Proceedings of Machine Learning Research 71:43–55, 2017 KDD 2017: Workshop on Anomaly Detection in Finance

Ensemble-Based Anomaly Detection Using Cooperative Learning

R. F. Kashef RKASHEF@IVEY.CA
IVEY Business School, University of Western Ontario

Abstract

Using the same process and functionality to solve both clustering and outlier discovery is highly desired. Such integration will be of great benefit to discover outliers in data and consequently obtain better clustering results after eliminating the set of outliers. It is known that the capability of discovering outliers using clustering-based techniques is mainly based on the quality of the adopted clustering. In this paper, a novel Cooperative Clustering Outlier Detection (CCOD) algorithm is presented. It involves multiple clustering techniques; the goal of the cooperative approach is to discover those outliers that are not detected by the single clustering-based outlier detection approaches using the methodology of cooperation. Undertaken experimental results show that the detection accuracy of the cooperative technique is better than that of the typical clustering-based FindCBLOF method over a number of artificial, gene expression and text document datasets.

Keywords: Ensemble-based Anomaly Detection, Cooperative Clustering, Internal Quality Measures

1. Introduction

Outlier detection refers to the problem of discovering objects that do not conform to expected behavior in a given dataset. These nonconforming objects are called outliers while the remaining objects are called inliers. A variety of techniques have been developed to detect outliers in several research applications including bioinformatics, data mining (Gunawardana et al., 2015; Chandola et al., 2012). More recently, the applications of anomaly detection methods have seen a proliferation in business intelligence where industries such as healthcare estimate fraudulent cases, abuse, and waste (van Capelleveen et al., 2016).. In addition, social media has dawned a new age of available information where geolocated data per Instagram has put methods of outlier detection in practice for the early detection of unusual events in urban areas (Dominguez et al., 2017).

Current approaches for detecting outliers using clustering techniques explore the relation of an outlier to the clusters in data. The clustering-based outlier detection technique, FindCBLOF (He et al., 2003) assumes that outliers form very small-sized clusters, and the detection accuracy of the FindCBLOF is mainly based on the quality of the adopted clustering technique. Factor-analysis can be used to determine and identify outliers from an orthogonal factor model, which can ultimately provide a natural way of clustering the anomalies with similar "abnormalities" into the same cluster (Xu and Wunsch, 2005).

In this paper, an Ensemble-Based outlier detection method namely Cooperative Clustering Outlier Detection (CCOD), is proposed and analyzed. Unlike current clustering-based

methods, such as FindCBLOF, the CCOD provides efficient outlier detection and data clustering capabilities. It uses the notion of cooperative learning (Kashef and Kamel, 2010) towards a better discovery of outliers. The algorithm of our outlier detection method is divided into four phases. The first phase provides individual clusterings. The second phase obtains a set of sub-clusters. The main objective of the third and four phases is to iteratively identify the set of possible and candidate outliers of objects. Empirical results obtained using gene expression and text document datasets indicate that the proposed method is successful in detecting more outliers compared to the FindCBLOF technique.

The rest of this paper is organized as follows: Different approaches for discovering outliers in data are presented in section 2. The CCOD algorithm is presented and analyzed in section 3. Experimental results are discussed in section 4. Finally, we draw some conclusions and outline future work in section 5.

2. Related Work and Background

The outlier detection problem can be defined as follows: Given a set of n objects, find the expected number of top objects ($TopRatio$) that are considerably dissimilar, exceptional, or inconsistent with respect to the remaining data (Papadimitriou et al., 2003). The subsequent sections illustrate a few of the more recent and well-known outlier detection approaches used in data mining practices.

2.1. Distance-based Outlier Detection

In distance-based outlier detection, an object x in a dataset is an outlier with respect to the parameters $MinPts$ and r, if no more than $MinPts$ objects in the dataset are at a distance r or less from x. This approach does not require any prior knowledge of the data distributions as the statistical methods do (Cao et al., 2014).

2.2. Distribution-based Outlier Detection

In these techniques, the data points are modeled using a stochastic distribution and points are determined to be outliers depending on their relationship with this model. Thus, an object is defined as an outlier if it is significantly different from the underlying distribution. A Gaussian mixture model is used in Yamanishi et al. (2004) to represent the normal behaviors and each datum is given a score based on changes in the model. A high score indicates a high possibility of being an outlier (Aggarwal, 2013).

