An Enhanced K-means Clustering Based Outlier Detection Techniques to Improve Water Contamination Detection and Classification

S. Visalakshi and V. Radha

Abstract In many data mining applications, the primary step is detecting outliers in a dataset. Outlier detection for data mining is normally based on distance, clustering and spatial methods. This paper deals with locating outliers in large, multidimensional datasets. The k-means clustering algorithm partitions a dataset into a number of clusters, and then the results are used to find out the outliers from each cluster, using any one of the outlier's detection methods. The k-means clustering algorithm is enhanced in three manners. The first is by using a different distance metric. The second and third enhancements are brought forward by automating the process of estimating 'k' value and initial seed selection using the enhanced clustering algorithm. Outliers are detected in the drinking water dataset after the clustering process is over. The results show that classification accuracy, speeds are improved and normalized root mean square error is reduced.

Keywords K-means · Similarity matrix · Dissimilarity co-efficient · Fixed-width clustering · Distance-based · Density-based

1 Introduction

An 'Outliers' is defined as an examination that is radically different from the other data in its set. Outliers are also referred as abnormalities, discordant, deviants or anomalies in data mining. According to Mendenhall et al. the term "Outliers" is "that lies very far from the middle of the distribution in either direction". Outlier detection must be performed during the preprocessing step for locating whether the data pre-

S. Visalakshi (✉) · V. Radha
Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore 641042, India
e-mail: visaraji@gmail.com

V. Radha
e-mail: radhasrimail@gmail.com

L.C. Jain et al. (eds.), *Computational Intelligence in Data Mining - Volume 1*,
Smart Innovation, Systems and Technologies 31, DOI 10.1007/978-81-322-2205-7_29

sented in the drinking water dataset are normal or abnormal. It has become an active research issue in data mining, which has important applications in the field of medical care, public safety and security, image processing, sensor/video network related surveillance, intrusion detection, monitoring criminal activities in e-commerce, monitoring water quality etc.

In my research, the main goal is to identify the contamination from the drinking water dataset. The basic event detection framework utilizes a data mining algorithm for studying the interactions between multivariate water quality parameters and detecting possible outliers. The classifier SVM is used for studying the interplay between multivariate water quality parameters and detecting possible outliers. The SVM make use of several models to detect complex outliers based on the characteristics of the support vectors obtained from SVM-models. SVM is an iterative approach and remove severe outliers in the first iteration itself. So from the next iteration it starts to learn from "cleaner" data and thus reveals outliers that were masked in the initial models [1].

The outlier detection method can be divided into uni-variant and multi-variant. Uni-variate means, considering only one variable or one parameter and multi-variate means, considering more than one variable and check for the relationship between variables. Uni-variate outlier is easy for detection and correction, but multi-variate are more difficult to detect and consumes more time for detection. Another method of outliers is parametric and non-parametric. A parametric method uses statistical models and non-parametric uses some outlier detection methods which are distance based, clustering based and spatial based.

According to [2, 3], the existing method for detecting outliers is classified based on the availability of labels present in training data sets and it is categorized namely: Supervised, Semi-Supervised and Unsupervised. In principle, models belonging to supervised or semi-supervised approaches, all the data must be trained before use, while in unsupervised approach training is not required. Additionally, in the supervised approach, training set should be provided with labels for anomalous or normal. In contrast, in the training set with normal object, labels alone are needed for semi-supervised approach. In other words, the unsupervised approach does not require any object label information. In the paper [4] the classification model treats outlier detection, classification and feature selection as separate step. Thus the proposed method combines all three methods.

In this research work, the outliers are detected before going for feature selection. Clustering is used to partition data into large or small clusters. Based on the classification, the small cluster is considered as outliers and removed safely from the dataset. The main goal of this paper is to remove the outlier effectively. For the experiment, the real time drinking water dataset is used for evaluation. Section 2 analyses the concepts of outlier detection and clustering. In Sect. 3 the proposed techniques are explained in detail. Section 4 reports the experimental results of enhanced technique. Finally conclusion is presented in Sect. 5.

