IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-4, April 2019 E-ISSN: 2347-2693

Multi-objective Optimization to Detect Outliers with Referential Point using Evolutionary Clustering Techniques

M. Anusha

Department of Computer Science, National College, Trichy, India

Corresponding Author: anusha260505@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7i4.731735 | Available online at: www.ijcseonline.org

Accepted: 16/Apr/2019, Published: 30/Apr/2019

***Abstract*—** Many real-world problems have multiple competing objectives and can often be formulated as multi-objective optimisation problems. Multi-objective evolutionary algorithms have proven very effective in obtaining a set of trade-off solutions for such problems. This research seeks Outliers detection is perceptibly different from or inconsistent with the remaining dataset is a major challenge in real-world multi-objective problem. In this paper, the problem of identifying deviation point in a data set that exhibit non-standard behaviour is referred to as outlier. Outlier detection turns out to be a challenging task due to insufficient data in finding features to describe absolute high data. This paper presents a reference point based outlier detection algorithm using multi-objective evolutionary clustering technique(MOODA). In this algorithm, it assigns a deviation degree to each data point using the sum of distances between referential points to detect distant subspaces where outliers may exist. Finally, experimental studies show that our proposed algorithm is more effective in terms of efficiency and accuracy by using UCI dataset.

***Keywords*—**Outlier detection,Clustering, Multi-objective optimization, Evolutionary algorithrms.

I. INTRODUCTION

Outlier detection is a challenging task in data mining to detect or identify data points that differs very greatly from rest of the data in a dataset. Outlier analysis searches for such highly deviating objects in contrast to regular objects. An outlier has highly deviating attribute values compared to its local neighbourhood [1]. The importance of outlier detection technique is to generate the data or sign fraudulent activity inside the data or alter data into significant information, which is used in wide variety of applications, like medical diagnosis, fraud detection, intrusion detection, marketing systems, astronomical spectroscopy and so on.

The algorithms such as statistical-based, clustering-based, distance-based, and subspace-based techniques [2, 3] were proposed to detect outliers. In statistical-based methods, it assumes that data follow a standard distribution, and it detects outliers by identifying objects that deviate from the distribution. The drawback of this method is that we will not always have a priori understanding regarding the underlying distribution of the datasets, especially for high-dimensional datasets [4, 5].

Outlier detection is very significant since most of real-world problem which change over time. The problem typically involves multiple objectives, such as noise, decision making and partition density. It is well known that evolutionary algorithms (EAs) are very suitable for multi-objective optimization problems [6] since they can generate a set of Pareto solutions in one run. Although there are researches on using EAs to solve outlier detection problems [7], there exist no studies on using multi-objective EAs for outlier detection problem (MOODP).In this paper, we propose a reference point based outlier detection algorithm (MOODA). The deviation point and threshold are chosen as the accuracy objectives and the improved cluster as the objective to evaluate the cluster compactness. Main contributions of the paper is to propose a multi-objective EA based approach for outlier detection based on distance to handle massive and high-dimensional dataset efficiently. In this method, it detects outliers by calculating the distance of point to its k-nearest neighbour set and then by using reference point deviating points where outliers may exist and then finally compares deviating point of data point to a threshold value to detect outliers. The rest of the paper is structured as follows. Section II, surveys the existing work related to outlier mining techniques. And in Section III, a referential point–based outlier detection technique is presented. Section IV presents experimental studies that describe the performance of the results on various datasets taken from UCI repository. And the conclusion is given in Section V.

II. RELATED WORK

At present there are many outlier detection algorithms have been proposed by researchers. The outlier detection algorithms can be categorized into following techniques: density-based, distance-based, distribution-based, and clustering-based outlier detection techniques [8, 9]. In density-based outlier detection technique, it detects an outlier if its local density differs from its neighbourhood. After that, so many variations of LOF algorithm have been projected. In local distance-based outlier factor [10] algorithm, it discovers outliers in scattered datasets by using the relative distance of an object to its neighbourhood, and in improving influenced outlierness [11] score, it considers both neighbours and reverse neighbours of an object while calculating its relative density distribution. Anusha et al. [12-15] focused on centroid based multi-objective clustering that lacks to reduce the number of clusters. Various feature selection and neighbourhood learning techniques are explained in [16, 17].Müller et al. [18] projected, a connectivity-based outlier factor (COF), with considering underlying patterns of the data detect outlier using the relative distribution through distance.

