Int. J. Nonlinear Anal. Appl. 12 (2021) No. 1, 701-712
ISSN: 2008-6822 (electronic)
http://www.ijnaa.semnan.ac.ir

Outlier Detection in Test Samples and Supervised Training Set Selection

Navid Mohseni[a], Hossein Nematzadeh[b,*], Ebrahim Akbari[b]

[a]Department of Computer Engineering, Babol Branch, Islamic Azad University, Babol, Iran.
[b]Department of Computer Engineering, Sari Branch, Islamic Azad University, Sari, Iran.

Abstract

Outlier detection is a technique for recognizing samples out of the main population within a data set. Outliers have negative impacts on classification. The recognized outliers are deleted to improve the classification power generally. This paper proposes a method for outlier detection in test samples besides a supervised training set selection. Training set selection is done based on the intersection of three well known similarity measures namely, jacquard, cosine, and dice. Each test sample is evaluated against the selected training set for possible outlier detection. The selected training set is used for a two-stage classification. The accuracy of classifiers are increased after outlier deletion. The majority voting function is used for further improvement of classifiers.

Keywords: Outlier detection, Training set selection, Similarity measures

1. Introduction

Training Set Selection (TSS) becomes important when the data set is large scale, so that the large number of instances is available in the data set [1]. These data could even constantly being produced. In such a case, data mining researchers tend to decrease the train set size by selecting the most representative data to construct the new reduced train set. The reduced train set is then used to construct a learning model with classifiers. It is expected that the resulted model could improve the accuracy of prediction and sometimes decrease the execution time [2, 3, 4, 5]. Generally, TSS also has wide range of applications in image processing including image set classification [6] and face recognition [7]. An outlier is an observation that has an abnormal distance from other values with same labels. The number of outliers in a data set is not usually much. However, recognizing and removing outliers could improve the accuracy of the classifiers [8]. There exist statistical methods

*Corresponding author
Email addresses: Mohseni.navid@yahoo.com (Navid Mohseni), hn_61@yahoo.com (Hossein Nematzadeh), Akbari@iausari.ac.ir (Ebrahim Akbari)

Received: 8 July 2020 *Accepted:* 14 December 2020

that could be used for outlier detection like Linear Regression Model (LRM) [9], Principle Component Analysis (PCA) [10]. Since there is no one best outlier detection method, many researches have been done in this field [8, 11, 12, 13]. This paper combines outlier detection in test samples while doing training set selection. Researches in outlier detection generally detect the outliers within the train set. However, it is plausible that an unseen or a test sample can be an outlier. Thus, this paper investigates the existence of an outlier within a test set rather than a train set. To our best of knowledge there is not a work which simultaneously detects outliers in test samples and selects the training set with regard to the test set. The rest of this section investigates the recent training set selection and outlier detection works separately.

Acampora et al [14] proposed a pareto-based multi-objective optimization approach based on accuracy and reduction rate. It is believed that even though Support Vector Machine (SVM) is widely used in data mining, but it suffers scalability which directly affects the complexity (execution time and memory). The proposed method was used in the preprocessing step. In consequence, the training set was reduced by selecting the most representative samples. Thus, by increasing the sample size, the execution time and the memory size is not increased drastically. The experimental results show that the proposed method improves the accuracy and reduction rate of Shell Extraction algorithm (SE) [15]. Likewise, Verbiest et al [5] further improved the SVM's accuracy using training set selection within a wrapper approach. The subsets of new training set were evaluated based on SVM training accuracy within a wrapper TSS. Five TSS methods have been proposed which were originally developed for K-Nearest Neighbor (KNN). The experimental results revealed that Generational Genetic Algorithm achieved the best performance among the five proposed methods. The paper also had two major findings. First, evolutionary based algorithm could increase the SVM accuracy. However, the execution time could be increased due to consecutive number of iterations. Second, wrapper TSS outperformed filter TSS on experimental data sets. Mohammed et al [4] proposed a three-step method for TSS using swarm intelligence. First, data sampling was done to reduce the size of input train set. Second, this reduced set was used to train a group of classifiers with bagging and distance based feature sampling. Third, swarm intelligence meta-heuristics were used (moth-flame optimization algorithm (MFO), the grey wolf optimizer (GWO) and the whale optimization algorithm (WOA)) to enhance the fusion process by assigning weights to the classifiers. The objective function was Matthews Correlation Coefficient (MCC). The proposed method fitted perfectly with both binary and multi-class data sets. WOA achieved the worse results in ensemble size for prediction accuracy. Even though, the weight assignment execution time can be controlled by tuning algorithmic parameter, the overall execution time is considerably high. Hence, the proposed method is applicable in offline learning. It can be concluded that meta-heuristic based TSS usually have high time complexity. The literature is investigated to recognize the recent work in outlier detection as follows.

