STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **6**, June 2018, pp 178–188.
Published online in International Academic Press (www.IAPress.org)

Weighted Clustering for Anomaly Detection in Big Data

Rasim Alguliyev, Ramiz Aliguliyev, Yadigar Imamverdiyev, Lyudmila Sukhostat*

Institute of Information Technology, Azerbaijan National Academy of Sciences, Azerbaijan

Abstract In this paper, a new method for anomaly detection based on weighted clustering is proposed. The weights that were obtained by summing the weights of each point from the data set are assigned to clusters. The comparison is made using seven datasets (of large dimensions) with the k-means algorithm. The proposed approach increases the reliability of data partitioning into groups. Experimental results show that the proposed approach becomes more efficient with increasing size of the analysed dataset.

Keywords Clustering, Weighted Clustering, Clustering Evaluation Metrics, Big Data; Anomaly Detection; K-means

AMS 2010 subject classifications 62H30, 62-07

DOI: 10.19139/soic.v6i2.404

1. Introduction

With the emerging computer technology, the amount of data a person can work with has increased significantly. For a long time, scientists have been developing algorithms aimed at simplifying the work with data, identifying new, previously unknown knowledge stored in the data. Computing systems are limited in the storage, analysis, and processing of Big data due to its volume, speed or diversity. Large amounts of data can be described by adding veracity and value to volume, variety, and velocity [1]. Thus, with the help of five V, it is possible to find a new understanding of the existing Big data. However, today the amount of stored data is becoming too large to be processed by traditional algorithms. Researchers have to resort to various tricks, such as working with parts of available data, using a priori knowledge about the available data. Thus, working with Big data raises the need to formalize new methods used by researchers, creates new algorithms and software tools that use the power of previously created tools to work with data of large volumes and dimensions. The problem of anomaly (outliers) detection is one of the problems of Big data analysis. It is widely used in the following areas: intrusion detection in computer networks, fraud detection in banking transactions, medicine, monitoring the movement of trains, failure detection of spacecraft systems, etc. [2].

One of the most convenient and understandable approaches for Big data processing is clustering. The problem of clustering is becoming more and more relevant in many areas. Currently, many different clustering algorithms have been developed. Their complexity depends on the dimensionality of the data, the volume of the clustering set, the scope of application, and so on.

In 2003, three requirements for data flows clustering algorithms were formulated [3]: 1) data compression and expression of the compressed data; 2) processing new data points in a fast and incremental way; 3) distinguishing outliers quickly and clearly. A lot of works are devoted to the development and application of clustering algorithms (mainly various modifications of the k-means algorithm are used) to data flows [4, 5]. Cluster analysis can be used

*Correspondence to: Lyudmila Sukhostat (Email: lsuhostat@hotmail.com). Institute of Information Technology, Azerbaijan National Academy of Sciences. 9A, B. Vahabzade Street, Baku AZ1141, Azerbaijan.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

to get an idea of the data, generate a hypothesis, and detect anomalies and classification. Applications are often defined in terms of outliers (for example, in case of fraud detection, anomaly detection in the network, etc.), in which case a direct approach is likely to be more effective [6, 7]. The difficulty of anomaly (outliers) detection is to label patterns of normal and abnormal behaviors that are not easy to obtain [8, 9]. However, the chosen approach for anomaly detection can only be suitable for a certain range of tasks, but not for all [10].

The aim of this paper is to develop a clustering approach for anomaly detection in real Big data. Working with Big data requires large computational resources. For a previously known number of clusters, according to [11], we propose an algorithm based on weighted clustering.

The rest of the paper is organized as follows. Section 2 gives a literature review of existing works on clustering large amounts of data. The proposed weighted clustering method is described in Section 3. In Section 4, datasets and clustering evaluation metrics are presented. The experimental results and discussion are given in Section 5, followed by conclusions in Section 6.

2. Related Work

In [12] a new clustering algorithm was proposed. It shows good results in accordance with the three metrics of clustering (volume, variety, and veracity). This approach works with large amounts of data and showed a compromise between the quality of clustering and runtime.

In order to work with Big data, an approach combining the principal component analysis (PCA) and the k-means algorithm was proposed [13]. Due to the randomized preconditioning transformation, it is possible to achieve accurate and reliable estimates in the data sparsification process.

The proposed sparsified K-means algorithm returns both assignments and cluster centers in a single pass over the data, while the state-of-the-art feature-based algorithms require at least two passes. Preconditioning and sampling technique could be used to either speed up computation for in-core memory problems, or to create one-pass variants for out-of-core or streaming problems.

A new approach combining K-means and tree-based classification [14] was proposed in order to analyse and visualize the high-dimensional time series. Hellinger distance between density functions is heavily used in the proposed analysis. Sensible results were obtained as a result of experiments on real datasets.

The approach in [15] solves the problem of initialization in the clustering algorithm. At the training stage, the min-max problem was solved iteratively. At each iteration, the weights were updated.

The experimental results have shown the robustness of bad initializations and its efficacy compared to k-Means, k-Means++ [16] and K-Harmonic Means [17].

A modified k-means algorithm (KMOR) was proposed for data clustering and outlier detection in [18]. The experiments were performed on synthetic and real datasets. The proposed algorithm was compared with ODC (Outlier Detection and Clustering) algorithm [19]. The results showed the superiority of the KMOR algorithm for accuracy and run-time.