2.3. Density-based Outlier Detection

Density-based methods have been developed for finding outliers in spatial data. Outliers are objects having a low local density of an object's neighborhood of objects. Density-based methods can be grouped into two categories namely multidimensional metric space-based methods and graph-based spatial outlier detection methods where the definition of spatial neighborhood is based on Euclidean distance and graph connectivity respectively (Aggarwal, 2013). LOF (Breunig et al., 2000) is intuitively a measure of the difference in density between an object and its neighborhood objects. In a multidimensional dataset, it is more meaningful to assign for each object a degree of being an outlier. Local outliers are the set of

objects, which relative to their local neighborhoods have low densities of the neighborhoods. Let *MinPts* specifies the minimum number of objects in the neighborhood of a specific point. The *MinPts*-distance neighborhood of x contains every object whose distance from x is not greater than the *MinPts*-distance. These objects are called the *MinPts*-nearest neighbors of x, $N_{MinPts}(x)$. The reachability distance of object x with respect to object y is defined as $reach\_dist_{MinPts}(x,\ y) = max\{MinPts\text{-distance}(x),\ \text{distance}(x,\ y)\}$. If object x is far away from y, then the reachability distance between the two is simply their actual distance. However, if they are close, the actual distance is replaced by the *MinPts*-distance of x. The local reachability density of an object x, $lrd(x)$, is the inverse of the average reachability distance from the *MinPts*-nearest neighbors of x.

$$\mathrm{lr}d_{\mathrm{MinPts}}\left(x\right) = \left(\frac{\sum_{y \in N_{\mathrm{MinPts}}(x)} \mathrm{reach\_dist}_{\mathrm{MinPts}}(x,y)}{|N_{\mathrm{MinPts}}\left(x\right)|}\right)^{-1} \tag{1}$$

The LOF is a measure of outlying-ness that is calculated for each object. LOF is the average of the ratios of the local reachability density of x and those of x's MinPts nearest neighbors. The local outlier factor of an object x is defined as:

$$\mathrm{LOF}_{\mathrm{MinPts}}\left(x\right) = \left(\frac{\sum_{y \in N_{\mathrm{MinPts}}(x)} \frac{\mathrm{lr}d_{\mathrm{MinPts}}(y)}{\mathrm{lr}d_{\mathrm{MinPts}}(x)}}{|N_{\mathrm{MinPts}}\left(x\right)|}\right) \tag{2}$$

Local outliers are objects having considerable density difference from their neighboring objects, i.e. they have high LOF values.

2.4. Deviation-based Outlier Detection

Deviation-based outlier detection does not use statistical tests or distance-based measures to identify exceptional objects. Instead, it identifies outliers in a group of objects as those objects that do not fit to the general characteristics of that group, i.e., the variance of the group is minimized when removing the outliers. The sequential exception technique simulates the way in which humans can distinguish unusual objects from among a series of supposedly similar objects (Arning et al., 1996).

2.5. Clustering-based Outlier Detection

In clustering-based outlier detection, Inliers are defined as objects that belong to large and dense clusters, while outliers either do not belong to any cluster or form very small clusters. The main concern of clustering-based outlier detection algorithms is to find clusters and outliers, which are often regarded as noise that should be removed in order to make more reliable clustering (He et al., 2003; Jiang et al., 2001). Some clustering algorithms find outliers as a by-product of clustering algorithms as DBSCAN (Ester et al., 1996) and ROCK Guha et al. (2000). In the Find Cluster-Based Local Outlier Factor (FindCBLOF) algorithm , outliers are returned as objects with higher CBLOF values where the CBLOF is a measure of both the size of the cluster the object belongs to and the distance between the object and its closest cluster (if the object lies in a small cluster) (He et al., 2002). The efficiency of the FindCBLOF approach for detecting outliers in data is constrained

to the quality of the adopted clustering technique. In Guha et al. (2000) and Kashef and Kamel (2008), it has been experimentally proven that better clustering solutions reveal better detection of outliers using the notion of CBLOF.