Lejeune et al [11] proposed a shape based outlier detection in multivariate functional data using mapping functions from differential geometry. Basically, outlier detection is difficult within multivariate functional data because of individual behavior and dynamic correlation of parameters. Interpretable functional curve shape features were proposed and extracted from synthetic data sets. The results showed the efficiency of the proposed method against functional-depth-based-methods. Wang and Mao [13] proposed ensemble outlier detection model for an adaptive K-Nearest Neighbor instead of the traditional K-Nearest Neighbor algorithm. The adaptive K-Nearest Neighbor exploited support vector data description and investigated the area in which class probabilities are constant with regard to the test pattern. The solution ensemble model selected more competent individuals for each test pattern using posterior probabilities of classifiers and outperformed traditional ensemble methods such as majority voting. The experimental results were conducted on UCI data sets. Tang

and He [12] proposed a density based outlier detection. To do so the Relative Density Based Outlier Score (RDOS) was introduced as a measurement of local outlierness of objects. Kernel Density Estimation (KDE) was used to estimate the density distribution at the location of an object. Reverse and shared nearest neighbors were used besides the traditional K-Nearest Neighbor. Fourteen real life data sets (large scale and small scale data sets) were selected to evaluate the proposed methods including healthcare and non- healthcare data sets. The Area Under Curve (AUC) and Receiver Operator Characteristic (ROC) analysis shows the superiority of the proposed method in majority of data sets. Christy et al proposed [16] a cluster based outlier detection algorithm for healthcare data sets in which two outlier detection algorithms were introduced for outlier detection and removal based on an outlier score namely, Distance-Based outlier algorithm and Cluster-Based outlier algorithm. The similarity clustering was done to recognize the outliers in key subset attributes rather than the set of all attributes. The results on three data sets extracted from R package (Esoph, Diabetes, and KosteckiDillon) showed that Cluster-Based outlier algorithm outperformed Distance-Based outlier algorithm. In addition, the random removal of data objects does not affect on the values of the F-Score and Likelihood.

All in all, investigation of the literature does not reveal any work that detects outliers on test samples which is one of the main objectives of this research. Briefly the contributions of the paper are as follows:

- Outlier detection within the test set
- Training Set Selection (TSS)
- A two-stage supervised classification of the test sample

The rest of this paper is organized as follow. First, the fundamental concepts regarding the similarity measures are explained in Section 2. Next, the proposed method is introduced in Section 3. Then, the experimental results and further discussions are illustrated with figures and tables in Section 4. Finally, the conclusion remark is given in Section 5.

2. Preliminaries

The similarity measures within the proposed method are Jaccard, Cosine, Dice. Thus, the definitions and rlated expressions are given in this section. Assuming A and B are two vectors then the Jaccard similarity can be calculated as in Equation (2.1) [17].

$$Jaccard(A,B) = \frac{A \cdot B}{\|A\|^2 + \|B\|^2 - A \cdot B} \tag{2.1}$$

Likewise, Dice similarity measure is expressed in Equation (2.2) [18].

$$Dice(A,B) = \frac{2|A \cdot B|}{\|A\|^2 + \|B\|^2} \tag{2.2}$$

So that $A \cdot B$ is the dot product and $\|A\|^2$ is defined as Equation (2.2)

$$\|A\|^2 = \sum_{i=1}^{n} A_i^2, \qquad n = vector\ size \tag{2.3}$$

Finally, the Cosine similarity measure used in the proposed method is defined as follow in Equation (2.4) [19].

$$Cosine(A,B) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}, \qquad n = vector\ size \tag{2.4}$$

3. Proposed method

Figure 1 clearly shows the procedure of the proposed method in general. First, the data set is split into train and test set. Then, regarding each sample in the test set the similarity of the instance is calculated with respect to the train set using a predefined threshold (α) in (0.6, 0.7, 0.8, 0.9) to construct TSS_r in which TSS_1 refers to $TSS_{jaccard}$, TSS_2 refers to TSS_{Cosine}, and TSS_3 refers to TSS_{dice}. Obviously, the greater α is, the more similar the test sample is to the train set. Assuming TSS_r is less than 0.05 of the original train set, the related test sample is recognized as an outlier which means that it has a long distance to its neighbors otherwise the test sample is classified based on the respective TSS_r for each similarity measure (L_r^*). Finally, majority voting function is applied over (L_r^*) to calculate L^*. It is expected that both L_r^* and L^* improve the classification accuracy. The ratio of 0.05 is set up experimentally.