In work [20] it is emphasized that the developed solutions allow for estimating not only the temporal data centroid but also its weighting vector, which indicates the representativeness of the centroid elements. The impact of the isotropy and isolation of clusters on the effectiveness of the clustering methods was also discussed. The proposed solutions can be directly applicable to any other variations of k-means.

An alternative clustering approach was proposed in [21]. It is quite simple and consists in fitting the data to the clustering model. The method is designed as a clustering algorithm where the initial structure is not important.

The main task of [22] is to increase the performance of the k-means algorithm for large datasets. An experimental comparison was made with other clustering methods. The results have shown high performance of the proposed hierarchical k-means (H-K-means) algorithm.

3. Proposed Approach

In this section, the authors propose a clustering method for anomaly detection and describe the details of this algorithm.

Let us denote the following notations: $X = (x_1, x_2, ..., x_n)$ are the points in the dataset, where n is the total number of data points in the dataset, $X = (x_{i1}, x_{i2}, ..., x_{im}) \in R^m$ is the point in the dataset, where m is the dimension of data points, $C = (C_1, C_2, ..., C_k)$ are clusters, where C_p $(p = \overline{1,k})$ is the p^{th} cluster and k is the number of clusters. The task is to minimize the following function in order to detect anomalies in the dataset as follows:

$$f(x) = \sum_{p=1}^{k} \sum_{x_i \in C_p} |C_p|_W * \|x_i - O_p\|^2 \to min, \tag{1}$$

where $|C_p|_W$ is the weight of the p^{th} cluster.

In this case, the clusters weight is determined as a sum of the weights of all points in the cluster:

$$|C_p|_W = \sum_{x_i \in C_p} w(x_i),\ p = 1, 2, ...k, \tag{2}$$

where weights of points are calculated on the basis of their distance from the center of all points in the dataset

$$w(x_i) = \|x_i - O\|,\ O = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{3}$$

and the center of the p^{th} cluster (O_p) is defined as

$$O_p = \frac{1}{n_p} \sum_{x_i \in C_p} x_i,\ n_p = |C_p|,\ p = 1, 2, ..., k. \tag{4}$$

The algorithm of the proposed method for anomaly detection is as follows:

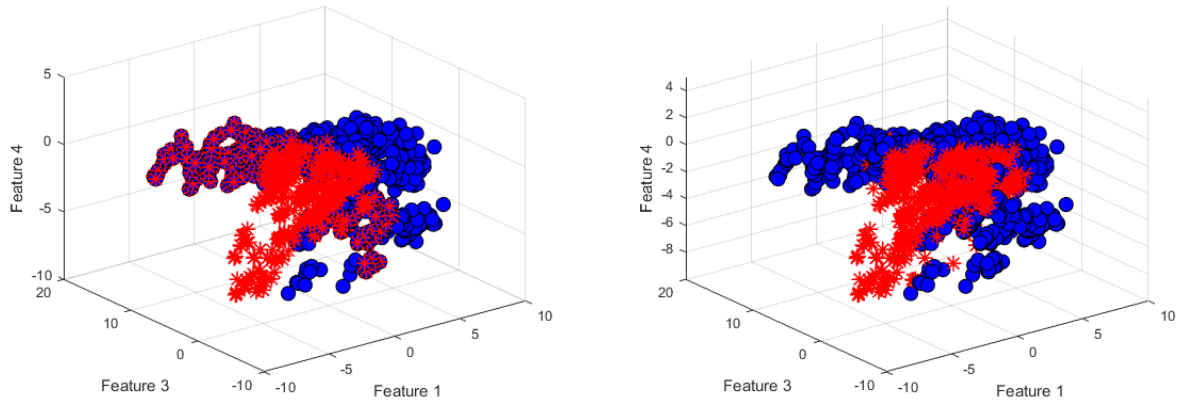
Input: $X = (x_1, x_2, ..., x_n)$,
$w = (w_1, w_2, ..., w_n)$,
k: number of clusters.
Output: Vector of cluster indices $IDX = (idx_1, idx_2, ..., idx_n)$.
Step 1: Find the center of all points of the dataset (O)
Step 2: Calculate the weights of all points x_i according to (3)
Step 3: $s = 0$
Step 4: Calculate the value of the function according to (1) taking into account (2)

$$f^{(s)} = \sum_{p=1}^{k} \sum_{x_i \in C_p} |C_p|_W * \left\| x_i - O_p^{(s)} \right\|^2 \tag{5}$$

Step 5: $s = s + 1$
Step 6: Repeat steps 3-5 until the convergence condition is met:

$$\left| \frac{f^{(s+1)} - f^{(s)}}{f^{(s)}} \right| \leqslant \varepsilon, \tag{6}$$

where s is the number of iterations.
Step 7: Return the values of IDX
End



(a) (b)

Figure 1. Comparison of experimental results on a real dataset (Banknote authentication dataset): (a) k-means algorithm, (b) proposed approach.

In the algorithm, each cluster is represented by its center, and the goal is to find a solution that minimizes the distance between each point and the cluster center to which it is assigned [23].

