classification performance, which highlights to have a good and reasonable classification filtering for noise detection. This paper aims to extend our previous model, namely, the k-means support vector machine [24]. It also proposes a CLCF model using k-means clustering algorithm and five different classification filtering algorithms on four real data sets to recognize the class noisy instances. Furthermore, the proposed model increases the clustering efficiency and overall performance. The model is constructed in three phases. The first phase is noise detection, which is based on clustering technique to identify misclassified instances in each cluster. The second phase is noise filtering, which applies five classification filtering algorithms to obtain the real noisy instances. Third phase is noise classification that employs two different techniques, namely, the removing and relabeling for classifying noisy instances. Experiments were conducted to measure the performance of the model using evaluation criteria. Figure 1 presents the general view of proposed model.

In brief, the main contributions of the present study are as follows:

1. Investigating misclassified instances issues in k-means clustering algorithm.

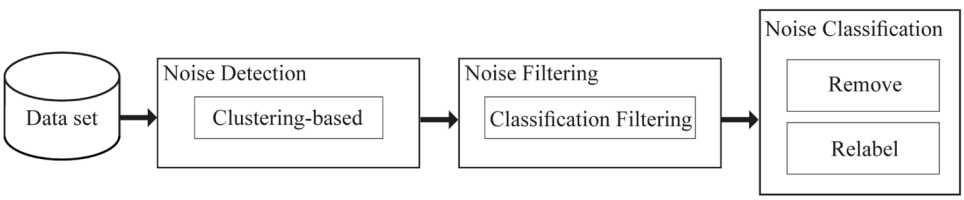

Fig. 1 General view of the CLCF model

2. Proposing a new CLCF model that comprises k-means clustering algorithm as well as classification filtering algorithms for class noise detection in binary datsets.

The paper is organized as follows. Section 2 presents the preliminaries knowledge. Section 3 describes the proposed model. The data sets and performance measurement are described in Sect. 4. The discussion on the results is described in Sect. 5. Finally, Sect. 6 concludes this paper with a brief summary and suggestions for future works.

2 Preliminaries knowledge

In this section, the methods required for noise detection and noise filtering are introduced. The selection of classifiers applied for the filtering is explained as well.

2.1 K-means clustering algorithm

In this research, one crisp clustering technique, namely, k-means is applied to recognize the misclassified instances, which are then assumed as noisy instances. Applying the k-means is very common because it is theoretically simple and memory efficient and is computationally fast [25]. The flowchart of k-means clustering algorithm is illustrated in Fig. 2.

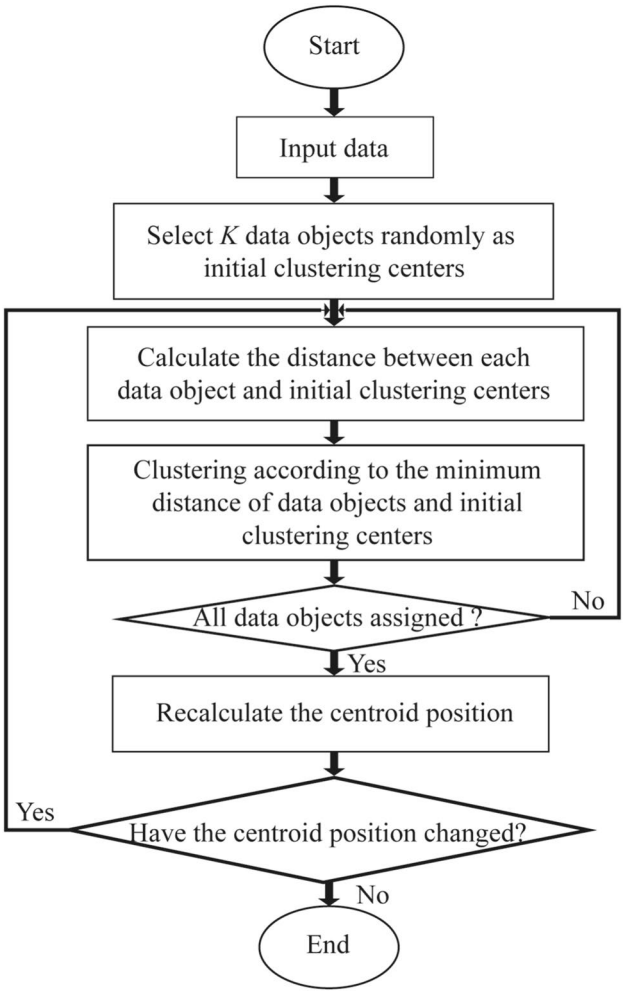

Fig. 2 The flowchart of k-means algorithm

2.2 Classification filtering algorithms

In this study, five classifiers with different learning paradigms among the most popular supervised learning techniques to identify noisy instances [26, 27] were used as classification filtering algorithms.