2 Outlier Detection (OD) and Clustering

Clustering analysis and Outlier Detection are two related tasks which will go hand in hand. Clustering finds the major patterns in a dataset and sorting will be performed according to the data, whereas outlier detection aims at capturing the exceptional cases which deviate from the majority patterns. Taking binary decision on whether the object present in the dataset is an outlier or not are becoming a challenging task for the real time dataset. Some of the general terms used in clustering are Cluster (ordered list of objects, which will have common characteristics), Distance (calculating the distance between two points or two elements), Similarity (similarity between two documents SIMILAR (D_i, D_j), Average Similarity (similarity measure will be computed for all the documents (D_i, D_j), except i = j an average value will be obtained), Threshold (finding out the lowest possible input value of similarity which is required to join two objects in one cluster), Similarity Matrix (to find out the similarity between two objects, the similarity function SIMILAR (D_i, D_j) will be used and result will be represented in the form of matrix), Dissimilarity Co-efficient (distance between two clusters) and Cluster seed (first object or first point of cluster is defined as initiator of that cluster and this initiator is known as cluster seed) [1]. Clustering is an important and popular tool for outlier analysis. Most of the techniques presented in outlier will rely on the key assumption that normal objects belong to either large/dense clusters, while small clusters are considered as outliers [5, 6]. Many researchers argue that clustering algorithm is not an appropriate choice for outlier detection. There is no single and specific algorithm for detecting outlier. Therefore, many approaches have been proposed and existing algorithms are also enhanced to improve the outlier detection metrics.

The approaches are classified into four major categories, namely, Distance-based approach, Density-based approach, Distribution-based approach, Deviation-based and clustering-based approach. Distance-based approach, outliers are detected, based on the distance between two points. Density-based approach helps to find out non-linear shapes and structure based on the density. Distribution-based approach helps to detect clusters with arbitrary shape and it does not require any input parameters. It can also handle large amount of spatial data. Clustering-based approach considers small sizes of clusters as outliers. Deviation-based approach helps to identify outliers which deviate from the selected objects [4, 7, 8].

In this paper, enhancement of k-means is proposed in the first phase and the second phase analyses detection of outliers. The main focus of this work is to identify outlier using distance-based clustering, which results in discovering normal and abnormal clusters. Classic k-means algorithm is very sensitive in nature. The selection of initial cluster prototypes will converge to suboptimal solution, only when the initial prototypes are chosen properly and the value of k must be specified

in advance. While solving the real-world applications fixing the value of k will be difficult. For such real-time applications [8], first-width clustering algorithm is used to perform clustering process, and classifies outliers as erroneous value or interesting event. The main disadvantage of this algorithm is that it will not consider the intra clusters. The width must be specified before processing starts. For high-dimensional real time data, this algorithm will not work. The required number of distance calculation is high. The selection of centroid is important and if it is not initialized correctly, the classified cluster might have outlier. To overcome this, the Bayesian Information Criterion is included with Modified Dynamic Validity Index (MDVI). Generally, the k-means clustering is used to cluster and classify normal and abnormal clusters. The process of computation is high when k-means is applied to high dimensional dataset. To overcome this issue, a different distance calculation mechanism is applied. In the proposed algorithm all issues are handled and solved. The k-means is enhanced and there is no need to specify the k value; it is automatically assigned to the variable k and the centroid seed selection is estimated automatically. Proposed research is discussed in the following section.

3 Proposed Work

One of the best top ten algorithms in data mining is k-means, which is simple and scalable in nature. Clustering algorithm partitions the dataset into k clusters and it has the two main objectives of making each cluster as compact as much as possible [9]. This paper proposes a new cost function, and distance measure, based on the values present in the dataset. The traditional k-mean algorithm divides the data set X into k clusters and calculates the centroid of each cluster. k value must be assigned before the clustering process. The distance must be calculated with each instance and each instance is to be assigned to the cluster with the nearest seed. Finally threshold % for each cluster, and the distance between each point of cluster from centroid are calculated. When the distance is greater than the threshold value it will considered as "outlier" [9, 10]. The main objective of this paper is to enhance the k-mean clustering algorithm, handling the issues of first width clustering algorithm and evaluating the proposed algorithm with the real time application.