Figure 1 shows a comparison of the proposed approach and the k-means algorithm using Banknote authentication dataset as an example. The following characteristics of the dataset were considered: variance of Wavelet Transformed (WT) image, the kurtosis of WT image and entropy of image. The dataset contains two clusters: blue filled circles (”normal” values) and red stars (anomalies). Figure 1 visually shows the effectiveness of the proposed approach.

4. Datasets and Evaluation Metrics

This section compares the performance of the proposed method and k-means algorithm. First, we will describe the datasets that were used to conduct the experiments.

4.1. Datasets

The experiments were performed on six datasets from the UCI repository [24, 25], including Diabetic Retinopathy Debrecen dataset (Diabetic), MAGIC Gamma Telescope DataSet (Magic04), Banknote authentication, Credit card clients, Forest CoverType dataset (Covertype) and Phishing dataset, and NSL-KDD dataset [26]. These datasets are medium and large in size and are used in many research areas.

Diabetic Retinopathy (DR) Debrecen Dataset contains features extracted from the Messidor image set to predict whether an image contains signs of diabetic retinopathy or not [27, 28]. It contains 19 features (the Euclidean distance of the center of the macula and the center of the optic disc, the binary result of the AM/FM-based classification, etc.) with 1151 samples. The 20th attribute is a class label, i.e. 1 (contains signs of DR) and 0 (no signs of DR). In this paper samples with class label equal to 1 were considered as anomalous values.

MAGIC 04 Dataset was generated to simulate registration of high energy gamma particles in an atmospheric Cherenkov telescope, taking advantage of the radiation emitted by charged particles produced inside the electromagnetic showers initiated by the gammas, and developing in the atmosphere. The dataset was generated by Corsika program [29]. It was collected from 19020 samples. It contains 10 features (axis of the ellipse, the ratio of the highest pixel, projected onto the major axis, etc.). The 11th attribute is a class label, i.e. gamma (signal) and hadron (background). In the paper, samples with the gamma class label were considered as an anomaly.

Banknote Authentication Dataset was extracted from images that were taken from genuine and forged banknote-like specimens [25]. The images have a size of 400x 400 pixels. WT tool was used to extract features from images. The dataset contains four features (variance of WT image, the skewness of WT image, the kurtosis of WT image, and entropy of image) with 1372 samples. Samples with labels of counterfeit banknotes were taken as anomalies.

Credit Card Clients Dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005 [30]. It contains 30000 samples. This research employed a binary variable, default payment (Yes=1, No=0), as the response variable. The dataset contains 23 features: the amount of the given credit, gender (male/female), education, marital status, age (year), history of past payment, etc.

NSL-KDD Dataset of attack signatures [26] was constructed based on KDD-99 database [31]. To conduct research in the field of intrusion detection, a set of communication data was compiled and covered a wide range of various intrusions simulated in an environment that mimics the US Air Force network. The database contains training (125973 samples) and test (22544 samples) sets. Each instance has 42 attributes. Labels are assigned to each instance either as an "attack" type or as "normal" behavior. The total number of samples was 148517 (NSL-KDD_All).

Covertype Dataset includes information about four wilderness areas located in the Roosevelt National Forest of northern Colorado (USA) [32]. These areas represent forests with minimal human-caused disturbances, so that existing forest cover types are more a result of ecological processes rather than forest management practices. This dataset contains 54 features (elevation in meters, distance to nearest surface water features, soil types, etc.) with 581012 samples. Dataset classes include Spruce/Fir (1), Lodgepole Pine (2), Ponderosa Pine (3), Cottonwood/Willow (4), Aspen (5), Douglas-fir (6) and Krummholz (7). In this paper, 1-6 classes were considered as normal values, and samples with class label Krummholz - as anomalies.

Phishing Dataset contains 11055 phishing websites [33]. It includes 30 attributes (using the IP address, URL length, abnormal URL, website forwarding, etc.). This data belongs to one of the two classes labeled as Phishy (-1) and Legitimate (1).

4.2. Clustering Evaluation Metrics

As a result of clustering algorithms application, it is necessary to estimate the quality of the obtained partitions. To do this, the quality assessment indices were considered. Six quality metrics having different nature were selected for the analysis [34].

Assume that the dataset N is divided into classes $C^+ = \left(C_1^+, ..., C_{k^+}^+\right)$ (true clustering), and, using the clustering procedure, clusters $C = (C_1, ..., C_k)$ can be found in the dataset, where k^+ is the initial number of classes, k is the number of clusters that need to be found [35].

A comparison of the clustering solutions is based on counting the pairs of points. A decision ("normal"/abnormal behavior) will be made based on the results. The most well-known clustering distance metrics based on data point pairs are the purity [36, 37], the Mirkin metric [38], the partition coefficient [39], the variation of information [40], the F-measure [41] and the V-measure [41].

Purity. The purity of the cluster C_p $\left(p = \overline{1,k}\right)$ gives the ratio of the dominant class size in the cluster to the cluster size itself [36, 37, 42]. The value of the purity is always in the interval $\left[\frac{1}{k^+}, 1\right]$. The purity of the entire collection of clusters can be evaluated as a weighted sum of the individual cluster purities:

$$purity(C) = \frac{1}{n}\sum_{p=1}^{k} \max_{p^+=1,...,k^+} \left|C_p \bigcap C_{p^+}^+\right|. \tag{7}$$

A higher purity value indicates a better clustering solution.