2.2.1 Support vector machine (SVM)

The basis of the support vector machine (SVM) [28] is the procedure of learning a linear hyperplane from a training set separating positive examples from negative ones. SVM can be considered as a binary classifier. In addition, in numerous existing works such as [4, 8, 18, 29–31], SVM has been used for detecting noise. SVM is a popular classification filtering method broadly used to detect class noise [2].

2.2.2 Naïve Bayes (NB)

One of the statistical classifier is Bayesian classifier which is known as simple probabilistic classifiers. By using this technique, the probability of an instance which belongs to a particular class is predicted [32]. Many existing works have applied NB for noise detection such as [4, 7, 8, 26].

2.2.3 Random forest (RF500)

The random forest (RF) learner with 500 decision trees was considered in this study because of its strong performance compared to well-known classifier [33]. To construct randomized decision tree, the RF classifier uses bagging and the 'random subspace method' [34]. The outputs of ensembles of these randomized, unpruned decision trees are combined to produce the final prediction. Many existing works have applied RF for noise detection such as [4, 8, 26, 31].

2.2.4 K-nearest neighbor (KNN)

The k-nearest neighbor classifier reduces hyper spheres in the space of instances by allocating the majority class of the k-nearest instances based on a defined metric [35]. It is an effective, simple classification algorithm [36] and

it has been widely used in the domain of noise detection [3, 4, 18, 31, 37–41].

2.2.5 Neural network (NN)

Multilayer perceptron (MLP) is a technique which feed forward neural networks trained with the standard back propagation procedure. They need to train favorable results since they are supervised networks. The MLPs is used in majority of neural network applications [42]. Neural network has been widely used in the domain of noise detection [3, 5, 19, 26, 38, 43]. It is well-known classification filtering methods for class noise detection [2].

3 Proposed ClCF model for class noise detection and classification

The proposed CLCF model is described for the detection of noisy instances. This model consists of three main phases: noise detection, noise filtering, and noise classification. Figure 3 illustrates the overall architecture of the proposed model. The clustering technique and classification filtering are integrated to detect and filter noisy data. The model phases are explained in detail next.

3.1 Phase 1: noise detection (clustering-based)

In this phase, the k-means (KM) clustering technique [44] is applied on four real data sets to recognize the misclassified instances. K-means clustering technique distributes input vectors into separated clusters by means of similarity and distance measurement [45]. All input vectors are assembled into distinct centers by means of minimizing objective function based on Eq. 1.

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} \left(x_j - \mu_i\right)^2 \quad (1)$$

where k is the number of cluster (S_i) and i = 1, 2 … k, and μ_i displays the centers of the clusters. First, the intensity distribution is computed, and then initial centroids are created using K random intensities. The following equation shows the iterative algorithm for clustering based on their intensities.

$$c^{(i)} = \min_j x^{(i)} - \mu_j^2 \quad (2)$$

The misclassified instances referred to the lowest number of a certain class label in each cluster, which is then counted as class noise.

Let the noisy data "X" consists of n datum $(x_1, y_1), \ldots, (x_n, y_n)$, $x \in R^n$ and $y \in \{1, -1\}$. $Y = \{y_a = x_i \mid L(x_i) = 1\}$ where $a = 1, \ldots, A$ (A is the number of samples that their labels are +1), $p = \{p_t = x_i \mid L(x_i) = -1\}$ where $t = 1, \ldots, U$ (U is the number of samples that their labels are −1) and $n = U + A$. $L(x_i)$ represents the class label for each sample $L(x_i) = \{label(x_i) \mid Label(x_i) = 1 or -1\}$.

