A REVIEW ON OUTLIER DETECTION AND REMOVAL TECHNIQUES

R. S. Sonawane[1] & Prof. S. S. Banait[2]
[1] M.E. Student, [2] Associate Professor
[1] Department of computer Engineering,
[1] K. K. Wagh Institute of Engineering Education & Research, Nashik,
Savitribai Phule Pune University, Maharashtra, India.

Received: December 07, 2018 Accepted: January 07, 2019

ABSTRACT*: To improve the clustering accuracy it is necessary to remove the outliers from clusters and for this purpose Outlier detection is used. Outlier detection is an important information analysis function in its own right. Extended k means algorithm with outlier detection is proposed which focuses on controlling the number of outliers and also focuses on subspace clustering. The proposed method also identifies clusters embedded in subspaces of the original data space. For establishment and convergence an iterative procedure, we have design to optimize the objective function in proposed work. By performing numerical experiments on synthetic data and real data, thus it improves the effectiveness and efficiency in the proposed algorithm.*

Key Words: *Data mining, k-means clustering, outlier detection, Subspace clustering.*

I. Introduction

In data mining clustering is a technique of identifying similar data groups in a dataset. Data points in each group are comparatively more similar to the data points of that group than those of the other group. The goal of clustering is to identify groups or clusters from a set of objects. Data clustering is a unsupervised learning process. One of the most important task of data mining is to cluster the dataset in such a way that data points in one cluster is different than data points of another cluster [12].

Outlier is the observation point that is different from other observation point. Thus outlier detection is also very important task in data mining for data analysis [13]. The main objective of outlier detection is to detect abnormal data from dataset. Outlier detection technique is used for intrusion detection, fraud detection, medical and public health. Outlier detection has also been applied to other domain also like image processing, sensor networks, biology, speech recognition, etc. There are multiple clustering techniques are available like hierarchical method, partitioning method, grid based method. There are also some algorithms are used for high dimensional data like subspace clustering, projection technique or co-clustering technique. There are multiple outlier detection approaches like distance based approach, classification based approach, clustering based approach, and proximity based approach, statistical based approach.

In the past fifty- sixty years many clustering algorithm have been developed, but among these algorithms, k-means algorithm is one of the oldest and most widely used algorithm. But k-means algorithm also has multiple drawbacks and one of the drawback is that k-means algorithm is very sensitive to noisy data and outliers. This drawback motivated to use k-means algorithm with outlier detection and removal. The idea of this k-means with outlier removal algorithm is that introducing an additional cluster .This additional cluster contains all clusters. This algorithm partition the dataset into k+1 clusters ,which include k clusters and one additional cluster to store the data points which can not fit in to normal clusters.

II. LITURATURE WORK

In this literature work, various strategies and techniques of outlier detection are discussed.
To deal with the challenges imposed by outliers (abnormal data points), many learning algorithms have been proposed. For instance, the novel unsupervised approach for outlier detection, this novel unsupervised approach by using a modified clustering algorithm method is detect the outliers and removed these outliers from dataset for these purpose to improve the accuracy of an algorithm. This novel unsupervised approach for outlier detection used in multiple applications like intrusion detection, fraud detection, medical and public health [1]. Some soft subspace clustering algorithms was sensitive to some scaling parameters, thus a novel soft subspace clustering algorithm was proposed for objective function by using log transformed distance. Entropy weighting method was used for the drawback like sensitive to the scaling parameter. Many soft subspace clustering was proposed like log transformed entropy weighting k-means, EWKM

algorithm and locally adaptive clustering algorithm [2].For grouping high dimensional data, subspace clustering algorithm with automatic feature grouping was proposed. In this, to capture the feature groups a new component was introduced. And to group features of high dimensional data automatically new iterative process was defined [3].

To provide outlier detection and data clustering simultaneously, outlier removal clustering algorithm (ORC) was proposed. In ORC algorithm, there were two stages. In first stage, it consists of purely k-means process and in second stage it removes the outliers which are far from its cluster centers [4]. For identifying outlier cluster based local outlier factor technique was used. The new algorithm findCBLOF was proposed to identify outliers from the local data [5]. To calculate variable weights automatically, k-means type clustering algorithm was proposed. For calculate variable weights automatically new clustering process was introduced also a formula was proposed to calculate this weights. This method was used where large and complex real data used [6].

For the purpose of outlier detection, k-mode clustering technique was proposed. Some clustering algorithm was sensitive for the selection of initial cluster centers. Thus initialization of k-modes clustering algorithm was proposed to overcome the drawback of selection of initial cluster centers [7]. Outlier detection is very important task in data mining. Various outlier detection techniques like clustering based outlier detection methods were implemented to remove the outliers. This consists of two stages, 1. One - pass clustering algorithm was used to cluster dataset. 2. Outlier factor was used to determine outlier cluster [8].

An important problem in cluster analysis is the cluster structure which is defined by the selection of variable. Automated k-means clustering process was proposed with identification of outlier and selection of variable [9]. Multiple outlier detection method was implemented to detect outliers to improve accuracy of an algorithm like clustering based method which was proposed to identify outliers. The basic k-means algorithm was used to partition the data into multiple clusters. The point near to cluster centers were not outliers, and then calculate distance based outlier score for remaining points which far from cluster centers [10]. Noise clustering was defines as robust clustering method, which was proposed to create clusters of data set which reduce errors occurs due to outliers. Noise clustering method identifies outliers according to certain distance [11].