Mirkin metric. The Mirkin metric is defined as follows [38]:

$$M(C, C^+) = \frac{1}{n^2}\left(\sum_{p=1}^{k}|C_p|^2 + \sum_{p^+=1}^{k^+}\left|C_{p^+}^+\right|^2 - 2\sum_{p=1}^{k}\sum_{p^+=1}^{k^+}\left|C_p \bigcap C_{p^+}^+\right|^2\right). \tag{8}$$

The smaller the metric value, the better clustering.

F-measure. Another evaluation measure, also known as the "clustering accuracy", is based on the F value of the cluster C_p and the class $C_{p^+}^+$ $\left(p^+ = \overline{1,k^+}\right)$, that is the harmonic mean of the precision and the recall. Precision and recall are computed as follows [35]:

$$P\left(C_p, C_{p^+}^+\right) = \frac{\left|C_p \bigcap C_{p^+}^+\right|}{|C_p|}, \tag{9}$$

$$R\left(C_p, C_{p^+}^+\right) = \frac{\left|C_p \bigcap C_{p^+}^+\right|}{\left|C_{p^+}^+\right|}. \tag{10}$$

Thus, the F-measure has the following form:

$$F\left(C_p, C_{p^+}^+\right) = \frac{2P\left(C_p, C_{p^+}^+\right) R\left(C_p, C_{p^+}^+\right)}{P\left(C_p, C_{p^+}^+\right) + R\left(C_p, C_{p^+}^+\right)}. \tag{11}$$

The F-measure of the cluster C_p is the maximum F-value attained at any class in the entire set of classes $C^+ = \left(C_1^+, ..., C_{k^+}^+\right)$. The F-measure of the entire dataset is considered to be the weighted sum of the individual cluster F-measures. That is,

$$F\left(C\right) = \sum_{p=1}^{k} \frac{|C_p|}{n} \max_{C_{p^+}^+ \in C^+} F\left(C_p, C_{p^+}^+\right). \tag{12}$$

The higher the F-measure, the better clustering solution.

Partition coefficient (PC). This coefficient is used to compare $C = (C_1, ..., C_k)$ and $C^+ = \left(C_1^+, ..., C_{k^+}^+\right)$ distributions [39]. According to [42], PC is calculated as:

$$PC\left(C, C^+\right) = \frac{1}{kk^+} \sum_{p=1}^{k} \sum_{p^+=1}^{k^+} \left(\frac{\left|C_p \bigcap C_{p^+}^+\right|}{|C_p|}\right)^2. \tag{13}$$

A higher value of $PC\left(C, C^+\right)$ indicates a better clustering solution.

Variation of information (VI). This metric measures the amount of information that the authors gain and lose when going from the clustering C to another clustering C^+ [40, 42].

$$VI\left(C, C^+\right) = \frac{1}{n \log n} \sum_{p=1}^{k} \sum_{p^+=1}^{k^+} \left|C_p \bigcap C_{p^+}^+\right| \log\left(\frac{|C_p| \left|C_{p^+}^+\right|}{\left|C_p \bigcap C_{p^+}^+\right|^2}\right). \tag{14}$$

In general, the smaller the VI, the better clustering solution.

V-measure. The V-measure is an entropy-based measure that explicitly measures how successfully the criteria of homogeneity and completeness have been satisfied [37]. The homogeneity can be defined as

$$hom(C) = \begin{cases} 1, & if\ H\left(C^+|C\right) = 0 \\ 1 - \frac{H\left(C^+|C\right)}{H\left(C^+\right)}, & else \end{cases}, \tag{15}$$

where

$$H\left(C^+|C\right) = -\sum_{p=1}^{k} \sum_{p^+=1}^{k^+} \frac{\left|C_p \bigcap C_{p^+}^+\right|}{n} \log\left(\frac{\left|C_p \bigcap C_{p^+}^+\right|}{\sum_{p^+=1}^{k^+} \left|C_p \bigcap C_{p^+}^+\right|}\right), \tag{16}$$

$$H\left(C^{+}\right)=-\sum_{p^{+}=1}^{k^{+}}\frac{\sum_{p=1}^{k}\left|C_{p}\bigcap C_{p^{+}}^{+}\right|}{k^{+}}\log\left(\frac{\sum_{p=1}^{k}\left|C_{p}\bigcap C_{p^{+}}^{+}\right|}{k^{+}}\right). \tag{17}$$

$H\left(C^{+}|C\right)$ is equal to 0 when each cluster contains only members of a single class, a perfect homogeneous clustering. In the degenerate case when $H\left(C^{+}\right)$ is equal to 0, when there is only a single class, the homogeneity is defined to be 1.