3. Ensemble-Based Outlier Detection using Cooperative Clustering (CCOD)

In this section, we introduce a new ensemble-based outlier detection algorithm called Cooperative Clustering Outlier Detection (CCOD). It uses the notion of cooperative clustering (Kashef and Kamel, 2010) towards better discovery of outliers. The proposed algorithm takes the dataset X of d dimensional n vectors and c clustering algorithms (A_1, A_2, ..., A_c). The purpose of our method is not only to perform clustering but at the same time to discover outliers based on cooperation between various clustering algorithms. The CCOD algorithm detects outliers recursively from level 2 (k=2) until level k (i.e., desired number of clusters), we recall this procedure as a bottom-up scenario which takes place in four phases. The first phase obtains the intersection (or agreement) between the adopted clustering techniques (i.e., sub-clusters), the second phase represents each sub-cluster with a histogram representation of the pair-wise similarities between objects in the same sub-cluster, the third phase identifies a possible set of outliers by assigning a cooperative outlier factor to each object in each sub-cluster, and finally the last phase returns the overall set of candidate outliers that affects the homogeneity of the merging process. The following sections describe the various phases of the proposed algorithm.

3.1. Generation of Sub-Clusters and Similarity-Histogram

A new set of disjoint sub-clusters Sb is generated as a form of the co-occurrence of objects among the multiple c clusterings. Thus, the underlying model indicates the agreement (or intersection) among the various clustering techniques for clustering the data into a set of k clusters. Each sub-cluster is represented by a concise statistical representation called Similarity Histogram (SH) of the pair-wise similarities distribution in a collection of objects (used by Kashef and Kamel (2010)).

3.2. Identification of Possible Outliers using Cooperative Outlier Factors

Our outlier detection method takes into account the following facts on outliers: (1) Outliers either do not belong to any cluster or form very small clusters, (2) Outliers may exist in large clusters, and (3) Outliers may affect on the homogeneity of the clustering results. Let Sb denotes the set of generated sub-clusters arranged in an increasing order of their sizes. Assume |Sbi| refers to the size of the sub-cluster Sbi . Given a dataset X of n objects , the boundary of large and small sub-cluster, v, is chosen such that the number of objects in large sub-clusters exceeds a fraction α_{sb} of the dataset X:

$$(|Sb_0| + |Sb_1| + \ldots + |Sb_{\,v}|) \geq (|X| * \alpha_{Sb}) \tag{3}$$

Equations 4 and 5 distinguish between strong and weak agreements (i.e., large and small sub-clusters).

$$\text{Large Subcluster Set (LSS)} = \{Sb_0,\ Sb_1, \ldots, Sb_v\} \tag{4}$$

$$\text{Small Subcluster Set (SSS)} = \{Sb_{v+1},\ Sb_{v+2},\ \ldots, Sb_{n_{\text{Sb}}-1}\ \} \tag{5}$$

Two types of outliers called Intra-Outlier and the Inter-Outlier are defined as follows: Intra-Outlier is an objects x having large distances from objects in the same sub-cluster Sb_i, if $\forall$ y $\in Sb_i$, x $\neq$y, |Sim(x,y)$< \delta$| is maximum, where δ is a similarity threshold and |Sim(x,y)$< \delta$| is the number of pair-wise similarities that are lower than δ. Intra-Outliers can be found in both large and small sub-clusters. Inter-Outliers are objects of a small sub-cluster that have large distances from objects in large sub-clusters. Each object is assigned an outlier factor called Cooperative Outlier Factor (COF) (Eq.6). For objects in small sub-clusters, the COF combines the weights of being Intra-Outlier and Inter-Outlier. The distances between the InterOutliers and large sub-clusters are calculated as the distance to their centroids. The cosine similarity measure is used to determine the similarities between objects.

$$\text{COF}\,(x) = \begin{cases} (1 - Sim\,(x, c_j)) + \frac{|\text{Sim}(x,y)<\delta|}{|Sb_i|-1},\ \forall y \in Sb_i,\ Sb_i \in SSS \\ \frac{|\text{Sim}(x,y)<\delta|}{|Sb_i|-1},\ \forall y \in Sb_i,\ Sb_i \in LSS \end{cases} \tag{6}$$

Objects with high values of the COF are considered as local outliers within the set of sub-clusters. The discovered set of local outliers provides the possible set of outliers PO_k, for the whole set of k clusters. The key difference between CCOD and FindCBLOF is that, in the former, the COF is assigned to objects within the set of sub-clusters, which composes an additional confidence of being an outlier, where the set of sub-clusters acts as an agreement between the multiple clusterings. In addition, the COF takes into account that outliers may exist in both large and small sub-clusters, where it detects both Intra and Inter outliers in the set of sub-clusters.