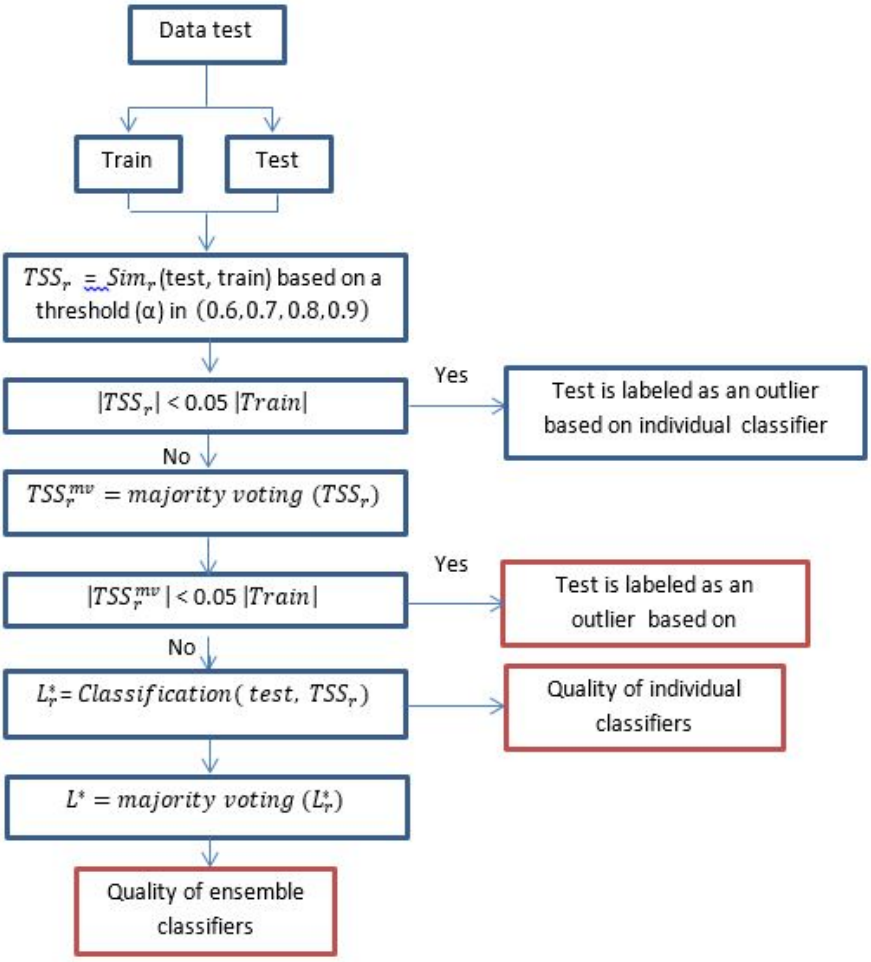

Figure 1: Illustration of the proposed method

The general view of the proposed method in Figure 1 can be formulated as follow. Assuming the data set in Table 1, each sample can be shown as in Equation (3.1). Likewise, each feature vector F_j can be shown as Equation (3.2)

Table 1: A sample data set

	F_1	F_2	F_3	$\cdots$	F_m	LABEL
x_1	f_1^1	f_1^2	f_1^3	$\cdots$	f_1^m	1
x_2	f_2^1	f_2^2	f_2^3	$\cdots$	f_2^m	2
x_3	f_3^1	f_3^2	f_3^3	$\cdots$	f_3^m	3
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\square \cdot \square$ $\cdots$ $\square \cdot \square$	$\vdots$	$\vdots$
x_n	f_n^1	f_n^2	f_n^3	$\cdots$	f_n^m	l

$$\bar{x}_i = (f_{i1}, f_{i2}, \cdots, f_{in}) \qquad i = 1, 2, \cdots, n \tag{3.1}$$

$$\bar{F}_j = (f_{1j}, f_{2j}, \cdots, f_{nj})^t \tag{3.2}$$

If $L = \{L_1, L_2, \cdots, L_k\}$ is a set of label, then the samples belonging to each label is indicated as follow in Equation (3.3).