Completeness is symmetric to homogeneity. The completeness can be defined as

$$comp(C)=\begin{cases}1, & if\ H\left(C|C^{+}\right)=0\\ 1-\frac{H\left(C|C^{+}\right)}{H(C)}, & else\end{cases}, \tag{18}$$

where

$$H\left(C|C^{+}\right)=-\sum_{p^{+}=1}^{k^{+}}\sum_{p=1}^{k}\frac{\left|C_{p}\bigcap C_{p^{+}}^{+}\right|}{n}\log\left(\frac{\left|C_{p}\bigcap C_{p^{+}}^{+}\right|}{\sum_{p=1}^{k}\left|C_{p}\bigcap C_{p^{+}}^{+}\right|}\right), \tag{19}$$

$$H\left(C\right)=-\sum_{p=1}^{k}\frac{\sum_{p^{+}=1}^{k^{+}}\left|C_{p}\bigcap C_{p^{+}}^{+}\right|}{k^{+}}\log\left(\frac{\sum_{p^{+}=1}^{k^{+}}\left|C_{p}\bigcap C_{p^{+}}^{+}\right|}{k^{+}}\right). \tag{20}$$

V-measure of the clustering solution is calculated by finding the harmonic mean of homogeneity and completeness as follows:

$$V(C)=\frac{2hom(C)comp(C)}{hom(C)+comp(C)}. \tag{21}$$

The computation of the homogeneity, the completeness, and the V-measure are completely independent from the number of classes and clusters, the size of the dataset and the clustering algorithm.

5. Experimental Results and Discussion

To evaluate the performance of the proposed approach, a number of experiments were implemented in Matlab 2016a on a 64-bit Windows-based system with an Intel core (i7), 2.5 GHz processor machine with 8 Gbytes of RAM.

Experimental datasets Diabetic, Magic04, Banknote authentication, Credit card clients, NSL-KDD_All, Covertype, and Phishing were used as initial data. The characteristics of the datasets are presented in Table 1. Six quality metrics having different nature were selected for the analysis.

Table 1. Summary of the datasets.

Dataset	*Number of instances*	C_1^+	C_2^+	*Number of attributes*
Diabetic	1151	611	540	19
Magic04	19020	12332	6688	10
Banknote authentication	1372	762	610	4
Credit Card Clients	30000	23364	6636	23
NSL-KDD_All	148517	71463	77054	41
Covertype	581012	20510	560502	54
Phishing	11055	4898	6157	30

During the preprocessing, the values in the datasets were standardized to have a mean of 0 and a standard deviation of 1. All datasets contain two classes: C_1^+ and C_2^+. Samples included in the C_1^+ class are taken as anomalies.

The results of the proposed approach (PA) and k-means (KM) algorithm based on six metrics are presented in Table 2. Purity, Mirkin metric, PC, VI, F-measure, and V-measure were considered as evaluation metrics.

In the proposed algorithm, each cluster is represented by its center, and the task is to find a solution that minimizes the distance between each point and the center of the cluster to which it is assigned, taking into account the weights of the clusters [43].

The proposed approach showed the best results for all metrics on the Covertype dataset: Purity = 96.47%, Mirkin metric = 36.69%, PC = 47.12%, VI= 5.06%, F-measure = 69.24% and V-measure = 1.0000.

Table 2. Comparison of the proposed approach and k-means on different datasets.

Dataset	*Purity*		*Mirkin*		*F-measure*		*VI*		*PC*		*V-measure*	
	PA	*KM*	*PA*	*KM*	*PA*	*KM*	*PA*	*KM*	*PA*	*KM*	*PA*	*KM*
Diabetic	0.5352	0.5308	0.4975	0.5000	0.6552	0.5931	0.1562	0.1783	0.2509	0.2519	1.0004	1.0003
Magic04	0.6484	0.6484	0.4992	0.4953	0.5380	0.5479	0.1349	0.1342	0.2722	0.2713	1.0000	1.0000
Banknote authentication	0.6312	0.6122	0.4656	0.4748	0.6308	0.5637	0.1814	0.1780	0.2681	0.2615	1.0003	1.0003
Credit Card Clients	0.7788	0.7788	0.4609	0.4276	0.6482	0.4234	0.1072	0.0930	0.3250	0.3295	1.0000	1.0000
NSL-KDD_All	0.5384	0.5188	0.4971	0.4993	0.6846	0.6801	0.0720	0.0637	0.2766	0.2502	1.0000	1.0000
Covertype	0.9647	0.9647	0.3669	0.4315	0.6924	0.5859	0.0506	0.0580	0.4712	0.4625	1.0000	1.0000
Phishing	0.5880	0.5569	0.4845	0.4953	0.5817	0.5888	0.0998	0.1213	0.2747	0.2523	1.0000	1.0000

The best result based on the purity metric was obtained for the Covertype dataset and gained 96.47%, which coincided with the value of the same metric on this dataset for the approach we proposed.

According to the Mirkin metric (42.76%), the lowest value was achieved for the Credit card clients dataset. The highest result for the NSL-KDD_All dataset according to F-measure metric is 68.01%. Covertype dataset showed the best results for k-means clustering based on VI (5.80%) and PC (46.28%) metrics.

V-measure does not have a discriminating ability, i.e. its value on different datasets is almost the same for the methods. From this, it can be concluded that the use of the V-measure is not useful for evaluating the results of clustering. Therefore, in the following comparisons, it was not considered. A comparison of the performance of the proposed approach with the k-means algorithm is shown in Table 3.