3.3. Discovery of Candidate Outliers through Merging of Sub-clusters

In order to obtain k desired clusters from the set of n_sb ($\text{n}_\text{sb} \geq$k) sub-clusters, the two most similar sub-clusters are merged first into a new cluster. The quality of merging is calculated by the coherency of the new histogram H_ij. In Eq.7, the histogram of the newly generated cluster is constructed by adding the corresponding counts of each bin from the two merged histograms, and also by adding the additional pair-wise similarities (i.e., | Sim(x, y)|) that are obtained during merging of the two sub-clusters that were not calculated in each individual histogram. Such that the following condition is satisfied: {((bin - (NumBins/2))*BinSize) < Sim(x,y) $\leq$ ((bin - (NumBins/2))*BinSize+BinSize)}.

$$H_{\text{ij}}\,(\text{bin}) = H_i\,(\text{bin}) + H_j\,(\text{bin}) + |\text{Sim}\,(x, y)|\,, \forall x \in Sb_i,\ y \in Sb_j,\ bin = 0, 1, \ldots, NumBins - 1 \tag{7}$$

Let $|\text{Sb}_\text{i}|$ and $|\text{Sb}_\text{j}|$ be the cardinality of sub-clusters Sb_i and Sb_i, respectively. Then, the number of pair-wise similarities, n_sim, in the newly generated histogram is calculated as: $n_{sim}(Sb_i,Sb_j) = (|\text{Sb}_\text{i}|+|\text{Sb}_\text{j}|)*(|\text{Sb}_\text{i}|+|\text{Sb}_\text{j}|\text{-}1)/2$. The merging cohesiveness factor (mcf) between any two sub-clusters is computed by calculating the ratio of the count of similarities

weighted by the bin similarity above a certain similarity threshold δ to the total count of similarities in the new merged histogram. The *mcf* ($\mathrm{Sb_i}$, $\mathrm{Sb_j}$) is calculated as:

$$\mathrm{mcf}\left(Sb_i, Sb_j\right) = \frac{\left(\sum_{bin=binThreshold}^{numBins=1}\left((bin * binSize) - 1 + \left(\frac{\mathrm{binSize}}{2}\right)\right) * H_{\mathrm{ij}}\left(\mathrm{bin}\right)\right)}{n_{\mathrm{Sim}}\left(Sb_i, Sb_j\right)} \tag{8}$$

where *binThreshold* is the bin corresponding to the similarity threshold δ. The higher the *mcf*, the more coherent is the newly generated cluster. The set of possible outliers is tested against the merging process in order to identify the candidate outliers for the set of k clusters. For each object o in the set $\mathrm{PO_k}$, if removing o results in a selection of two other sub-clusters with better homogeneity (i.e., higher value of the *mcf*), then o becomes a candidate outlier. Finally, the selected two most similar sub-clusters are then merged. This step is repeated until the number of clusters equals k. The CCOD algorithm detects outliers in an iteratively bottom-up fashion which starts form level l =2 (i.e., number of partitions l= 2) to level l=k (i.e., number of partitions l= k). It detects a set of outliers at level l that are considered as candidate outliers at level l and possible outliers at level l+1. The resulting candidate outliers are sorted according to their COF. The *TopRatio* candidate outliers are finally obtained, and the set of l cooperative clusters are returned.