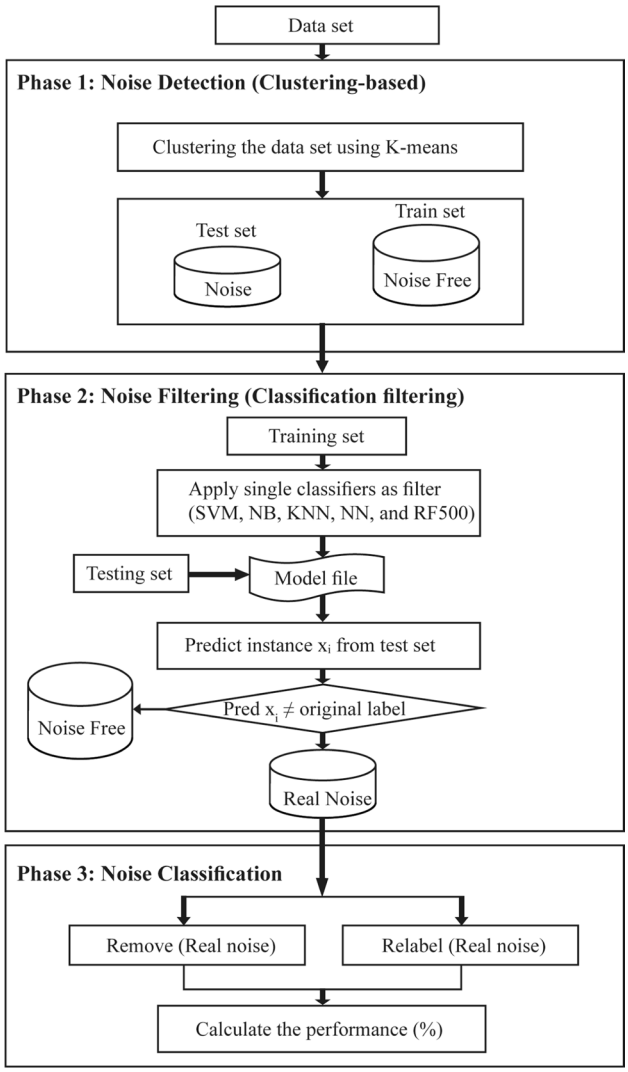

Fig. 3 Overall architecture of the CLCF model

Definition 1 Suppose "M" is a cluster includes instances with class label "+1" $M = \{L(x_a) | a = 1, \ldots, A\}$ and "H" is a cluster includes instances with class label "−1 "$H = \{L(x_t) | t = 1, \ldots, U\}$. Assume "M" is a cluster which is partitioned into two classes M_1 and $M_2 = M - M_1$, where $|M| = b$, $|M_1| = b_1$ and $|M_2| = |M - M_1| = b_2$. Then, the following statements are used to detect noisy instances:

$$(b_1 < b_2) \Rightarrow (\forall x_i \in M_1,\ x_i \text{ is noise}) \wedge (\forall x_i \in M_2,\ x_i \text{ is noisefree}) \quad (3)$$

$$(b_2 < b_1) \Rightarrow (\forall x_i \in M_2,\ x_i \text{ is noise}) \wedge (\forall x_i \in M_1,\ x_i \text{ is noisefree}) \quad (4)$$

3.2 Phase 2: noise filtering (classification filtering)

In this phase, five classification filtering algorithms are applied to detect noisy instances. These classifiers are support vector machine (SVM), random forests 500 (RF500), Naïve Bayes (NB), neural network (NN) and K-nearest neighbor (KNN, k = 10) respectively. Based on the first phase, the noisy and noise free sets are detected. Then, the noise set is considered as the testing set (T) and the noise free set is considered as the training set (Tr). The training data sets are separately classified using each classifier to create a model. The testing data sets are then predicted based on the created models. If the predicted label for each testing instance is not equal with its original label, the instance is known as "real noise" otherwise "noise free". The classification filtering problem is presented as follows:

Definition 2 Suppose φ is a classifier algorithm and $B = \varphi(T, Tr) = \{b_i\}_{i=1}^{n}$ and b_i is the predicted label of instance (x_i) from the test set (T) and $|T| = n$. The Eq. (5) demonstrates how the real noisy instances are identified. The procedure of classification filtering is presented in Fig. 4 to simplify the Definition 2. It shows how classification filtering can detect noisy instances in data sets.

$$b_i = \begin{cases} L(x_i) & x_i \text{ is noise free} \\ -L(x_i) & x_i \text{ is noise} \end{cases} \quad (5)$$

3.3 Phase 3: noise classification

Two approaches are used to deal with noisy samples, which are "removing" and "relabeling" techniques. The removing approach omits all detected noisy samples after noise filtering procedure and produces a new decreased data set. The relabeling approach assigns a new label to all detected noisy objects after noise filtering procedure by switching their label and keeps the original size of the data set. The proposed CLCF algorithm is illustrated in Fig. 5.

4 Experimental setup

The experimental data sets and the performance evaluation criteria used in this study are discussed here. The accuracy of the CLCF algorithms in terms of removing and relabeling techniques on the Pima, Heart (statlog), Wisconsin and Ionosphere data sets [6] are presented as well. The experiment was applied based on 10 runs for each data set to achieve average evaluation criteria. The average performance was calculated in terms of accuracy using SVM-RBF kernel algorithm and tenfold cross validation.