Tang et al. [19] proposed a distance-based approach for outlier detection which is based on the well-known nearest neighbour algorithm; it detects outliers by calculating distance of an object with its neighbours. The distances among objects need to be calculated in either raw data space or feature subspace. In distribution-based approach, it considers an object as an outlier if deviates too much from a standard distribution. And in cluster-based approach [20], it considers an object as an outlier if it does not belong to any cluster. Clustering techniques in detect outliers in large attributed graphs. However, clustering-based approaches must build a clustering model, which limits the outlier detection performance. Various feature selection and neighbourhood learning techniques are explained in [21].

Statistical methods [22, 23] attempt to fit the distributions on the training data. Then any data which has low probability under this distribution will be determined as an outlier. However, this kind of methods depends on the choice of distribution. An unsuitable distribution may result in a bad detection performance. Neighbour-based methods assume that normal data has relatively more neighbours than the outlier data [24, 25]. In [24], Breunig et al. adopted a density-based local outlier factor (LOF) to address this issue. In addition, by normalizing LOF, Kriegel et al. [25] propose the local outlier probabilities to detect outliers. However, the search for the nearest neighbours prohibits such methods to be applied to high-dimensional data due to the curse of dimensionality.

In [26], Azami et al. converted the OCSVM scores into the outlier probabilities. Quinn et al. [27] attempted to find the boundary based on a squared-loss function which is similar to OCSVM method. However, this kind of methods needs to calculate the kernels, which has high computational complexity. Thus, kernel-based methods are both time-consuming in training and testing. Reconstruction-based methods assume that there are strong correlations in the features of the normal data. By minimizing the reconstruction errors of the normal training data, the reconstruction-based methods can capture these strong correlations. Therefore, the normal test data will have relatively smaller reconstruction errors, while outliers will have larger reconstruction errors.

III. A REFERENTIAL POINT–BASED OUTLIER DETECTION

The referential point based [28] outlier detection algorithm is proposed for efficient outlier detection in high dimensional data. Let the input to the proposed algorithm be an n-dimensional data set M consisting of k data objects with the topographies T = {t1, t2, ..., tn}. The objective is to determine a suitable feature subset Ts ⊂ T using reference points to detect the outliers present in the input data in an efficient manner. Accordingly, the outlier detection method employed here consists of two major tasks, the first task is to apply the closeset reference point on the input data to determine closest neighbour for identifying a relevant subset of topographies Ts. The next task is to apply the proposed outlier detection algorithm on this low-dimensional data for establishing the efficacy of the selected subset of topographies. Point of deviation is measured in the subspaces. Let us assume that given a population P that includes O entities and a number of elements. The deviation of each data point in a dataset is calculated by threshold measure to which the point diverges from the rest of the data point in the same subspace and point is called deviation point (DP) in subspace s. DP is s the sum of the distances between a data point and its k-nearest neighbour set in a data subspace. Mathematically, the deviation point p in subspace s is calculated as:

$$DP(p)=\sum_{1=1}^{n} dist(p,p_i) \quad (1)$$

where pi is closest neighbour to the set of point p in subspace s. And distance between any two points can be calculated as:

$$Dist(px,py)=\sqrt{\sum((px_i-py_i)/(max_i-min_i))2} \quad (2)$$

In Equation. (2), Max_i and Min_i denote maximum and minimum values of data point in $_i$th dimension. The algorithm for calculating a deviation point of data point in a subspace, which is based on distance, is described in Algorithm 1. Algorithm starts by categorizing the objects with its corresponding entities in the population P. then the closest neighbor is identified for each data point and the

deviation point for the object is identified using threshold value. Based on the threshold, it calculates the deviation point of each point as sum of the distance between data point and its k-nearest neighbor set. After getting the outlying degree of each point, it stores in DP subspace.