$$y^{1p} = \{x_1^{lp}, x_2^{lp}, \cdots, x_n^{lp}\} = \left\{x_j^{lp}\right\}_{j=1}^{n} \tag{3.3}$$

where

$$\cup_{p=1}^{k} y_0^{lp} = \sum_{p=1}^{k} n_p = n, \qquad p = 1, 2, \cdots, k$$

Assuming $train = (x_1, x_2, \cdots, x_n)$ and a test sample is denoted as $T = (f_1, f_2, \cdots, f_n)$, then based on a similarity measure δ^r (jaccard, cosine, and dice) the similarity between test T and each x_i is calculated as in Equation (3.4).

$$w_i^r = \delta^r(T_i, x_i), \quad r = 1, 2, \cdots, h, \;\; i = 1, 2, \cdots, n \tag{3.4}$$

In which h is the number of similarity measure based on w_i^r. Each member of train set is selected as TSS^r if $w_i^r \geq \alpha$ as shown in Equation (3.5). α is a predefined threshold.

$$TSS_r = \{z_1^r, z_2^r, \cdots, z_{sr}^r\}, \qquad z_r \leq n, \;\; r = 1, 2, \cdots, n \tag{3.5}$$

Based on each similarity measure r, if $TSS_r = \emptyset$ or the size of TSS_r is sufficiently small with respect to the size of the original train set ($0.05 \times |Train|$ as stated in Figure 1), then that test sample T_i is recognized as an outlier. Algorithm. 1 shows how a test sample (T) can be predicted generally using a classifier and the selected train set using a similarity measure δ^r. The proposed method utilizes three similarity measures, thus $h = 3$ and the threshold (α) is selected from $(0.6, 0.7, 0.9)$ as stated in Figure 1. In addition the value of β is 0.05 within the proposed method. Using Algorithm. 1 the need for classifier can be waived if all members of TSS^r has one common label (L_p). In that case the predicted label for the test sample T is also L_p otherwise, the classifier is needed to determine the label of the test sample T. In other words, classification of the test sample is done in two stages. First, TSS is examined to evaluate the chance of classification without using the classifier. Second, if using a classifier is inevitable, a base classifier is used.

If the majority of measures identify a test sample as an outlier, that sample should be recognized as an outlier accordingly. This can be achieved using the majority voting of three measures. Algorithm 2 clearly shows both application of majority voting for outlier detection and classifiers' prediction L^* within Figure 1.

Algorithm 1. Outlier detection and classification of a test sample for each similarity measure

Input:
1. $Train = (x_1, x_2, \cdots, x_n)$
2. $T = (f_1, f_2, \cdots, f_m)$
3. $(\delta^1, \delta^2, \cdots, \delta^h)$h is the maximum number of similarity measure
4. $threshould$ (α)

Output:label of test T as an outlier or label class L_p

For $r \leftarrow 1$ **to** h
  FOR $i \leftarrow 1$ **to** n
    $w_i^r \leftarrow \delta^r(T, x_i)$
    IF $w_i^r > \alpha$ **THEN**
      $TSS_r \leftarrow x_i$
    END IF
  END FOR
  $TSS^r \leftarrow \{x_1^r, x_2^r, \cdots, x_{sr}^r\}$
  IF $TSS_r = \emptyset$ **or** $|TSS_r| < \beta \times |train|$ **THEN**
    T is labeled as an outlier based on each similarity of r
  END IF
  IFall members of TSS_r are related to one label (L_p) **THEN**
    $Label(T) \leftarrow L_p$
  ELSE
    $L_p \leftarrow Classifier(T, TSS_r)$
    test T is labeled via TSS_r using a Classifier
  END IF
END FOR

Algorithm 2. Outlier detection and classification of a test sample using majority voting

Input:$L = (L^1, L^2, \cdots, L^h), test(T)$

Output:L^*

$S \leftarrow sum(L == 0)$
IF $S \geq h/2$ **THEN**
  $L^* \leftarrow 0(outlier)$
ELSE
  $S_p \leftarrow sum(L == L_p)$ $(p = 1, 2, \cdots, h)$
END IF
IF$S_p \geq= h/2$**THEN**
  $L^* \leftarrow L_p$
END IF

Even though it is expected that by the end of Algorithm.1 the classification accuracy improves, majority function is further applied to ensemble all similarity measures in order to increase the measurement criteria.