4.1 Data sets

To test and evaluate the CLCF model, four real experimental data sets were used. Three medical data set namely Pima, Wisconsin and Heart (statlog) along with one non-medical data set namely ionosphere are used from UCI

Algorithm 2: CLCF Algorithm

Initialize: f=true, D is data set, k is number of clusters, C is cluster, N is noise set, NF is noise free, Tr is training set, T is test set, RN is real noise set, RNF is real noise free set

Output: Acc_remove, Acc_relabel

```
1: while f=true do
2:        C_k=clustering (D,K)
3: for i=1 to k do
4:        if num(C_i1 < C_i2 ) then
5:              N ← C_i1   and  NF← C_i2
6:        else
7:              NF←C_i1   and   N←C_i2
8:     End if
9:End for
10: Tr ← NF
11: T← N
12: (RN, RNF)=classification filtering (T,Tr)
13:  Acc_remove=SVM (RNF)
14:  Acc_relabel= SVM(RNF,RN)
15: End while
```

Fig. 5 Proposed CLCF algorithm

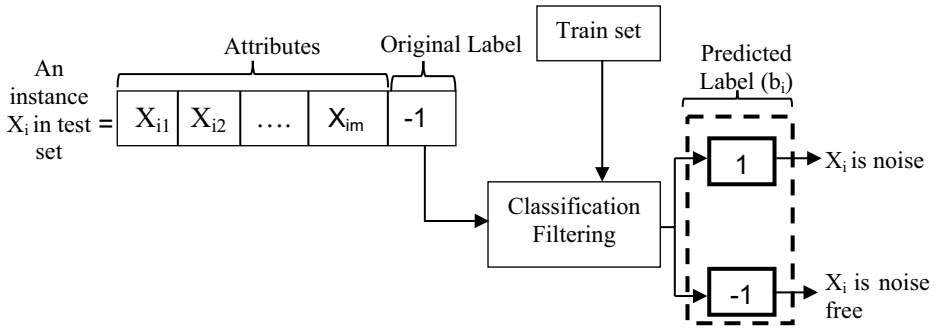

Fig. 4 The procedure of classification filtering

Table 1 Distribution of data sets [6]

Data set	#Ex	#Features	#Class
Pima	768	8	2
Wisconsin	683	9	2
Heart (statlog)	270	13	2
Ionosphere	351	34	2

repository [46]. We used one non-medical data set in order to evaluate our proposed model in different areas. All the data sets are related to binary classification problem. Table 1 lists the data sets used in this research with the number of classes (#Class), number of features (#Feature), and number of examples (#Ex).

4.2 Performance measures

The accuracy formula is applied to calculate the performance of the proposed technique in classification [24, 48] using the confusion matrix. In the following formula, True Negative refers to correctly rejected samples, True Positive (TP) refers to correctly identified samples, False Positive refers to incorrectly identified samples and False Negative (FN) means incorrectly rejected samples:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (6)$$

5 Results and discussion

The accuracy of each data set on CLCF model using k-means with five classification filtering algorithms are analyzed separately and illustrated in Figs. 6, 7, 8 and 9. The best K = 3 is determined experimentally from K = [2, 10]. The results of four data sets including the Pima, Wisconsin, Heart and Ionosphere data sets are illustrated and explained as follow.

The accuracy achieved by five different algorithms, namely, the KM-SVM, KM-KNN, KM-NB, KM-RF500 and KM-NN with two different classification techniques, namely, the removing and relabeling on the Pima data set are illustrated in Fig. 6. Although the KM-KNN with 90.843% accuracy was the best algorithm using the removing technique in all five CLCF algorithms, the Fig. 6 shows that the relabeling technique outperformed the removing technique in all the five CLCF algorithms. Finally, the best accuracy which is highlighted in Table 2 is achieved by KM-RF500 using relabeling technique with 94.817% on the Pima data set.

The accuracy achieved by five different algorithms, namely, the KM-SVM, KM-KNN, KM-NB, KM-RF500 and

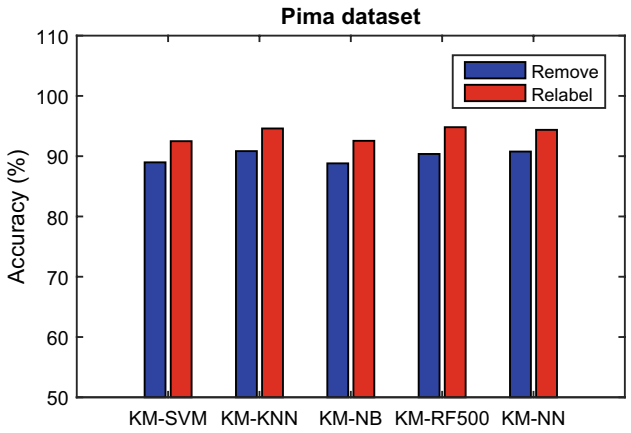

Fig. 6 Comparing accuracy of the CLCF model using k-means with five classification filtering algorithms on the Pima data set