The algorithm for outlier detection using reference point is illustrated in Algorithm 1, where it takes two inputs—first is DP and threshold value β, which is calculated as:

Algorithm 1 Multi-objective Optimization to Detect Outliers with Referential Point using Evolutionary Clustering Techniques

Input: Population P, subspace s
Output: deviation point DP

1. O=|P|
2. For i=1;i<=P;i++
3. For j=1;i<=O;i++
4. If (ref_pt)<=th
5. Break;
6. Pij= compute (D[i],[j])
7. OD (pij)= ∑dist(p,pi)
   1=1toO
8. OD-Mat= OD (pij)
9. End if
10. End for
11. End for
12. Return DP as result

$$\beta=\mu * \sqrt{\sum}(DPT)2 \quad (3)$$

where T⊆ P and

$$OD_T=1/O * \sum DP_T(\beta(j)) \quad (4)$$
$$j=1\text{-}m$$

IV. RESULTS AND DISCUSSION

To evaluate the performance and efficiency of the proposed algorithm MOODA, the experiments are conducted using personal computer which uses Windows 7 as operating system. The MOODA is implemented using MATLAB 7.0. We analyzed the proposed algorithm using various real life datasets. Cluster validity index called Shiloutee index is used to validate the result. Shiloutee index value lies between the interval [-1 1]. A value close to 1 means the cluster objects are similar and 0 means dissimilar clusters such that the objects lie far from the clusters, while -1 indicates that the sample are misclassified.

A. Datasets

Four real-life data sets are used for experiments. A short description of the data sets in term of size, dimension and number of clusters is provided in Table 1.

Table 1. Description of Real-life data sets.

Data sets	*Size of the data sets*	*Number of dimensions*	*Number of clusters*
Ionosphere	351	33	2
Iris	150	4	3
Wine	178	13	3
Seed	210	7	3

The real-life data sets are obtained from UCI Machine Learning Repository

B. Parameter Setting

The number of clusters parameter is fixed for the particular data sets. For the proposed algorithm, the crossover rate is 0.95, mutation rate is 0.01 and population size is 200.

C. Detection of Outlier Accuracy

The Table 2. shows the performance metric values obtained by Shiloutee index for the four data sets respectively. It is proved from the Table 2 the proposed Multi-objective Optimization to Detect Outliers with Referential Point using Evolutionary Clustering Techniques is performing well.

Table 2. Performance Metric Value for Real-life Data sets.escription of Real- life data sets.

Data sets	*No. dimensions*	*No.Outliers*	*Shiloutee Index*
Ionosphere	33	5	0.11
Iris	4	5	0.08
Wine	13	10	0.25
Seed	7	48	0.18

In order to evaluate the proposed algorithm, it is necessary to define a measure of agreement between two partitions of same data sets. Table 3. shows the results obtained from NLMOGA ,FS-NLMOGA and MOODA. The quality of the cluster is evaluated using Silhoutee index. From the result, it is certain that clustering accuracy is less than FS-NLMOGA and NLMOGA where result produces good deviating result for MOODA except the result obtained for wine data set. This is because of reference point based feature selection is applied in input section. The Fig 1, shows the outlier accuracy between three algorithms. Hence, we can conclude that MOODA is more efficient than NLMOGA and FS-NLMOGA for feature selected clustering problem. The algorithm can also accomplish high dimensional data set.

Table 3. The Results of Clustering Accuracies of NLMOGA, FS-NLMOG, MOODA using Silhoutee index

Data sets	*NLMOGA*	*FS-NLMOGA*	*MOODA*
Ionosphere	0.14	0.96	0.11
Iris	0.16	0.82	0.08
Wine	0.15	0.86	0.25

Data sets	*NLMOGA*	*FS-NLMOGA*	*MOODA*
Seed	0.36	0.74	0.18

D. Performance Analysis for Outlier Detection

The Table 3 shows the deviation point for the four real-life data sets respectively. Since, the proposed MOODA uses reference point based selection, it is proved that proposed algorithm is performing well. Fig.1.showsthe cluster compact with outlier between the cluster classes of the real-life data sets