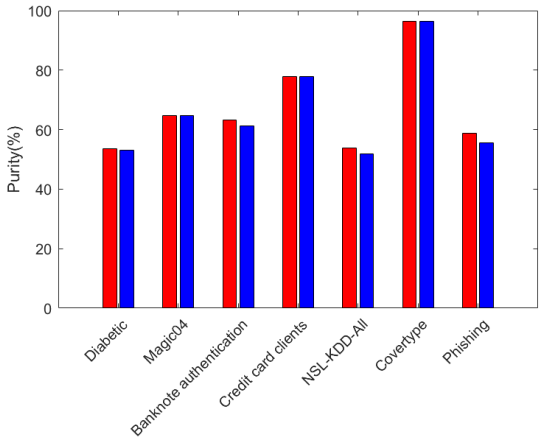
Table 3. Performance evaluation compared between the proposed approach and k-means algorithm.

Dataset	*Purity (%)*	*Mirkin (%)*	*F-measure (%)*	*VI (%)*	*PC (%)*
Diabetic	0.83 (+)	0.50 (+)	10.47 (+)	12.39 (+)	0.40 (-)
Magic04	0	0.79 (-)	1.81 (-)	0.52 (-)	0.33 (+)
Banknote authentication	3.10 (+)	1.94 (+)	11.90 (+)	1.91 (-)	2.52 (+)
Credit Card Clients	0	7.79 (-)	53.09 (+)	15.27 (-)	1.37 (-)
NSL-KDD_All	3.78 (+)	0.44 (+)	0.66 (+)	13.03 (-)	10.55 (+)
Covertype	0	14.97 (+)	18.18 (+)	12.76 (+)	1.88 (+)
Phishing	5.58 (+)	2.18 (+)	1.21 (-)	17.72 (+)	8.88 (+)

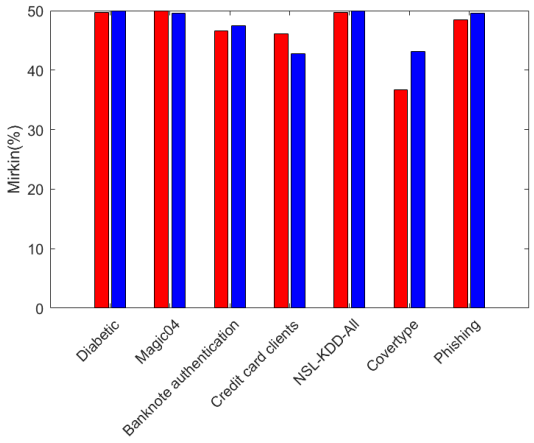
Here the authors use relative improvement for comparison:

$$\frac{our\_method - another\_method}{another\_method} * 100\%. \tag{22}$$

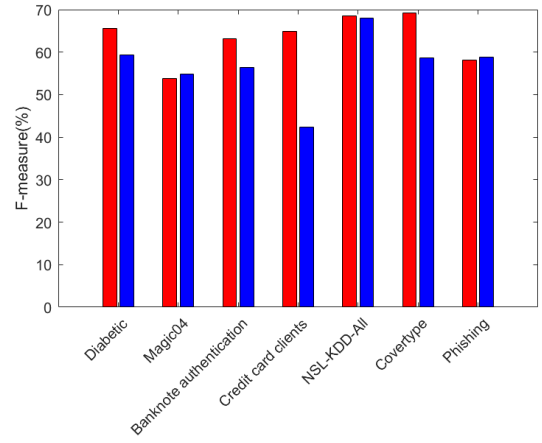
In the Table 3 "+" means that the result outperforms and " - " means the opposite. A comparison of the evaluation metrics values for the proposed approach (red bars) and the k-means algorithm (blue bars) on seven datasets is more clearly illustrated in Fig. 2.



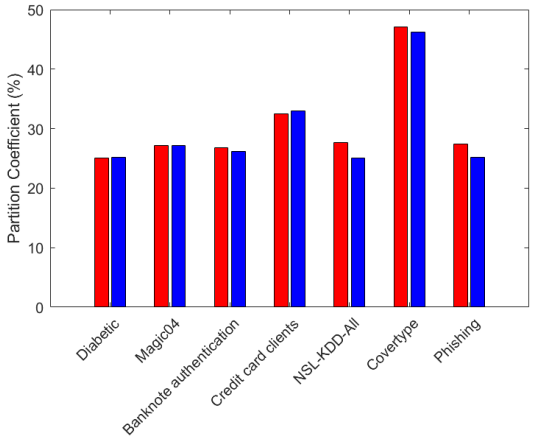
(a)



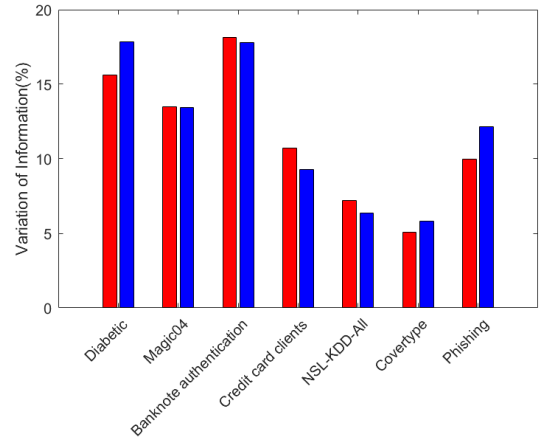
(b)



(c)



(d)



(e)

Figure 2. The comparison of the methods based on evaluation metrics.