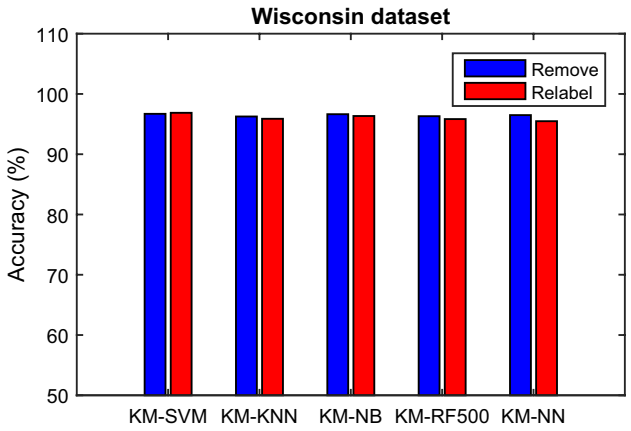

Fig. 7 Comparing accuracy of the CLCF model using k-means with five classification filtering algorithms on the Wisconsin data set

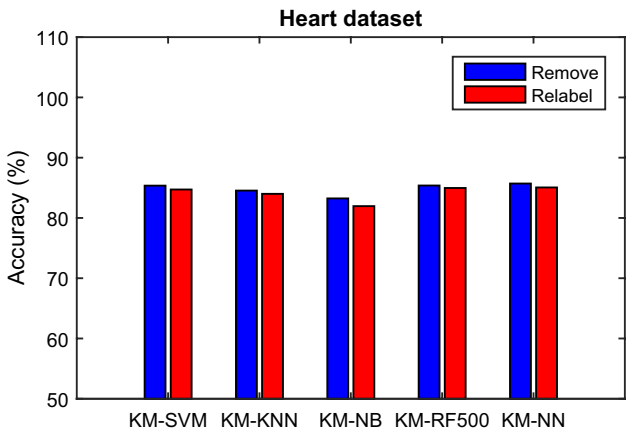

Fig. 8 Comparing accuracy of the CLCF model using k-means with five classification filtering algorithms on the Heart data set

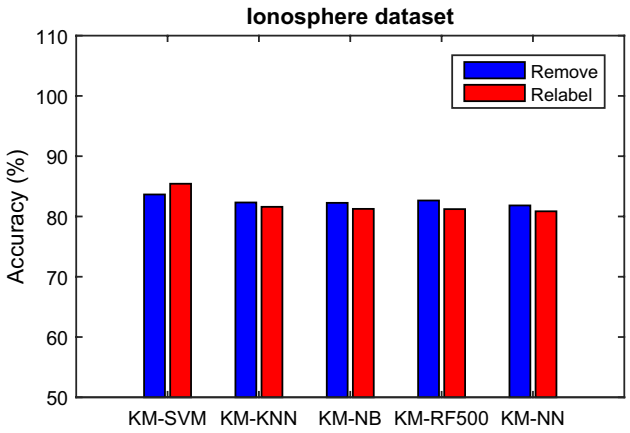

Fig. 9 Comparing accuracy of the CLCF model using k-means with five classification filtering algorithms on the Ionosphere data set

Table 2 Accuracy of the removing and relabeling techniques of the CLCF model using k-means with five classification filtering algorithms on the Pima data set

Data set	Methods	Accuracy	
		Remove	Relabel
Pima	KM-SVM	88.987 ± 0.035	92.497 ± 0.034
	KM-KNN	90.843 ± 0.038	94.606 ± 0.023
	KM-NB	88.811 ± 0.038	92.550 ± 0.025
	KM-RF500	90.371 ± 0.039	**94.817 ± 0.027**
	KM-NN	90.770 ± 0.037	94.374 ± 0.025

Table 3 Accuracy of the removing and relabeling techniques of the CLCF model using k-means with five classification filtering algorithms on the Wisconsin data set

Data set	Methods	Accuracy	
		Remove	Relabel
Wisconsin	KM-SVM	96.704 ± 0.024	**96.877 ± 0.024**
	KM-KNN	96.262 ± 0.024	95.880 ± 0.031
	KM-NB	96.634 ± 0.025	96.345 ± 0.023
	KM-F500	96.316 ± 0.025	95.830 ± 0.027
	KM-NN	96.497 ± 0.025	95.483 ± 0.027

KM-NN with two different classification techniques, namely, the removing and relabeling on the Wisconsin data set are illustrated in Fig. 7. Although The KM-SVM with 96.704% accuracy was higher using the removal technique in all five CLCF algorithms, but KM-SVM using relabeling outperformed the removing technique. Finally, the best accuracy which is highlighted in Table 3 is achieved by KM-SVM using relabeling technique with 96.877% on the Wisconsin data set.