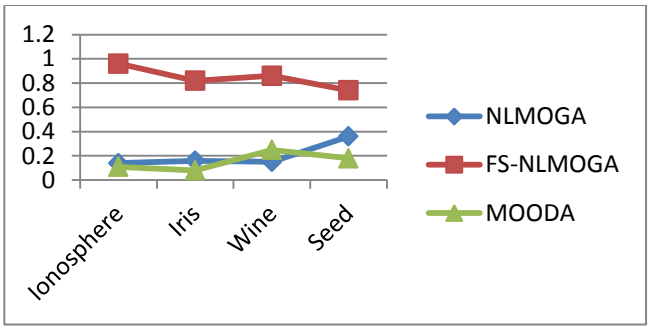

Fig. 1. Outlier detection of the real-life data sets

E. Pareto front for data sets

The Pareto fronts obtained by proposed algorithm are represented in Fig.1. It is inferred that the proposed algorithm works well for satisfying the main objectives considered in this paper. From the Fig.1, it is proved that the MOODA provides maximum deviation accuracy simultaneously. The algorithm shows a better result for the data sets Iris, Seed, and Ionosphere when compared with Wine

V. CONCLUSION AND FUTURE SCOPE

This paper proposes a reference point based outlier detection algorithm for multi-objective optimization problem. In contrast to conventional multi-objective genetic algorithm, MOODA maximizes deviation points between clusters simultaneously. These functions are optimized simultaneously using feature selected criterion. In this proposed algorithm, the deviation point calculation is based on the deviation point to detect outliers by comparing with the threshold value. The outlier shows high accuracy than NLMOGA and FS-NLMOGA. The performance of the proposed algorithm is tested with several real-life benchmark UCI repository data sets. The result indicates that the algorithm can simultaneously optimize the chosen objectives by minimizing the intra-cluster distance and maximizing the inter-cluster distance with high accuracy with chosen deviation point.

REFERENCES

[1] M.H.Marghny and A.I. Taloba, "*Outlier Detection using Improved Genetic K-means".* 2011. International Journal of Computer Applications. Vol .**28**, Issue .**11**, pp.**33-36**, **2011**.

[2] R.Baklouti, M. Mansouri, M. Nounou, H. Nounou, A.B. Hamida, "*Iterated robust kernel fuzzy principal component analysis and application to fault detection*", J. Comput. Sci. Vol. **15**, pp. **34–49 2016**.

[3] W.D. Fisher, T.K. Camp, V.V. Krzhizhanovskaya, "*Anomaly detection in earth dam and levee passive seismic data using support vector machines and automatic feature selection*." J. Comput. Sci, Vol. **20**, pp.**143–153** ,**2017**.

[4] M. Sakurada, T. Yairi, Anomaly, "*detection using autoencoders with nonlinear dimensionality reduction*", in: Proceedings of the 2014 ACM on Machine Learning for Sensory Data Analysis (MLSDA 2014), pp. **4**, **2014**

[5] J. T. Andrews, E. J. Morton, L. D. Griffin, "*Detecting anomalous data using auto-encoders*", International Journal of Machine Learning and Computing, Vol. **6**,Issue.**1**,pp. **21-26**,**2016**.

[6] S. Wu and S. Wang, "*Information-theoretic outlier detection for large-scale categorical data*," IEEE Trans on Knowledge and Data Engineering (TKDE), Vol. **25**, Issue. **3**, pp. **589–602**, **2013**.

[7] V. Chandola, A. Banerjee, and V. Kumar, "*Anomaly detection for discrete sequences: A survey*," IEEE Trans on Knowledge and Data Engineering (TKDE), Vol. **24**, Issue. **5**, pp. **823–839**, **2012**.

[8] M. Hubert,P.J.Rousseeuw, P. Segaert, "*Multivariate functional outlier detection*.", Stat. Methods Appl. Vol. **24**,Issue. **2**, pp. **177–202**,**2015**.

[9] S. Hido, Y. Tsuboi, H. Kashima, M. Sugiyama, and T. Kanamori, "*Statistical outlier detection using direct density ratio estimation," Knowledge and Information Systems"*, Vol. **26**, Issue. **2**, pp. **309–336**, **2011**.

[10] B. Perozzi, L. Akoglu,P. Iglesias Sánchez,E.Müller, "*Focused clustering and outlier detection in large attributed graphs*.", In: Proceedings of the 2014 ACM International Conference on Knowledge Discovery And Data Mining(SIGKDD), pp. **1346–1355**, **2014**.

[11] H.P.Kriegel, P. Kroger, E. Schubert, A. Zimek, "*Outlier detection in arbitrarily oriented subspaces*.", In: Proceedings of the 2012 IEEE International Conference on Data Mining (ICDM), pp. **379–388**, **2012**.