3.4. Complexity Analysis

Assume $T^{A_1}(l)$, $T^{A_2}(l), \ldots T^{A_c}(l)$ is the computational time complexity of the clustering techniques $\mathrm{A_1}$,$\mathrm{A_2}$,..,$\mathrm{A_c}$, respectively, for a given number of clusters l=2,3,..,k. The time complexity of the FindCBLOF($\mathrm{A_i}$) algorithm is $T^{A_i}(l)$+O(n) where n is the total number of objects. Starting from Phase 1, the c clustering approaches are employed in the CCOD algorithm, thus Phase 1 takes the computational time complexity of the clustering approach with the maximum processing time , (i.e., Phase 1 is of order $\mathrm{O}(\max(T^{A_1}(l), T^{A_2}(l), \ldots T^{A_c}(l)))$. Finding the set of sub-clusters {$\mathrm{Sb_i}$, i=0,1,2,,..nsb- 1} is of order O(n), and building histograms is of order $\mathrm{O}(|\mathrm{Sb_i}|^2)$ thus the complexity of Phase 2 is $\mathrm{O}(\mathrm{n}+|\mathrm{Sbi}|^2)$. In Phase 3, the complexity of assigning the COF to objects is $\mathrm{O}(|\mathrm{Sbi}|^2)$. Finally, in Phase 4, finding the most homogenous sub-clusters to be merged is of order $\mathrm{O}(\mathrm{n_{sb}}^2)$, this merging step is repeated for each possible outlier. Thus Phase 4 is of order $\mathrm{O}(\mathrm{LocalTopRatio}*\mathrm{n_{sb}}^2)$.

4. Experimental Results

In this section, the detection accuracy of the CCOD($\mathrm{A_i}$,$\mathrm{A_j}$) is compared with that of the FindCBLOF($\mathrm{A_i}$) (He et al., 2003) approach using the k-means (KM) (Hartigan and Wong, 1979) , BKM (Bisecting k-means) (Savaresi and Boley, 2001), or PAM (Partitioning Around Medoids) (Barnett et al., 1994) algorithms. Where $\mathrm{A_i}$, $\mathrm{A_j}$ $\in${KM, BKM, PAM}.

4.1. Datasets

Experiments were performed on a number of artificial (HBK dataset (Hawkins et al., 1984) and Wood Dataset (Leroy and Rousseeuw, 1987)), gene expression (Breast Cancer (West et al., 2001)) and documents datasets (Yahoo) with various characteristics and degree of

Table 1: Experimental Datasets

Dataset	***n***	***K***	***d***
HBK	75	3	4
Wood	20	2	6
BC	7129	4	49
Yahoo	2,340	20	28,298

Table 2: Parameter Settings

Parameter	***Value/Range***	***Algorithm/Dataset***
α	90%	*FindCBLOF*
β	5	*FindCBLOF*
α_{sb}	70%	CCOD
MinPts	10	LOF
δ	[0.1, 0.25]	Gene Expression
δ	[0.2, 0.3]	Document datasets

outliers. The HBK is artificially generated random dataset with 75 observations in four dimensions. The dataset contains 14 outliers. The Wood dataset consists of 20 observations with data points 4, 6, 8, and 19 being outliers. The Yahoo dataset is a collection of Reuter's news articles from the Yahoo! News website (Boley et al., 1999). The settings of the parameters of the adopted approaches as well as the experimental datasets are shown in Tables 1 and 2.

4.2. Significance of Results

Since the actual underlying means and standard deviations are not known, we have used a two-sample t-statistic in which the population standard deviations are estimated by the sample standard deviations sd_1 and sd_2. Let q_1 and q_2 be the two samples of the evaluation measure q (number of detected outliers or SI (Separation Index) (Maulik and Bandyopadhyay, 2002)) for the results of both A_1 and A_2, respectively. Our null hypothesis (which we will argue to be rejected) is defined as follows:

$$H_0 : q_1 = q_2 \text{ (No significant improvement in q using } A_1) \tag{9}$$

where $q1$ is the average q value of A_1 clustering over n_1 samples, and $q2$ is the corresponding average value of q of A_1 clustering over n_2 samples. The confidence interval of the difference between the two means at a confidence level α is given by:

$$(q_1 - q_2) \pm t^{\text{critical}} \sqrt{\frac{sd_1^2}{n_1} + \frac{sd_2^2}{n_2}} \tag{10}$$