Table 4 Accuracy of the removing and relabeling techniques of the CLCF model using k-means with five classification filtering algorithms on the Heart data set

Data set	Methods	Accuracy	
		Remove	Relabel
Heart	KM-SVM	85.360 ± 0.069	84.723 ± 0.060
	KM-KNN	84.543 ± 0.060	84.001 ± 0.054
	KM-NB	83.243 ± 0.067	81.965 ± 0.062
	KM-F500	85.389 ± 0.054	84.977 ± 0.046
	KM-NN	**85.715 ± 0.060**	85.066 ± 0.062

Table 5 Accuracy of the removing and relabeling techniques of the CLCF model using k-means with five classification filtering algorithms on the Ionosphere data set

Data set	Methods	Accuracy	
		Remove	Relabel
Ionosphere	KM-SVM	83.652 ± 0.055	**85.439 ± 0.058**
	KM-KNN	82.320 ± 0.065	81.603 ± 0.063
	KM-NB	82.262 ± 0.055	81.265 ± 0.060
	KM-F500	82.651 ± 0.073	81.220 ± 0.065
	KM-NN	81.839 ± 0.067	80.869 ± 0.062

The accuracy achieved by five different algorithms, namely, the KM-SVM, KM-KNN, KM-NB, KM-RF500 and KM-NN with two different classification techniques, namely, the removing and relabeling on the Heart data set are illustrated in Fig. 8. The KM-NN with 85.715% accuracy was higher using the removing technique in all five CLCF algorithms. Also, the KM-NN with 85.066% using relabeling was higher in all five different algorithms. Finally, the best accuracy which is highlighted in Table 4 is achieved by KM-NN using removing technique with 85.715% on the heart data set.

The accuracy achieved by five different algorithms, namely, the KM-SVM, KM-KNN, KM-NB, KM-RF500 and KM-NN with two different classification techniques, namely, the removing and relabeling on the Ionosphere data set are illustrated in Fig. 9. Although, the removing technique outperformed the relabeling technique in four CLCF algorithms but the KM-SVM using relabeling with 85.439% outperformed the removing technique. The best accuracy is highlighted in Table 5.

To guide the interpretation of Figs. 6, 7, 8 and 9, a comparison of the results from the CLCF model using k-means with five classification filtering algorithms are presented in Tables 2, 3, 4 and 5 for each data set (with the best results highlighted in bold and their standard deviations). It shows the best results in terms of accuracy for each data set. Based on the empirical results, the best CLCF algorithm

is KM-RF500 for the Pima data set using the relabeling technique. However, the removing technique achieved the better result in the Wisconsin data set, the KM-SVM using the relabeling outperformed comparing with other techniques. The best CLCF algorithm is KM-NN using the removing technique for the Heart data set. Although, the removing technique achieved the better result in the Ionosphere data set, the KM-SVM using the relabeling outperformed comparing with other techniques.

We analyzed and compared the best CLCF algorithm of each data set (after noise reduction) with the results of each data set before noise reduction. All the four data set are evaluated before noise reduction using SVM-RBF kernel algorithm and tenfold cross validation. Table 6 shows a comparison between accuracy of data sets before and after noise detection. The results from this table present that Pima and Heart data sets include more noisy instances in compare to Wisconsin and Ionosphere data sets. This table also shows our proposed CLCF model outperformed and increased the overall performance. This research has used three medical data sets, which were Pima, Wisconsin, Heart (statlog). The consistency between the results provides strong support for the validity of the proposed algorithms to classify class noisy instances in medical areas. The results also in Ionosphere data set show its efficiency for decision-making systems in various domains.

5.1 Comparison of the results

As Table 7 shows, a comparison between the proposed technique and existing techniques indicates that the proposed technique leads to more accurate classification. The results show an improvement when compared with the existing approaches and the preceding comparative analysis. A close analysis of Tables 2, 3, 4, 5, 6 and 7 suggests that the proposed model is able to accurately detect class noisy instances and enhance the overall performance.

5.2 Significant test on the CLCF results

In this section, the statistical tests are used to examine the significance of the differences in the means of performance before and after noise detection. A *t* test has been performed on accuracy results to observe whether differences between two methods are statistically significant at level of $\propto = 0.05$. The *p* value of paired *t* test that compares CLCF model with before noise detection for each data set is shown in Table 8. It shows all the differences are statistically significant.

Table 6 Comparative analysis of the best result of the CLCF model and the results obtained before noise reduction on four data sets

Data set	Accuracy	
	Before noise reduction	After noise reduction
Pima	71.496 ± 0.062	94.817 ± 0.027
Wisconsin	95.014 ± 0.025	96.877 ± 0.024
Heart	75.185 ± 0.095	85.715 ± 0.060
Ionosphere	82.337 ± 0.058	85.439 ± 0.058

Table 7 Accuracy comparisons with existing techniques

Data set	Previous study		The proposed model	
	Ref.	Acc.	Acc.	Tech.
Pima	[5]	79.77	94.81	KM-SVM
	[12]	75.1		
	[13]	78.07		
	[13]	96.62		
	[13]	96.57		
Heart (Statlog)	[47]	84.07	85.71	KM-NN
	[13]	84.81		
	[12]	81.7		
Ionosphere	[12]	84.6	85.43	KM-NN
	[43]	82.8		