4. Experimental results

The proposed method has been tested on twelve data sets including binary and multi-labeled data sets in Table 2. Three data sets from twelve data sets of Table 2 are well-known artificial data sets illustrated in Fig 2 in which each label is distinguished with a separate color. The calculated results are based on 5-fold cross validation. Support Vector Machine (SVM) classifier with linear kernel is used for binary data sets. In addition, Random Forests (RF) is used for calculation of measurement criteria of multi-labeled data sets. The proposed method is implemented on a platform with 500 Gigabyte SSD Hard Disk, 12 Gigabyte RAM and a Core i5 CPU. Precision, recall, fscore and binary accuracy are measurement criteria for binary data sets [20]. Likewise, Normalized Mutual Information (NMI) and multi-labeled accuracy [21] are measurement criteria for multi-labeled data sets.

Table 2: Data sets

Data sets	Instance	Feature	Labels
Bcwisconsin	569	30	2
Wisconsin	194	33	2
Australian	690	14	2
German	1000	24	2
Heart	270	13	2
Yeast	1484	8	10
CTG	2126	21	10
Ecoli	336	7	8
Pv	210	18	7
Pathbas	300	2	3
R15	600	2	15
Spr	312	2	3

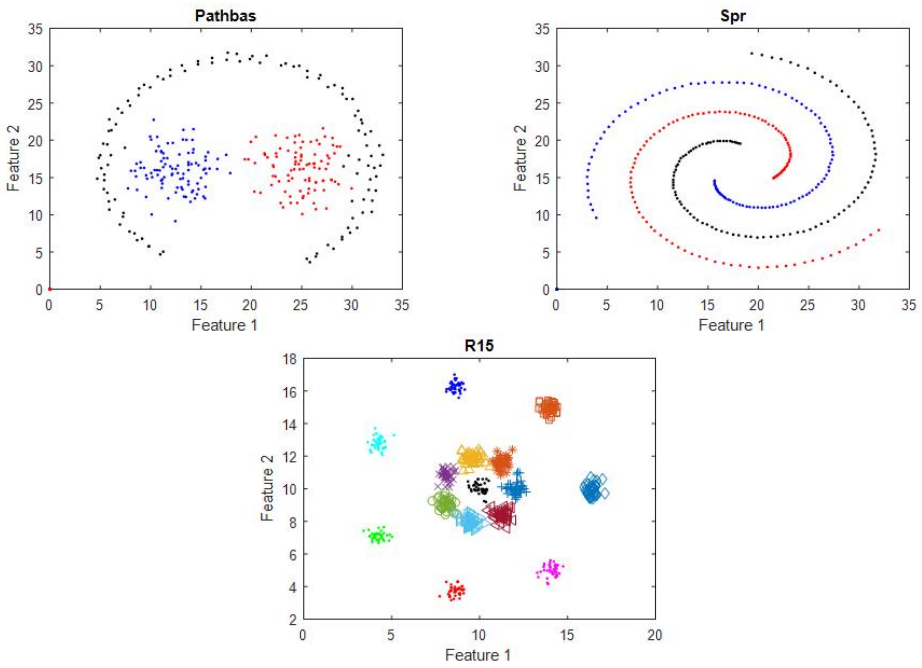

Figure 2: The two dimensional view of the artificial data sets in Table 1

Table 3 and 4 show the measurement criteria for binary and multi-labeled data sets before applying the proposed method. Table 5 shows the results after applying the proposed method for binary data sets so that the best α is specified for each data set. The respective measurements have been calculated based on the related best α. Table 5 shows that even though the accuracy of Jaccard, Cosine, and Dice have been increased with the proposed method, the majority voting function even improved this increment. Thus, it can be concluded that application of majority voting is useful and can improve the measurement criteria as indicated in Table 5. In addition, the number of outliers detected in the test set is also specified in the last column of Table 5. The proposed method did not recognize any test outliers in Wisconsin and German. Likewise, majority voting has increased or at least preserved the accuracy of individual measures. Thus, majority voting has been applied on calculation of accuracy and NMI for multi-labeled data sets in Table 6 as well. No outliers have been detected in Pathbad, R15, and Spr. The scattering illustrations of points in three artificial data sets of Fig 2 approve the number of detected outliers as well. The execution time did not differ before and after applying the proposed method which is an interesting aspect of the proposed method.