To cluster the numerical data and to detect the outliers from numerical data a three stage k-means algorithm was proposed. 1. In first stage the process of forming clusters was determined. 2. In second stage outliers were found from each cluster and the centroid was removed. 3. In third stage, the clusters, densities were similar and some of the part which overlapped was merged [14]. Detecting outliers is a task of detecting data points which are different or abnormal is very challenging task real world KDD applications. Some of the outlier detecting method was not applied to the scattered real-world dataset because of ineffectiveness of method. A novel local distance based outlier factor was proposed, which detect outliers in scattered real-world dataset [15]. A unified approach was discovered for simultaneously grouping and for finding outliers from data. This approach was formalized as generalization of k-means problem. This approach proved that the problem was NP- hard. This approach presents a practical polynomial time algorithm which was guaranteed to converge a local optimum [16]. After that this approach was extended to all distance measures that can be expressed in the form of a Bregman divergence. There was a topic on the valuation of guarantees embedded in single variable annuity contract. A novel approach was proposed to fill the gap which is based on data grouping and machine learning which afterword used for large portfolio of variable annuity contacts. Variable annuity (VA), also known as segregated fund, guaranteed investment fund, unit-linked life insurance, or equity-linked life insurance, was a very popular insurance product [17]. There was multiple data clustering method was proposed and also machine learning method was proposed.

III. CONCLUSION

In this work various outlier detection algorithm are analyzed. The use of this algorithm is to cluster data points as normal data points and outlier data points. To improve performance our proposed algorithm uses k+1 cluster (additional one cluster) which improves accuracy of an algorithm. The use of proposed algorithm reduces run time and removes outliers. KMOR algorithm is capable to cluster data points and detect outlier points at the same time.

IV. Acknowledgment

Authors would like to thank Prof. Dr. K. N. Nandurkar, Principal and Prof. Dr. S. S. Sane, Head of Department of Computer Engineering, K.K.W.I.E.E.R., Nashik for their kind support and suggestions. We would also like to extend our sincere thanks to all the faculty members of the department of computer engineering and colleagues for their help.

References

1. M. Ahmed , A. Naser , A novel approach for outlier detection and clustering improvement, in: Proceedings of the 8th IEEE Conference on Industrial Electronics and Applications (ICIEA), 2013, pp. 577–582 .
2. G. Gan , K. Chen , A soft subspace clustering algorithm with log-transformed distances, Big Data and Inf. Anal. 1 (1) (2016) 93–109 .
3. G. Gan , M.K.-P. Ng , Subspace clustering with automatic feature grouping, Pat- tern Recognit. 48 (11) (2015) 3703–3713 .
4. V. Hautamäki , S. Cherednichenko , I. Kärkkäinen , T. Kinnunen , P. Fränti , Improv- ing k-means by outlier removal, in: Proceedings of the 14th Scandinavian Con- ference on Image Analysis, SCIA'05, 2005, pp. 978–987
5. Z. He , X. Xu , S. Deng , Discovering cluster-based local outliers, Pattern Recognit. Lett. 24 (9–10) (2003) 1641–1650 .
6. J. Huang , M. Ng , H. Rong , Z. Li , Automated variable weighting in k -means type clustering, IEEE Trans. Pattern Anal. Mach. Intell. 27 (5) (2005) 657–668 .
7. F. Jiang , G. Liu , J. Du , Y. Sui , Initialization of k-modes clustering using outlier detection techniques, Inf. Sci. 332 (2016) 167–183 .
8. S.-Y. Jiang , Q. An , Clustering-based outlier detection method, in: Fifth Inter- national Conference on Fuzzy Systems and Knowledge Discovery, 2, 2008, pp. 429–433 .
9. S.-S. Kim , Variable selection and outlier detection for automated k-means clus- tering, Commun. Statis. Appl. Methods 22 (1) (2015) 55–67 .
10. R. Pamula , J. Deka , S. Nandi , An outlier detection method based on clustering, in: Second International Conference on Emerging Applications of Information Technology, 2011, pp. 253–256 .
11. F. Rehm , F. Klawonn , R. Kruse , A novel approach to noise clustering for outlier detection, Soft Comput. 11 (5) (2007) 4 89–4 94 .
12. F. Jiang , G. Liu , J. Du , Y. Sui , Initialization of k-modes clustering using outlier detection techniques, Inf. Sci. 332 (2016) 167–183 .
13. S. Chawla, A. Gionis, k -means: a unified approach to clustering and outlier detection, SIAM, pp. 189–197.
14. Y. Zhou , H. Yu , X. Cai , A novel k-means algorithm for clustering and outlier detection, in: Second International Conference on Future Information Technology and Management Engineering, 2009, pp. 476–480 .
15. K. Zhang , M. Hutter , H. Jin , A new local distance-based outlier detection approach for scattered real-world data, in: Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, PAKDD '09, 2009, pp. 813–822 .
16. S. Chawla, A. Gionis, k -means: a unified approach to clustering and outlier detection, SIAM, pp. 189–197.
17. G. Gan , Application of data clustering and machine learning in variable annuity valuation, Insurance: Math. Econ. 53.