Ref = Reference, *Acc* = accuracy, *Tech* = technique

6 Conclusions

This paper investigates misclassified instances issues and proposes the CLCF model for class noise detection and classification using k-means algorithm and five different classification algorithms. It has been confirmed that clustering technique and classification filtering can lead to more reliable and accurate results for class noise detection and classification. The proposed model was applied on four binary data sets with low dimension and the performance was evaluated in terms of accuracy. It has been assumed that the data sets do not have missing values and the proposed model is not robust against outliers. It was shown that the proposed CLCF model was successful in identifying class noisy samples in comparison with the results obtained before noise detection. In addition, in order to improve data quality, two different techniques

Table 8 *p* value of paired *t* test that compares CLCF model with before noise detection model

Data sets	*p* value	Significant?
Pima	1.18E−13	Y
Wisconsin	2.08E−18	Y
Heart	5.17E−10	Y
Ionosphere	0.032237957	Y

for noise classification, namely, the removing and relabeling are applied. The main limitation of this model is the crisp nature of k-means algorithms for allocating cluster membership to data instances. Any data item, especially highly structured data, belongs to one of the existing clusters based on the minimum distance while this scenario does not work generally for real world data. There are three trends for future works. First, it would be useful to apply other clustering techniques to overcome the limitation of the k-means. Second, a new method for noise recognition can be developed. Finally, because the proposed method works with two-class data sets, another direction of this research is proposing a method for classifying multiclass data sets.

Acknowledgements The authors would like to thank those who have supported this research for their useful comments during its completion.

Compliance with ethical standards

Conflict of interest The author declare that they have no competing interests.

References

1. Zhu X, Wu X (2004) Class noise vs. attribute noise: a quantitative study of their impacts. Artif Intell Rev 223:177–210
2. Frénay B, Verleysen M (2014) Classification in the presence of label noise: a survey. IEEE Trans Neural Netw Learn Syst 255:845–869
3. Miranda AL, Garcia LPF, Carvalho AC, Lorena AC (2009) Use of classification algorithms in noise detection and elimination. In: International conference on hybrid artificial intelligence systems. Springer, pp 417–424
4. Sluban B, Lavrač N (2015) Relating ensemble diversity and performance: a study in class noise detection. Neurocomputing 1601:120–131
5. Lowongtrakool C, Hiransakolwong N (2012) Noise filtering in unsupervised clustering using computation intelligence. Int J Math Anal 659:2911–2920
6. Srimani PPK, Koti MS (2012) Outlier mining in medical databases by using statistical methods. Int J Eng Sci Technol 401:239–246
7. Catal C, Alan O, Balkan K (2011) Class noise detection based on software metrics and ROC curves. Inf Sci 18121:4867–4877
8. Sluban B, Gamberger D, Lavra N (2010) Advances in class noise detection. Front Artif Intell Appl 2151:1105–1106
9. Hodge V, Austin J (2004) A survey of outlier detection methodologies. Artif Intell Rev 222:85–126
10. Van Hulse JD, Khoshgoftaar TM, Huang H (2006) The pairwise attribute noise detection algorithm. Knowl Inf Syst 112:171–190
11. Xiong H, Pandey G, Member S (2006) Enhancing data analysis with noise removal. IEEE Trans Knowl Data Eng 183:304–319
12. Zeidat N, Wang S, Eick CF (2005) Dataset editing techniques: a comparative study. University of Houston, Houston
13. Smith MR, Martinez T, Giraud-Carrier C (2014) An instance level analysis of data complexity. Mach Learn 952:225–256
14. Thongkam J, Xu G, Zhang Y, Huang F (2008) Support vector machine for outlier detection in breast cancer survivability prediction. In: Advanced web and network technologies, and applications. Springer, pp 99–109
15. Jeatrakul P, Wong KW, Fung CC (2010) Data cleaning for classification using misclassification analysis. J Adv Comput Intell Intell Inform 143:297–302
16. Angelova A, Abu-Mostafa Y, Perona P (2005) Pruning training sets for learning of object categories. In: IEEE computer society conference on computer vision and pattern recognition, CVPR 2005, pp 494–501
17. Segata N, Blanzieri E, Delany SJ, Cunningham P (2010) Noise reduction for instance-based learning with a local maximal margin approach. J Intell Inf Syst 352:301–331
18. Segata N, Blanzieri E (2009) A scalable noise reduction technique for large case-based systems. In: International conference on case-based reasoning. Springer, Berlin, pp 328–342
19. Zeng X, Martinez T (2003) A noise filtering method using neural networks. In: IEEE international workshop on soft computing techniques in instrumentation, measurement and related applications, 2003, SCIMA 2003, pp 26–31
20. Sánchez JS, Barandela R, Marqués AI et al (2003) Analysis of new techniques to obtain quality training sets. Pattern Recogn Lett 247:1015–1022
21. Sabzevari M, Martínez-Muñoz G, Suárez A (2018) A two-stage ensemble method for the detection of class-label noise. Neurocomputing 275:2374–2383
22. Fränti P, Sieranoja S (2019) How much can k-means be improved by using better initialization and repeats? Pattern Recogn 93:95–112
23. He Z, Yu C (2019) Clustering stability-based evolutionary k-means. Soft Comput 231:305–321
24. Nematzadeh Z, Ibrahim R, Selamat A (2015) A method for class noise detection based on k-means and SVM algorithms. In: Intelligent software methodologies, tools and techniques. Springer, pp 308–318
25. Singh K, Malik D, Sharma N (2011) Evolving limitations in k-means algorithm in data mining and their removal. Int J Comput Eng Manag 121:105–109
26. Garcia LPF, Lorena AC, Carvalho ACPLF (2012) A study on class noise detection and elimination. In: 2012 Brazilian symposium on neural networks. Curitiba- PR. 20–25 Oct, pp 13–18
27. Farid DM, Harbi N, Rahman MZ (2010) Combining Naive Bayes and decision tree for adaptive intrusion detection. arXiv preprint arXiv:1005.4496
28. Meyer D (2004) Support vector machines: the interface to libsvm in package, p e1071
29. Li D-f, Hu W-c, Xiong W, Yang J-b (2008) Fuzzy relevance vector machine for learning from unbalanced data and noise. Pattern Recogn Lett 299:1175–1181
30. Wald R, Khoshgoftaar TM, Shanab AA (2014) The effect of noise level and distribution on classification of easy gene microarray data. In: Proceedings of the 2014 IEEE 15th international conference on information reuse and integration, pp 297–302
31. Dehariya S, Singh D (2013) An ensemble method based on particle of swarm for the reduction of noise, outlier and core point. Int J Adv Comput Res 31:1–5
32. Depeursinge A, Iavindrasana J, Hidki A et al (2010) Comparative performance analysis of state-of-the-art classification algorithms applied to lung tissue categorization. J Digit Imaging 231:18–30
33. Folleco A, Khoshgoftaar TM, Hulse JV, Bullard, L (2008) Software quality modeling: the impact of class noise on the random forest classifier. In: 2008 IEEE congress on evolutionary computation (IEEE world congress on computational intelligence). IEEE, pp 3853–3859
34. Van Hulse J, Khoshgoftaar T (2009) Knowledge discovery from imbalanced and noisy data. Data Knowl Eng 6812:1513–1542