Table 3: Measurement criteria before applying the proposed method on binary data sets using SVM

Data sets	Accuracy	Precision	Recall	Fscore
Bcwisconsin	0.95	0.97	0.95	0.96
Wisconsin	0.74	0.68	0.76	0.69
Australian	0.77	0.88	0.80	0.84
German	0.76	0.90	0.81	0.85
Heart	0.83	0.86	0.84	0.85

Table 4: Measurement criteria before applying the proposed method on multi-labeled data sets using RF

Data sets	Accuracy	NMI
Yeast	0.61	0.40
CTG	0.91	0.82
Ecoli	0.83	0.72
Pv	0.91	0.89
pathbas	0.97	0.92
R15	0.98	0.98
Spr	0.98	0.96

Table 5: Measurement criteria after applying the proposed method on binary data sets using SVM

Data sets	α	Jaccard Acc	Cosine Acc	Dice Acc	MV Acc	MV Precision	MV Recall	MV Fscore	Outliers detected
Bcwisconsin	0.9	0.95	0.96	0.96	0.96	0.98	0.96	0.97	6
Wisconsin	0.8	0.78	0.76	0.78	0.79	0.93	0.82	0.87	0
Australian	0.6	0.84	0.80	0.84	0.84	0.80	0.83	0.81	10
German	0.8	0.77	0.78	0.78	0.78	0.89	0.82	0.85	0
Heart	0.8	0.83	0.83	0.83	0.84	0.89	0.84	0.87	8

Table 6: Measurement criteria after applying the proposed method on multi-labeled data sets using RF

Data sets	α	Jaccard Acc	Cosine Acc	Dice Acc	MV Acc	MV NMI	Outliers detected
Yeast	0.9	0.62	0.62	0.62	0.63	0.41	6
CTG	0.9	0.91	0.91	0.91	0.92	0.83	4
Ecoli	0.9	0.85	0.84	0.84	0.86	0.75	8
Pv	0.9	0.89	0.90	0.91	0.93	0.91	14
Pathbas	0.9	0.98	0.98	0.97	0.99	0.95	0
R15	0.9	0.99	0.98	0.99	0.99	0.98	0
Spr	0.9	0.99	0.99	0.99	0.99	0.98	0

4.1. Discussion

For further discussion the relation between the number of detected outliers and α is evaluated in Figure 3. As the value of α increases from 0.6 to 0.9 the number of detected outliers within the test set is increased or at least remains unchanged in all data sets. This is rational because by increasing α, the proposed method becomes more rigorous in constructing the TSS. In other words, the greater α is, the smaller the TSS is. Accordingly, the smaller the TSS is, the greater the chance of the detecting the outlier is. In addition, the experimental investigation of outlier detection does not reveal any outliers as stated in Table 6. However, outliers are manually added to the aforementioned data sets to check either the proposed method could detect the outlier within the test sample or not. The proposed method succeeded to detect the outliers in three artificial data sets which are highlighted in red as shown in Figure 4. The test set is also marked with a circle around and the remaining samples are the train set.

5. Conclusion

This paper investigated the outlier detection within test samples as well as training set selection (TSS).The paper also addressed how using a classifier can be waived if all TSS members have a same label. Twelve bench mark data sets have been used including binary and multi-labeled data sets. The results showed an increase in classifiers' prediction accuracy. Majority function was applied for further improvement of classification. The proposed method in this paper follows a supervised approach. Training set selection as an unsupervised approach is one of the main direction of the future work.