4.3. Detection Accuracy of the CCOD Algorithm Using LOF

In this section, the comparison is established based on the number of detected outliers that are occurred in the *TopRatio* outliers detected by the LOF algorithm. We use the LOF as a basis for comparison. The number of the selected top outliers ranges from 10% to 30% of the dataset size. We selected the *LocalTopRatio* possible outliers as (log(nSb)/log(k))* *TopRatio*. For the non-cooperative FindCBLOF, the results are reported in terms of the mean and standard deviation evaluated over 20 runs. For the cooperative model CCOD(A_1, A_2), the reported result is a tuple (q , sd, t_1, t_2). t_1, and t_2 are the t-test values between the results of {CCOD(A_1, A_2) and A_1} and the results of {CCOD(A_1, A_2) and A_2} respectively. Tables 3 and 4 show the number of the discovered outliers using the cooperative detection algorithms compared to that of the FindCBLOF using individual clustering for the Breast Cancer and Yahoo datasets, respectively. In each table the value of the calculated t is greater than the critical value of t (from the t- distribution tables) which means that there is a statistical difference in the obtained results and thus the Null hypothesis (no significant difference) is rejected. Also, it can be shown that the cooperative detection algorithms, CCOD(KM,BKM),CCOD(KM,PAM), and CCOD(BKM,PAM) are able to detect more outliers than the FindCBLOF at different values of the TopRatio ranges from 10% to 30%.

Table 3: Number of Detected Outliers for the Breast Cancer Dataset

TopRatio	*FindCBLOF (KM)*	*FindCBLOF (BKM)*	*FindCBLOF (PAM)*	**CCOD (KM,BKM)**	**CCOD (BKM,PAM)**	**CCOD (BKM,PAM)**
10%	28(4)	31(4)	45(3)	**55 (3)**	**67(4)**	**79(5)**
				t1=24.14	*t1*=30.83	*t1*=33.52
				t2=21.46	*t2*=17.39	*t2*=26.08
15%	39(2)	48(4)	52(3)	**63(3)**	**79(5)**	**83(4)**
				t1=29.76	*t1*=33.21	*t1*=27.77
				t2=13.42	*t2*=12.27	*t2*=27.72
20%	50(3)	62(4)	69(3)	**83 (4)**	**98(6)**	**104(5)**
				t1=29.51	*t1*=32.00	*t1*=29.33
				t2=16.60	*t2*=19.33	*t2*=26.84
25%	68(3)	79(3)	81(4)	**97 (3)**	**107(5)**	**118(6)**
				t1=30.56	*t1*=29.91	*t1*=26.00
				t2=18.97	*t2*=18.16	*t2*=22.94
30%	76 (3)	84 (5)	93(4)	**116(6)**	**125(7)**	**139(6)**
				t1=26.66	*t1*=28.77	*t1*=31.49
				t2=18.32	*t2*=17.75	*t2*=14.26

Table 4: Number of Detected Outliers for the Yahoo Dataset

TopRatio	*FindCBLOF* *(KM)*	*FindCBLOF* *(BKM)*	*FindCBLOF* *(PAM)*	**CCOD** **(KM,BKM)**	**CCOD** **(KM,PAM)**	**CCOD** **(BKM,PAM)**
10%	35(4)	51(4)	31(3)	**67(3)**	**40(3)**	**58(3)**
				t1=28.62	*t1*=4.47	*t1*=6.26
				t2=14.32	*t2*=9.49	*t2*=28.46
15%	43(3)	62(4)	40(3)	**75(4)**	**47(4)**	**69(2)**
				t1=28.62	*t1*=3.58	*t1*=7.00
				t2=10.28	*t2*=6.26	*t2*=35.97
20%	64(4)	78(5)	57(3)	**94(5)**	**67(2)**	**85(3)**
				t1=20.95	*t1*=3.00	*t1*=5.36
				t2=10.12	*t2*=12.40	*t2*=29.51
25%	81(5)	99(4)	72(4)	**117(3)**	**86(3)**	**106(2)**
				t1=27.61	*t1*=3.84	*t1*=7.00
				t2=17.89	*t2*=12.52	*t2*=34.00
30%	94(4)	105(6)	87(5)	**124(7)**	**97(3)**	**113(3)**
				t1=16.64	*t1*=7.66	*t1*=5.33
				t2=9.21	*t2*=2.86	*t2*=19.94

4.4. Detecting True Outliers

Table 5 shows the number of the detected true outliers in the top 20 outliers for the FindCBLOF Versus that of the cooperative models on the HBK and Wood datasets. The cooperative models clearly identify the true outlier records compared to the other methods.