35. Cover T, Hart P (1967) Nearest neighbor pattern classification. IEEE Trans Inf Theory 131:21–27
36. Daza L, Acuna E (2007) An algorithm for detecting noise on supervised classification. In: Proceedings of WCECS-07, the 1st world conference on engineering and computer science, pp 701–706
37. Pechenizkiy M, Tsymbal A, Puuronen S et al (2006) Class noise and supervised learning in medical domains: the effect of feature extraction. In: 19th IEEE symposium on computer-based medical systems (CBMS'06), pp 708–713
38. Lan M, Tan CL, Su J, Lu Y (2009) Supervised and traditional term weighting methods for automatic text categorization. IEEE Trans Pattern Anal Mach Intell 314:721–735
39. Li Y (2003) Classification in the presence of class noise. Pattern Recogn 5:1–30
40. Li R-L, Hu Y-F (2003) Noise reduction to text categorization based on density for KNN. In: Proceedings of the 2003 international conference on machine learning and cybernetics (IEEE Cat. No. 03EX693), vol 5. IEEE, pp 3119–3124
41. Frénay B, Verleysen M (2014) Classification in the presence of label noise: a survey. IEEE Trans Neural Netw Learn Syst 251:845–869
42. Folorunsho O (2013) Comparative study of different data mining techniques performance in knowledge discovery from medical database. Int J Adv Res Comput Sci Softw Eng 33:11–15
43. Kordos M, Rusiecki A (2013) Improving MLP neural network performance by noise reduction. In: International conference on theory and practice of natural computing. Springer, Berlin, pp 133–144
44. Webb AR (2003) Statistical pattern recognition. Wiley, New York
45. Juang L-H, Wu M-N (2010) MRI brain lesion image detection based on color-converted k-means clustering segmentation. Measurement 437:941–949
46. Frank A, Asuncion A (2011) UCI machine learning repository, 2010. http://archive.ics.uci.edu/ml
47. Smith MR, Martinez T (2013) An extensive evaluation of filtering misclassified instances in supervised classification tasks, vol 11, pp 1312–3970. arXiv preprint arXiv:1312.3970
48. Nematzadeh Z, Ibrahim R, Selamat A, Nazerian V (2020) The synergistic combination of fuzzy C-means and ensemble filtering for class noise detection. Eng Comput 377:2337–2355