In this work, k-means clustering algorithm is used to cluster the dataset into k clusters. In the proposed algorithm the k-value is generated automatically by using enhanced Bayesian Information Criterion (BIC) along with Modified Dynamic Validity Index (MDVI). The Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) or Deviation Information Criterion (DIC) are used to determine the number of clusters and the distance metric used for this work is Euclidean distance. The distance from centroid to cluster will be processed until convergence is achieved. The below algorithm describes the process and steps to perform k-mean clustering.

Input: Dataset D with x_i (i = 1…n) data points **Output**: Clusters (C_1, …, C_k)

Step 1: Feature selection is applied.
Step 2: Estimation of K is automatic and the procedure is given below:

I. Pre-cluster real time dataset (drinking water dataset) using Birch algorithm.
II. BIC is computed for each cluster using Eq. (1), where J is a cluster.

$$BIC(J) = -2\sum_{j=1}^{J} \xi_j + m_J \log(N). \tag{1}$$

III. The ratio of change in BIC at each successive merging relative to the first merging determines the initial estimate and is calculated using Eq. (2).

$$dBIC(J) = BIC(J) - BIC(J+1). \tag{2}$$

From these initial estimates the change ratio of the J cluster is calculated using Eq. (3) as the ratio of dBIC(J) to the dBIC(1) of the first cluster.

$$R_1(J) = \frac{dBIC(J)}{dBIC(1)}. \tag{3}$$

If dBIC(1) < 0, then K_T = 1 and go to step 8 else calculate inter-cluster ratio and K_T = number of cluster for which the recorded ratio is minimum of all and repeat steps V, VI and VII for all K_T.

IV. Calculate modified inter and intra cluster ratio between cluster C_k and C_{k+1} using Eq. (4).

$$\text{IntraRatio(k)} = \frac{\text{Intra(k)}}{\text{MaxIntra}} \text{InterRatio(k)} = \frac{\text{Inter(k)}}{\text{MaxInter}}. \tag{4}$$

K is the pre-defined upper bound number of the clusters.

V. Calculate the modified dynamic validity index using Eq. (5).

$$MDVI = \min_{k = 1,\ldots k} \{\text{IntraRatio(k)} + \gamma * \text{InterRatio(k)}\}. \tag{5}$$

VI. K_T = Number of clusters for which the dynamic validity index is maximum Optimal k = K_T.

Step 3: Estimation of K initial seeds (C_j) is automatically generated.

I. Calculate K.

II. Compute the distances between objects in D is calculated using Eq. (6).

$$N(X_i) = \{any\, x_j : d(x_i, x_j) < x_j, D, i, j\}. \tag{6}$$

where $d(x_i, x_j)$ is the distance between x_i and x_j calculated using DLG and the average distance between all objects is calculated using the following equation.

III. Compute the average distance between all objects using Eq. (7).

$$\varepsilon = \frac{1}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d(x_i, x_j). \tag{7}$$

IV. Find neighborhood of objects in D
The coupling degree between neighborhoods of objects x_i, x_j is the ratio of number of objects neighbor to both x_i and x_j is calculated using Eq. (8).

$$\text{Coupling}(\mathrm{N}(x_i), \mathrm{N}(x_j)) = \frac{|N(x_i) \cap N(x_j)|}{N(x_i) \cap N(x_j)}. \tag{8}$$

V. If Coupling ($N(x_i)$,$N(x_j)$) < ε (average distance between all objects), then next centroid is found

$$\text{Add to C(Next Centroid)}$$

VI. If |No of centroids| < k, Go to Step 6, otherwise go to Step 9.

VII. End

Step 4: Steps 5–9 for each point x_i in D′ are repeated.