Based on the experimental results, it can be concluded that the proposed approach is superior to the k-means algorithm in four metrics (Mirkin metric, F-measure, PC and VI) for the Covertype dataset. Purity, Mirkin, F-measure and VI metrics showed good results for the Diabetic dataset, Purity, Mirkin metric, F-measure and PC for Banknote authentication dataset and NSL-KDD_All. The values of Purity metrics have been the same for two approaches on Magic04, Credit card clients, and Covertype datasets.

High evaluation results for the Phishing dataset were obtained based on Purity, Mirkin, PC and VI metrics. For Magic04 and Credit card clients datasets, the improvement was obtained based on PC (0.33%) and F-measure (53.09%) metrics, respectively.

6. Conclusion

In this paper, a new method for anomaly values detection based on weighted clustering was proposed. The aim of the algorithm presented in the paper was to improve the process of anomaly detection in large data sets. The weights that were obtained by summing the weights of each point from the dataset were assigned to clusters.

In the paper, the weight of each point was determined by its position according to the center in the entire dataset. It can be seen that the weighting improves the clustering solution. The comparison was made using seven datasets (of large dimensions) with the k-means algorithm. The quality of the clustering result was estimated using six evaluation metrics. An important feature of the proposed approach is that it increases the accuracy of anomaly detection based on clustering. The experimental results showed that the proposed algorithm more accurately detects anomalies compared to k-means and has practical significance.

By applying the proposed approach to data clustering, it is possible to increase the reliability of clusters partitioning into groups. It can be concluded that the proposed approach becomes more efficient with increasing size of the analysed dataset. We investigated the effect of cluster weights on performance and the accuracy of finding anomalies in Big data. It is important that this approach can be applied in various research fields. Future research will focus on the development and application of ensembles of clustering algorithms to anomaly detection.

Acknowledgement

This work was supported by the Science Development Foundation under the President of the Republic of Azerbaijan – Grant № EİF-KETPL-2-2015-1(25)-56/05/1.

Conflict of Interests

The authors declare that there is no conflict of interest with respect to research, authorship and publication of this paper.

REFERENCES

1. Y. Zhai, Y.-S. Ong, and I. W. Tsang, *The emerging "Big Dimensionality"*, IEEE Computational Intelligence Magazine, vol. 9, no. 3, pp. 14–26, 2014.
2. V. Chandola, A. Banerjee, and V. Kumar, *Anomaly detection: a survey*, ACM Computing Surveys, vol. 41, no. 3, pp. 1–58, 2009.
3. D. Barbara, *Requirements for clustering data streams*, ACM SIGKDD Explorations Newsletter, vol. 3, no. 2, pp. 23–27, 2003.
4. J. Chandrika and K. R. Ananda Kumar, *Dynamic clustering of high speed data streams*, International Journal of Computer Science, vol. 9, pp. 224–228, 2012.
5. Q. Quan, C.-J. Xiao, and R. Zhang, *Grid-based data stream clustering for intrusion detection*, International Journal of Network Security, vol. 15, pp. 1–8, 2013.
6. Y. Kou, C. Lu, S. Sirwongwattana, and Y. Huang, *Survey of fraud detection techniques*, in Proc. IEEE ICNSC Conference, Taipei, Taiwan, 2004.
7. A. Lazarevic, L. Ertoz, V. Kumar, A. Ozgur, and J. Srivastava, *A comparative study of anomaly detection schemes in network intrusion detection*, in Proc. SIAM Conference on Data Mining, San Francisco, CA, 2003.