[12] M.Anusha and J.G.R.Sathiaseelan, "*An Improved K-Means Genetic Algorithm for Multi-objective Optimization*", International Journal of Applied Engineering Research,pp. **228-231**, **2015**.

[13] M.Anusha and J.G.R.Sathiaseelan, "*An Empirical Study on Multi-Objective Genetic Algorithms using Clustering Techniques*", International Journal of Advanced Intelligence Paradigms. Vol. **8**, Issue. **3**, pp. **343-354**, **2016**.

[14] M.Anusha and J.G.R .Sathiaseelan, "*Feature Selection using K-Means Genetic Algorithm for Multi-objective Optimization*", Procedia Computer Science, Vol. **57**, pp. **1074-1080**,Elsevier B.V., Netherland, **2015**

[15] M.Anusha and J.G.R.Sathiaseelan, "*An Enhanced K-means Genetic Algorithms for Optimal Clustering*", In the Proceedings of the IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp.**580-584**, **2014**.

[16] M.Anusha and J.G.R. Sathiaseelan, "*Evolutionary Clustering Algorithm using Criterion-Knowledge-Ranking for Multi-objective Optimization*", Wireless Personal Communication, Springer, Vol.**94**, pp.**2009-2030**, Springer, USA. **2017**.

[17] M.Anusha and J.G.R. Sathiaseelan, "*Multi-objective Optimization Algorithm to the Analyses of Diabetes Disease Diagnosis*", International Journal Advanced Computer Science Application, Vol.**7**, pp.**485-488**, Thesai publishers, UK. **2016**.

[18] E.Müller,M.Schiffer,T.Seidl, "*Statistical selection of relevant subspace projections for outlier ranking*." In the Proceedings of 2011 IEEE International Conference on Data Engineering (ICDE),pp. **434–445**, April **2011**.

[19] J.Tang,Z. Chen, A. Fu, D. Cheung, "*Enhancing effectiveness of outlier detections for low density patterns*." In the Proceedings of Advances in Knowledge Discovery and Data Mining, pp. **535–548** ,**2002**.

[20] K. Zhang, M.Hutter, H, Jin, "*A new local distance-based outlier detection approach for scattered real-world data*.", In the Proceedings of Advances in Knowledge Discovery and Data Mining, pp. **813–822**, **2009**.

[21] F. Keller, E. Muller, K. Bohm, "*HiCS: high contrast subspaces for density-based outlier ranking*." In the Proceedings of 2012 IEEE International Conference on Data Engineering (ICDE), pp. **1037–1048**, **2012**.

[22] A. Dukkipati, D. Ghoshdastidar, J. Krishnan, "*Mixture modeling with compact support distributions for unsupervised learning*", in the Proceedings of 2016 IEEE International Joint Conference on Neural Networks(IJCNN), pp. **2706– 2713**, **2016**.

[23] E. Eskin, "*Anomaly detection over noisy data using learned probability distributions*",In Proceedings of the International Conference on Machine Learning, Citeseer, **2000**.

[24] H.-P. Kriegel, P. Kr¨oger, E. Schubert, A. Zimek, "*LoOP: local outlier probabilities*", in: Proceedings of the 2009 ACM Conference on Information and Knowledge Management, pp. **1649–1652**, **2009**.

[25] B Sch¨olkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, R. C. Williamson, "*Estimating the support of a high-dimensional distribution*", Neural Computation, Vol.**13**,Issue. **7**, pp.**1443–1471**, **2001**.

[26] M. El Azami, C. Lartizien, S. Canu, "*Converting svdd scores into probability estimates: Application to outlier detection*", Neurocomputing, Vol. **268**, pp. **64–75**, **2017**.

[27] J. A. Quinn, M. Sugiyama, "*A least-squares approach to anomaly detection in static and sequential data*", Pattern Recognition Letters, Vol. **40**,pp. **36–40**,**2014**.

[28] M.Anusha, "*Multi-Objective Optimization to Identify High Quality Clusters with Close Referential Point using Evolutionary Clustering Techniques*", Asian Journal of Computer Science and Technology, Vol.**7** Issue.**3**, pp. **68-71**, **2018**.