Step 5: Distance between each data point x_i and all k cluster centre is calculated using Eq. (9).

$$\mathrm{DLG}_{ij} = \min\left(\sum_{e=1}^{p} \mathrm{L}(p_e, p_{e+1})\right). \tag{9}$$

where **e** ∈ **E** and ranges from 1…p. Thus, DLG_{ij} satisfies the four conditions for a metric, that is, $D_{ij} = D_{ji}$; D_{ij} 0; D_{ij} $D_{ie} + D_{ej}$ for all x_i, x_j, x_p and $D_{ij} = 0$ iff $x_i = x_j$. Thus the new measure considers both global and local consistency and can adapt itself to the data structure. L the density length is computed using Eq. (10).

$$L(x_i, x_j) = \rho^{dist(x_i, x_j)} - 1. \quad (10)$$

where dist(x_i, x_j) is the Euclidean distance between x_i and x_j and >1 is the flexing factor (the value 8 is used during experimentation).

Step 6: The closest centre c_j is found and assigned x_i to cluster j.

Step 7: Label of cluster centre j along with the distance of x_i to c_j and stored in array Cluster[] and Dist[] respectively

Set Cluster[i] = j (j is the nearest cluster)
Set Dist[i] = DLG_{ij} (distance between x_i to closest cluster centre c_j)

Step 8: Cluster centres are recalculated.

Step 9: DLG_{new} of x_i is computed to new cluster centres until convergence

If DLG_{new} is less than or equal to DLG_{ij}, then x_i belongs to the same cluster j
else
DLG is computed with every other cluster centre and assign x_i to the cluster whose DLG is minimum
Set Cluster[i] = j and Set Distance[i] = DLG_{new}

Step 10: Output clustered results.

Features selection chooses distinctive features from a set of dataset. Selection of features helps to reduce the size of dataset and make the process simpler for all subsequent design [11].

Once the feature selection is over, pre-cluster the dataset using Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) algorithm which can handle large datasets. The ratio of change in cluster is calculated in BIC at each consecutive merging. The intra and inter cluster relationship are evaluated using the formula. To improve the performance of algorithm and to find the better cluster number the MDVI is used. Next process of algorithm is to select the initial seed selection for the assignment of data to cluster. The performance of initial seed selection will be based on the sum of square, difference between members of cluster, cluster center and normalized data size [8, 9]. The inter and intra cluster distance validity measure allows to determine the number of clusters automatically. For initial centroid selection, enhanced distance measure, Reverse Neighbor Node and coupling degree are used. The coupling degree is used to measure the similarity between two objects.

Thus conventional k-means algorithm begins with a decision on the value of k. Any initial partition which classifies the data into k clusters is assigned. The problem of Euclidean distance is overcome in the proposed work. Hence the Euclidean distance of both x_i and x_j is calculated with the threshold value of 8 is used for experiment. Thus the new measure considers both global and local consistency and can adapt itself to the data structure. The distance from centroid to cluster will process until convergence is achieved. The above algorithm outlines the

process of automatic key generation of k value and initial seed selection for clustering.

3.1 Outlier Detection (OD)

The second phase of the work is outlier detection. Detection of OD is basic issue in data mining. Outlier detection will remove 'noise' or 'unwanted' data from the dataset. Once the cluster formation has taken place with the help of enhanced k-mean clustering, then the output will be given as input for outlier detection. The dataset is partitioned into small and large clusters and the resultant cluster will be checked for anomalies, and if anomalies are present then those anomalies are safely removed from the whole dataset. For each cluster c_i in the cluster set C, a set of inter cluster distances $Dc_i = \{d(c_i, c_j): j = 1\ldots (|C| - 1), j = i\}$ is computed. Here, $d(c_i, c_j)$ is the Euclidean distance between centroids of c_i and c_j, and $|C|$ is the number of clusters in the cluster set C. Among the set of inter-cluster distances Dc_i for cluster c_i, the shortest K (parameter of KNN) distances are selected and using those, the average inter-cluster distance ICD_i of cluster c_i is computed using Eq. (11).

$$\mathrm{ICD_i} = \left\{ \begin{array}{ll} \frac{1}{K} \sum_{j=1,\neq i}^{K} d\left(c_i, c_j\right) & K \leq |C| - 1 \\ \frac{1}{|C|-1} \sum_{j=1,\neq i}^{|C|-1} d\left(c_i, c_j\right) & K > |C| - 1 \end{array} \right\}. \tag{11}$$

The average inter-cluster distance computation is enhanced. Instead of using the whole cluster set C to compute the average inter-cluster distance ICD_i for a cluster c_i, the presented algorithm uses the K-Nearest Neighbor (KNN) for cluster c_i. The advantage of this approach is that clusters at the edge of a dense region are not considered compared to clusters in the centre of the region. A cluster is identified as anomalous Ca $\subset$ C are defined as $C_a = \{c_i \in |C|\ ICD_i > AVG\ (ICD) + SD\ (ICD)\}$, where ICD is the set of average inter-cluster distances.