8. M. Xie, S. Han, B. Tian, and S. Parvin, *Anomaly detection in wireless sensor networks: a survey*, Journal of Network and Computer Applications, vol. 34, no. 2, pp. 1302–1325, 2011.
9. H. Nallaivarothayan, D. Ryan, S. Denman, S. Sridharan, and C. Fookes, *An evaluation of different features and learning models for anomalous event detection*, in Proc. DICTA Conference, Hobart, Australia, 2013.
10. J. J. Davis and A. J. Clark, *Data preprocessing for anomaly based network intrusion detection: a review*, Computers & Security, vol. 30, no. 6-7, pp. 353–375, 2011.
11. R. M. Alguliyev, R. M. Aliguliyev, A. Bagirov, and R. Karimov, *Batch clustering algorithm for Big data sets*, in Proc. AICT Conference, Baku, 2016.
12. H. Rehioui, A. Idrissi, M. Abourezq, and F. Zegrari, *DENCLUE-IM: a new approach for Big data clustering*, Procedia Computer Science, vol. 83, pp. 560–567, 2016.
13. F. Pourkamali Anaraki and S. Becker, *Preconditioned data sparsification for Big data with applications to PCA and k-means*, IEEE Transactions on Information Theory, vol. 63, no. 5, pp. 2954–2974, 2017.
14. R. S. Tsay, *Some methods for analyzing Big dependent data*, Journal of Business & Economic Statistics, vol. 34, no. 4, pp. 673–688, 2016.
15. G. Tzortzis and A. Likas, *The MinMax k-Means clustering algorithm*, Pattern Recognition, vol. 47, no. 7, pp. 2505–2516, 2014.
16. D. Arthur and S. Vassilvitskii, *K-Means++: the advantages of careful seeding*, in Proc. ACM-SIAM SODA Symposium, New Orleans, Louisiana, 2007.
17. A. Banerje and J. Ghosh, *Frequency-sensitive competitive learning for scalable balanced clustering on high-dimensional hyperspheres*, IEEE Transactions on Neural Networks, vol. 15, no. 3, pp. 702–719, 2004.
18. G. Gan and M. Ng, *k-means clustering with outlier removal*, Pattern Recognition Letters, vol. 90, pp. 8–14, 2017.
19. M. Ahmed and A. Naser, *A novel approach for outlier detection and clustering improvement*, in Proc. IEEE ICIEA Conference, Melbourne, Australia, 2013.
20. S. Soheily-Khah, A. Douzal-Chouakria, and E. Gaussier, *Generalized k-means-based clustering for temporal data under weighted and kernel time warp*, Pattern Recognition Letters, vol. 75, pp. 63–69, 2016.
21. M. I. Malinen, R. Mariescu-Istodor, and P. Franti, *K-means*: clustering by gradual data transformation*, Pattern Recognition, vol. 47, no. 10, pp. 3376–3386, 2014.
22. T. S. Xu, H. D. Chiang, G. Y. Liu, and C. W. Tan, *Hierarchical k-means method for clustering large-scale advanced metering infrastructure data*, IEEE Transactions on Power Delivery, vol. 32, no. 2, pp. 609–616, 2017.
23. F. Jiang, G. Liu, J. Du, and Y. Sui, *Initialization of K-modes clustering using outlier detection techniques*, Information Sciences, vol. 332, pp. 167–183, 2016.
24. J. Eggermont, J. N. Kok, and W. A. Kosters, *Genetic programming for data classification: partitioning the search space*, ACM SAC Symposium, pp. 1001–1005, 2004.
25. M. Lichman, *UCI Machine Learning Repository*, University of California, Available at http://archive.ics.uci.edu/ml, 2013.
26. P. Aggarwal and S. K. Sharma, *Analysis of KDD dataset attributes-class wise for intrusion detection*, Proc. Comp. Sci., vol. 57, pp. 842–851, 2015.
27. B. Antal and A. Hajdu, *An ensemble-based system for automatic screening of diabetic retinopathy*, Knowl.-Based Syst., vol. 60, pp. 20–27, 2014.
28. E. Decenciere, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay, B. Charton, and J.-C. Klein, *Feedback on a publicly distributed database: the messidor database*, Image Analysis & Stereology, vol. 33, pp. 231–234, 2014.
29. D. Heck, J. Knapp, J. N. Capdevielle, G. Schatz, and T. Thouw, *CORSIKA: a Monte Carlo code to simulate extensive air showers*, Forschungszentrum Karlsruhe GmbH, Karlsruhe, Germany, 1998.
30. I. C. Yeh and C. H. Lien, *The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients*, Expert Systems with Applications, vol. 36, no. 2, pp. 2473–2480, 2009.
31. J. McHugh, *Testing intrusion detection systems: a critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln laboratory*, ACM Trans. Inf. and Syst. Sec., vol. 3, pp. 262–294, 2000.
32. J. A. Blackard and J. D. Denis, *Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables*, Comp. and Elect. Agriculture, vol. 24, pp. 131–151, 2000.
33. R. Mohammad, F. A. Thabtah, and T. L. McCluskey, *Phishing websites dataset*, University of Huddersfield, Available at https://archive.ics.uci.edu/ml/datasets/Phishing+Websites, 2015.
34. R. M. Alguliyev, R. M. Aliguliyev, and L. V. Sukhostat, *Anomaly detection in Big data based on clustering*, Statistics, Optimization and Information Computing, vol. 5, no. 4, pp. 325–340, 2017.
35. R. M. Alguliev, R. M. Aliguliyev, T. Kh. Fataliyev, and R. Sh. Hasanova, *Weighted consensus index for assessment of the scientific performance of researchers*, COLLNET J. Scientometrics and Inf. Management, vol. 8, pp. 371–400, 2014.
36. F. Boutin and M. Hascoet, *Cluster validity indices for graph partitioning*, in Proc. ICIV Conference, pp. 376–381, 2004.
37. A. M. Rubinov, N. V. Soukhorukova, and J. Ugon, *Classes and clusters in data analysis*, Euro. J. Operational Research, vol. 173, pp. 849–865, 2006.
38. B. Mirkin, *Mathematical classification and clustering*, J. Global Optimization, vol. 12, pp. 105–108, 1998.
39. J. C. Bezdek and N. R. Pal, *Some new indexes of cluster validity*, IEEE Trans. Syst., Man and Cyber, Part B, vol. 28, pp. 301–315, 1998.
40. A. Patrikainen and M. Meila, *Comparing subspace clusterings*, IEEE Trans. Knowl. and Data Engin., vol. 18, pp. 902–916, 2006.
41. A. Rosenberg and J. Hirschberg, *V-measure: a conditional entropy-based external cluster evaluation measure*, in Proc. EMNLP-CoNLL Conference, pp. 410–420, 2007.
42. R. M. Aliguliyev, *Performance evaluation of density-based clustering methods*, Inf. Sci., vol. 179, pp. 3583–3602, 2009.
43. I. Eyal, I. Keidar, and R. Rom, *Distributed data clustering in sensor networks*, Distrib. Comput., vol. 24, pp. 207–222, 2010.