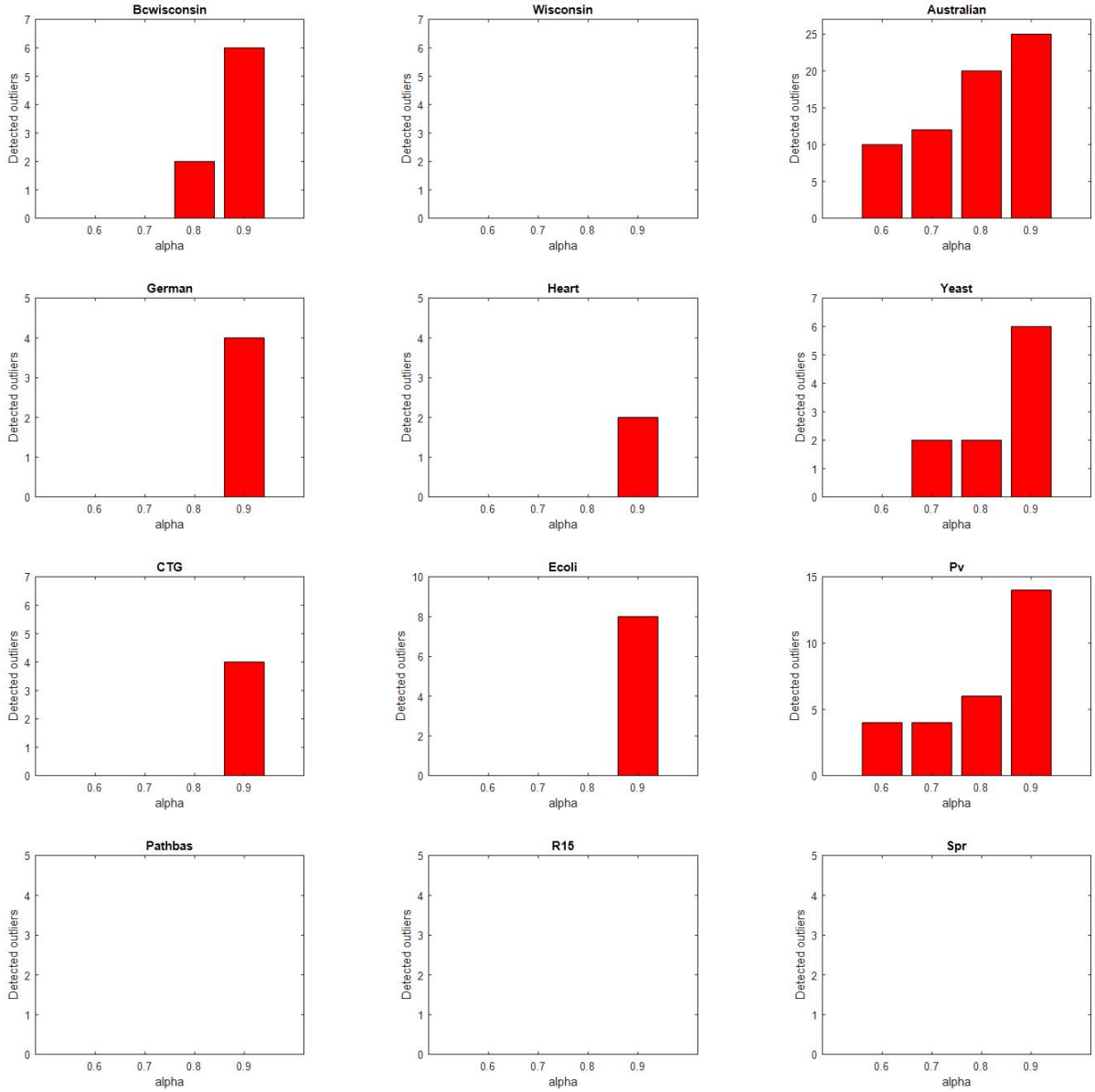

Figure 3: The effect of increasing alpha on detected outliers

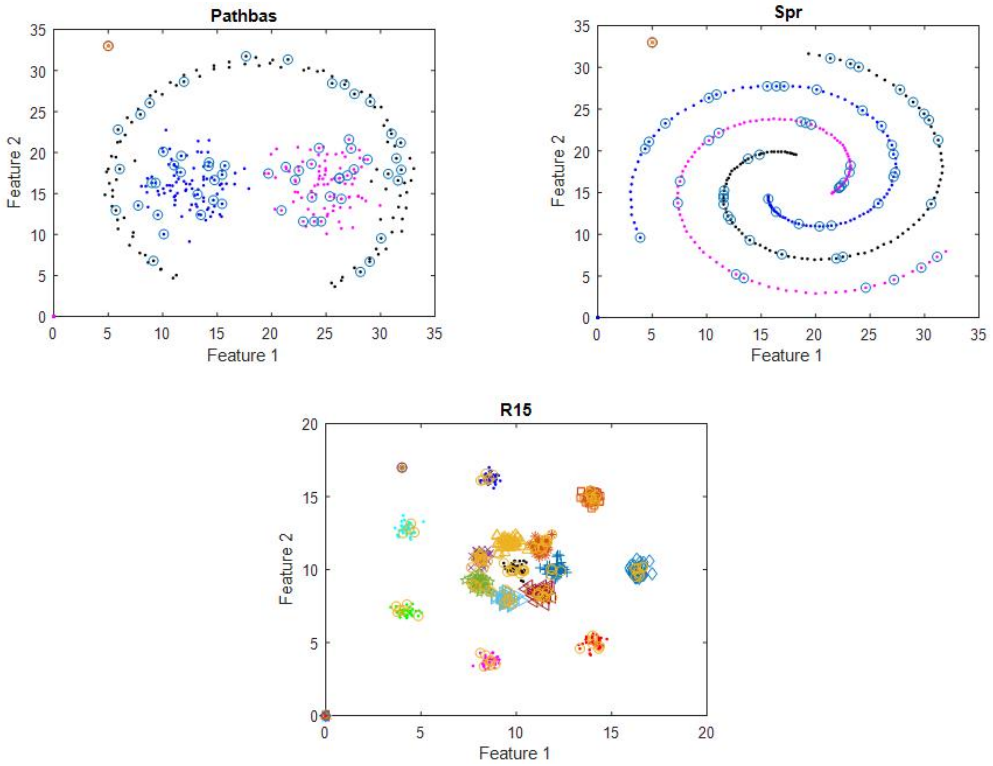

Figure 4: The illustration of outliers and test set and train set

References

[1] García-Pedrajas, N., Evolutionary computation for training set selection. WIREs Data Mining and Knowledge Discovery, 2011. 1(6): p. 512-523.
[2] Peng, S., et al., Optimal feasible step-size based working set selection for large scale SVMs training. Neurocomputing, 2020. 407: p. 366-375.
[3] Cano, J.R. and S. García, Training set selection for monotonic ordinal classification. Data & Knowledge Engineering, 2017. 112: p. 94-105.
[4] Mohammed, A.M., E. Onieva, and M. Woźniak, Training set selection and swarm intelligence for enhanced integration in multiple classifier systems. Applied Soft Computing, 2020. 95: p. 106568.
[5] Verbiest, N., et al., Evolutionary wrapper approaches for training set selection as preprocessing mechanism for support vector machines: Experimental evaluation and support vector analysis. Applied Soft Computing, 2016. 38: p. 10-22.
[6] Ren, Z., et al., Image set classification using candidate sets selection and improved reverse training. Neurocomputing, 2019. 341: p. 60-69.
[7] Santiago-Ramirez, E., et al., Optimization-based methodology for training set selection to synthesize composite correlation filters for face recognition. Signal Processing: Image Communication, 2016. 43: p. 54-67.
[8] Smiti, A., A critical overview of outlier detection methods. Computer Science Review, 2020. 38: p. 100306.
[9] Rath, S., A. Tripathy, and A.R. Tripathy, Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 2020. 14(5): p. 1467-1474.
[10] Chen, T., E. Martin, and G. Montague, Robust probabilistic PCA with missing data and contribution analysis for outlier detection. Computational Statistics & Data Analysis, 2009. 53(10): p. 3706-3716.