Once the anomaly is identified and removed then the two clusters are merged into single cluster, if it satisfies the rules which are given below. A pair of clusters c1 and c2 are similar if the inter-cluster distance d(c1, c2) between their centers is less than the width w. If c1 and c2 are similar, then a new cluster c3 is produced. The centre of c3 is the mean of the centers of c1 and c2 and whose number of data vectors is the sum of those in c1 and c2. In the proposed system, the merging procedure compares each cluster c_i with clusters $\{c_{i+1}, c_{i+2},\ldots\}$, and merges c_i with the first cluster c_j such that $d(c_i, c_j) < T$ and $j > i$. The value of T is set to 0.38 after experimentation with different values ranging between 0.01 and 2.02 in steps of 4. For detecting outlier, the k-means clustering algorithm is enhanced and KNN is used to find out the nearest neighbor. The DLG is used to consider both intra and inter distance between data points. For automatic estimation, BIC is enhanced using

MDVI. The centroid selection enhanced distance measure is combined with RNN and coupling degree. Finally, for dimensionality reduction, Principal Component Analysis is used. This section concludes that outliers are detected effectively with the help of the proposed technique. The enhanced k-means clustering will be able to discover clusters with correct arbitrary shape, it works well for large databases efficiently and lot of heuristics to determine the parameters. The enhanced technique helps to detect outliers efficiently and accurately.

4 Experimental Results

Clustering-based method determine cluster with shape, efficient to handle large database and determines the number of input parameters. Based on clustering, the exception will be considered as "noise", where it is bearable in some cases and sometimes that leads to inaccurate results. The two different season (summer and winter) real time datasets are collected from TWAD Board, Coimbatore, Tamil Nadu, India. The dataset contains various parameters which brief about the characteristics of drinking water. Some of the parameter used for research are Turbidity, EC, TDS, Ph, Ca etc. The SVM and BPNN classifiers are used in this experiment for training and testing. Enhanced k-means and outlier detection are evaluated with various metrics namely, Accuracy, Normalized Root Mean Square Error and Speed.

$$\boldsymbol{Accuracy} = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \times 100$$

$$\mathbf{NRMSE} = \frac{\sqrt{\boldsymbol{mean}[(\boldsymbol{y}_{\boldsymbol{true}} - \boldsymbol{y}_{\boldsymbol{imp}})^2]}}{\boldsymbol{variance}\, \boldsymbol{y}_{\boldsymbol{true}}}$$

Table 1 presents the classification accuracy of before and after OD to various % of outliers for two different seasons. It is obvious that the classification accuracy is improved compared with BPNN.

The normalized root means square error before and after outlier detection values are presented in Table 2 for two different seasons and it is evident that the normalized root mean square error value is minimized compared with BPNN.

The execution speed of before and after outlier detection is listed in the above Table 3 for summer and winter seasons and it is clear that the execution speed is reduced compared with BPNN. Thus the experimental results shows that proposed algorithm works effectively for detecting outlier in contaminated drinking water dataset.

Table 1 Classification accuracy (%)

Datasets	Outliers	BPNN before OD	BPNN after OD	SVM before OD	SVM after OD
Summer	0	79.80	81.23	83.79	84.66
	10	80.52	85.34	88.92	90.45
	20	81.27	84.34	87.17	87.46
	30	84.12	85.64	86.22	87.61
	40	80.97	82.34	84.97	85.93
Winter	0	79.83	81.76	82.87	84.16
	10	81.73	83.45	89.51	91.49
	20	82.24	85.54	86.42	87.17
	30	80.12	83.45	85.61	86.00
	40	79.34	80.16	84.73	85.72