Table 5: The number of true detected outliers in the top 20 outliers

Dataset	*Find CBLOF (KM)*	*Find CBLOF (BKM)*	*Find CBLOF (PAM)*	CCOD (KM,BKM)	CCOD (KM,PAM)	CCOD (BKM,PAM)
HBK	5 ± 1	7 ± 1	9 ± 1	9 ± 1	11 ± 1	12 ± 1
Wood	1 ± 1	1 ± 1	2 ± 1	2 ± 1	3 ± 1	3 ± 1

4.5. Triple Cooperation

An interesting observation is that across the four datasets, the detection accuracy of the triple cooperation CCOD(KM,BKM,PAM) is much better than that of the other pair-wise

cooperative algorithms at different values of the *TopRatio*. This better discovery of outliers is mainly based on the capability of obtaining better clustering solutions using the triple cooperative clustering than that of the pair-wise cooperative techniques.

Table 6: Number of Detected Outliers Using Triple Cooperation CCOD(KM,BKM,PAM)

TopRatio	*Yahoo*	*Breast Cancer*	*HBK*	*Wood*
10%	71(4)	88(6)	12(1)	3(1)
15%	86(5)	97(7)	12(1)	3(1)
20%	116(7)	127(8)	13(1)	4(1)
25%	124(3)	139(8)	13(1)	4(1)
30%	147(6)	152(7)	14(1)	4(1)

4.6. Enhancing Clustering Quality

In order to illustrate the significance of the cooperative detetction methods, the clustering performance of the KM, BKM, and PAM is compared to that of the cooperative detection models using the SI (Separation Index) (Maulik and Bandyopadhyay, 2002) before and after removing the discovered set of outliers at a variable number of *TopRatio* for the Breast Cancer and Yahoo datasets. The SI Index is used as internal quality measure, which does not require prior knowledge about the data. It is defined as the ratio of average within-cluster variance (cluster scatter) to the minimum pair-wise dissimilarity (measured by the cosine correlation measure) between clusters. The smaller the SI, the more separate the clusters.

Table 7: Percentage of Improvement in the SI

Dataset	CCOD (KM,BKM)	CCOD (KM,PAM)	CCOD (BKM,PAM)	CCOD (KM,BKM,PAM)
Yahoo	40%	32%	38%	70%
Breast Cancer	36%	47%	52%	60%

The CCOD(KM,BKM,PAM) achieves up to 60% improvement for Breast Cancer dataset, up to 70% improvement for Yahoo dataset at TopRatio=30% compared to that of the non-cooperative FindCBLOF after removing the discovered outliers.

5. Conclusions and Future work

In this paper, a novel clustering-based outlier detection algorithm (CCOD) is presented that uses the notion of cooperative clustering towards better detection of outliers. This approach is based on assigning a cooperative outlier factor to each object and recognizing the set of candidate outliers after each merging step in the cooperative clustering model. The CCOD algorithm relies on the fact that cooperative clustering outperforms non-cooperative clustering to achieve better detection of outliers in the data. Experimentally, the CCOD is applied on both gene expression datasets and text document datasets. Undertaken experimental results indicate that CCOD works better than the adopted traditional clustering-based outlier detection techniques with better improvement in the clustering quality after removing the discovered set of outliers. Future directions include: (1) Investigating the detection accuracy of the CCOD if the cooperative merging factor is assigned to each object after the cooperative clustering is performed, (2) Determining the proper value of the ratio between the LocalTopRatio and the TopRatio parameters in order to achieve the desired detection capability of the proposed algorithm, (3) Testing the scalability of the detection approach using more than three clustering techniques, and (4) Developing a distributed outlier detection using cooperative clustering to examine the performance of the cooperative clustering in detecting outliers in distributed networks and also finding the global set of outliers across the whole network.

References

Charu C Aggarwal. An introduction to outlier analysis. In *Outlier analysis*, pages 1–40. Springer, 2013.

Andreas Arning, Rakesh Agrawal, and Prabhakar Raghavan. A linear method for deviation detection in large databases. In *KDD*, pages 164–169, 1996.