[11] Lejeune, C., et al., Shape-based outlier detection in multivariate functional data. Knowledge-Based Systems, 2020. 198: p. 105960.
[12] Tang, B. and H. He, A local density-based approach for outlier detection. Neurocomputing, 2017. 241: p. 171-180.
[13] Wang, B. and Z. Mao, A dynamic ensemble outlier detection model based on an adaptive k-nearest neighbor rule. Information Fusion, 2020. 63: p. 30-40.

[14] Acampora, G., et al., A multi-objective evolutionary approach to training set selection for support vector machine. Knowledge-Based Systems, 2018. 147: p. 94-108.
[15] Liu, C., et al., An efficient instance selection algorithm to reconstruct training set for support vector machine. Knowledge-Based Systems, 2017. 116: p. 58-73.
[16] Christy, A., G.M. Gandhi, and S. Vaithyasubramanian, Cluster Based Outlier Detection Algorithm for Healthcare Data. Procedia Computer Science, 2015. 50: p. 209-215.
[17] Lu, M., et al., Scalable news recommendation using multi-dimensional similarity and Jaccard–Kmeans clustering. Journal of Systems and Software, 2014. 95: p. 242-251.
[18] Singh, A. and S. Kumar, A novel dice similarity measure for IFSs and its applications in pattern and face recognition. Expert Systems with Applications, 2020. 149: p. 113245.
[19] Ye, J., Improved cosine similarity measures of simplified neutrosophic sets for medical diagnoses. Artificial Intelligence in Medicine, 2015. 63(3): p. 171-179.
[20] Nematzadeh, H., et al., Frequency based feature selection method using whale algorithm. Genomics, 2019. 111(6): p. 1946-1955.
[21] Akbari, E., et al., Hierarchical cluster ensemble selection. Engineering Applications of Artificial Intelligence, 2015. 39: p. 146-156.