Table 2 Normalized root mean square error

Datasets	Outliers	BPNN before OD	BPNN after OD	SVM before OD	SVM after OD
Summer	0	0.8976	0.8765	0.4451	0.4389
	10	0.7921	0.7832	0.4049	0.3990
	20	0.8012	0.7986	0.4103	0.4035
	30	0.8123	0.7989	0.4282	0.4213
	40	0.9013	0.8967	0.4251	0.4190
Winter	0	0.8876	0.8675	0.4726	0.4664
	10	0.8012	0.7954	0.4410	0.4348
	20	0.8234	0.8123	0.4573	0.4512
	30	0.8100	0.8024	0.4415	0.4350
	40	0.9876	0.9765	0.4548	0.4486

Table 3 Execution speed (seconds)

Datasets	Outliers	BPNN before OD	BPNN after OD	SVM before OD	SVM after OD
Summer	0	6.69	6.11	5.13	5.07
	10	6.76	6.34	5.23	5.17
	20	7.23	7.05	5.41	5.35
	30	6.98	6.87	6.28	6.24
	40	7.45	7.25	7.12	7.07
Winter	0	6.89	6.50	6.37	6.32
	10	6.54	6.50	6.41	6.36
	20	7.15	7.10	7.05	6.98
	30	9.05	8.79	8.26	8.22
	40	9.34	9.10	8.97	8.93

5 Conclusion

This paper focused to detect outlier from dataset and aims to find objects which are different or contradictory from other data. An outlier detection method is proposed. The issues in k-means clustering algorithm are handled to enhance its clustering operations and to identify the outliers from the dataset are proposed. The first step is grouping of data into a number of clusters for this the average of inter cluster distance is used. Next outlier is identified from the resultant cluster. The experimental results show that the classification accuracy is improved and normalized root mean square error is minimized and it is evident that the proposed algorithm is efficient in identifying outliers to detect the contamination quickly. In future, Feature Selection algorithm can be combined with outlier detection to improve contamination detection.

Acknowledgments The authors express their gratitude to TWAD Board for their whole hearted support in providing dataset for research.

The author expresses their gratitude to Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, Tamil Nadu, India for the progress of research work.

References

1. Cateni, S., Colla, V., Vannucci, M.: Outlier detection methods for industrial applications. Advances in robotics. In: Automation and Control, pp. 274–275 (2008)
2. Ahmad, A., Dey, L.: A k-mean clustering algorithm for mixed numeric and categorical data. Data Knowl. Eng. **63**, 502–527 (2007)
3. Hodge, V.J., Austin, J.: A survey of outlier detection methodologies. Artif. Intell. Rev. **22**(2), 85–126 (2004)
4. Fawzy, A., Mokhtar, H.M.O., Hegazy, O.: Outliers detection and classification in wireless sensor networks. Egypt. Inf. J. **14**, 157–164 (2013)
5. Khan, F.: An initial seed selection algorithm for k-means clustering of geo-referenced data to improve replicability of cluster assignments for mapping application. Appl. Soft Comput. **12**, 3698–3700 (2012)
6. http://members.tripod.com/asim_saeed/paper.htm
7. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: a survey. ACM Comput. Surv. **41** (3), 1–58 (2009)
8. Pachgade, S.D., Dhande, S.S.: Outlier detection over data set using cluster-based and distance-based approach. Int. J. Adv. Res. Comput. Sci. Soft. Eng. **2**(6), 12–16 (2012)
9. Zhu, C., Kitagawa, H., Papadimitriou, S., Faloutsos, C.: Outlier detection by example. J. Intell. Inf. Syst. **36**, 217–247 (2011)
10. Shi, Y., Zhang, L.: COID: a cluster–outlier iterative detection approach to multi-dimensional data analysis. Knowl. Inf. Syst. **28**, 710–733 (2010)
11. Indira Priya, P., Ghosh, D.K.: A survey on different clustering algorithms in data mining techniques. Int. J. Mod. Eng. Res. **3**(1), 267–274 (2013)
12. Zhang, T., Ramakrishnan, R., Livny, M.: BIRCH: a new data clustering algorithm and its applications. Data Min. Knowl. Discov. **1**, 141–182 (1997)