Vic Barnett, Toby Lewis, et al. *Outliers in statistical data*, volume 3. Wiley New York, 1994.

Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and J"ö"rg Sander. Lof: identifying density-based local outliers. In *ACM sigmod record*, volume 29, pages 93–104. ACM, 2000.

Lei Cao, Di Yang, Qingyang Wang, Yanwei Yu, Jiayuan Wang, and Elke A Rundensteiner. Scalable distance-based outlier detection over high-volume data streams. In *Data Engineering (ICDE), 2014 IEEE 30th International Conference on*, pages 76–87. IEEE, 2014.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection for discrete sequences: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 24(5): 823–839, 2012.

Daniel Rodr"í"guez Dominguez, Rebeca P D"í"az Redondo, Ana Fern"á"ndez Vilas, and Mohamed Ben Khalifa. Sensing the city with instagram: Clustering geolocated data for outlier detection. *Expert Systems with Applications*, 78:319–333, 2017.

Martin Ester, Hans-Peter Kriegel, J"ö"rg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. Rock: A robust clustering algorithm for categorical attributes. *Information systems*, 25(5):345–366, 2000.

Yawwani Gunawardana, Shuhei Fujiwara, Akiko Takeda, Jeongmin Woo, Christopher Woelk, and Mahesan Niranjan. Outlier detection at the transcriptome-proteome interface. *Bioinformatics*, 31(15):2530–2536, 2015.

John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

Douglas M Hawkins, Dan Bradu, and Gordon V Kass. Location of several outliers in multiple-regression data using elemental sets. *Technometrics*, 26(3):197–208, 1984.

Zengyou He, Xiaofei Xu, and Shengchun Deng. Squeezer: an efficient algorithm for clustering categorical data. *Journal of Computer Science and Technology*, 17(5):611–624, 2002.

Zengyou He, Xiaofei Xu, and Shengchun Deng. Discovering cluster-based local outliers. *Pattern Recognition Letters*, 24(9):1641–1650, 2003.

Mon-Fong Jiang, Shian-Shyong Tseng, and Chih-Ming Su. Two-phase clustering process for outliers detection. *Pattern recognition letters*, 22(6):691–700, 2001.

Rasha Kashef and Mohamed S Kamel. Towards better outliers detection for gene expression datasets. In *Biocomputation, Bioinformatics, and Biomedical Technologies, 2008. BIOTECHNO'08. International Conference on*, pages 149–154. IEEE, 2008.

Rasha Kashef and Mohamed S Kamel. Cooperative clustering. *Pattern Recognition*, 43(6): 2315–2329, 2010.

Annick M Leroy and Peter J Rousseeuw. Robust regression and outlier detection. *Wiley Series in Probability and Mathematical Statistics, New York: Wiley, 1987*, 1987.

Ujjwal Maulik and Sanghamitra Bandyopadhyay. Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(12):1650–1654, 2002.

Spiros Papadimitriou, Hiroyuki Kitagawa, Phillip B Gibbons, and Christos Faloutsos. Loci: Fast outlier detection using the local correlation integral. In *Data Engineering, 2003. Proceedings. 19th International Conference on*, pages 315–326. IEEE, 2003.

Sergio M Savaresi and Daniel L Boley. On the performance of bisecting k-means and pddp. In *Proceedings of the 2001 SIAM International Conference on Data Mining*, pages 1–14. SIAM, 2001.

Guido van Capelleveen, Mannes Poel, Roland M Mueller, Dallas Thornton, and Jos van Hillegersberg. Outlier detection in healthcare fraud: A case study in the medicaid dental domain. *International journal of accounting information systems*, 21:18–31, 2016.

Mike West, Carrie Blanchette, Holly Dressman, Erich Huang, Seiichi Ishida, Rainer Spang, Harry Zuzan, John A Olson, Jeffrey R Marks, and Joseph R Nevins. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences*, 98(20):11462–11467, 2001.

Rui Xu and Donald Wunsch. Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678, 2005.

Kenji Yamanishi, Jun-Ichi Takeuchi, Graham Williams, and Peter Milne. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. *Data Mining and Knowledge Discovery*, 8(3):275–300, 2004.