Local Neighborhood-based Outlier Detection of High Dimensional Data using different Proximity Functions

Mujeeb Ur Rehman[1]
Department of CS
Khwaja Fareed University of Engineering and IT
Rahim Yar Khan - Pakistan

Dost Muhammad Khan[2]
Department of CS and IT
The Islamia University of Bahawalpur
Bahawalpur - Pakistan

***Abstract*—In recent times, dimension size has posed more challenges as compared to data size. The serious concern of high dimensional data is the curse of dimensionality and has ultimately caught the attention of data miners. Anomaly detection based on local neighborhood like local outlier factor has been admitted as state of art approach but fails when operated on the high number of dimensions for the reason mentioned above. In this paper, we determine the effects of different distance functions on an unlabeled dataset while digging outliers through the density-based approach. Further, we also explore findings regarding runtime and outlier score when dimension size and number of nearest neighbor points (min_pts) are varied. This analytic research is also very appropriate and applicable in the domain of big data and data science as well.**

Keywords*—High dimensional data; density-based anomaly detection; local outlier; outlier detection

I. Introduction

An outlier also known as anomaly could be defined as a data point that seems very dissimilar from other points based on some criteria [1,17]. This point should not be categorized as noise since it is likely to discover some very unexpected but useful information.

Outlier detection could be categorized in three different ways based on approaches [2,3], i.e. cluster-based, distance-based and density or local neighborhood-based. These approaches resemble each other as they operate on some notion of similarity. The only difference is the level of granularity or level of detail in terms of its analysis methodology. The local neighborhood approach differs from the global neighborhood method as shown in Fig. 1. Here point 1 (redpoint) is detected as an outlier for both approaches but the latter approach does not recognize point 2 (orange point) as an anomalous point.

Most of well-known outlier detection techniques work on full dimensional data. However, their performance gets deteriorated because of some intrinsic features present in data having a high number of dimensions [4]. Even techniques based on dimensionality reduction cannot resolve this problem as feature irrelevance/relevance is determined locally. Researches solved this inherent problem by formulating methodology on subspaces (a subset of attributes) [5].

However, it is not feasible to scan all subspaces within complete data as only the brute force technique assures all sets of attributes to explore anomalies inside data. But as far as its time complexity is concerned, it proves expensive enough. So there is desperate need to study and revise proximity functions to be applied on full feature space of data. A comparison of different proximity functions regarding high dimensional data answers the question of how outlier detection of high dimensional data could be coped with its inherent problems. Outliers have been classified either binary or scored which depends on the approach to be applied while exploring within datasets as shown in Fig. 2.

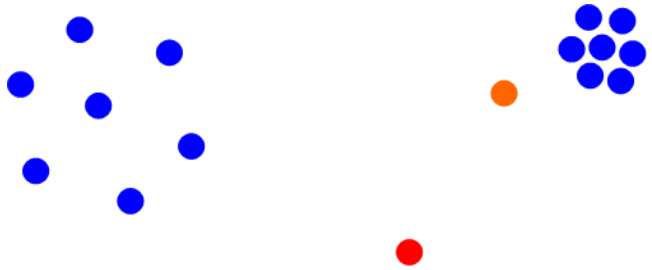

Fig. 1. An Outlier (Global vs Local) Figure.

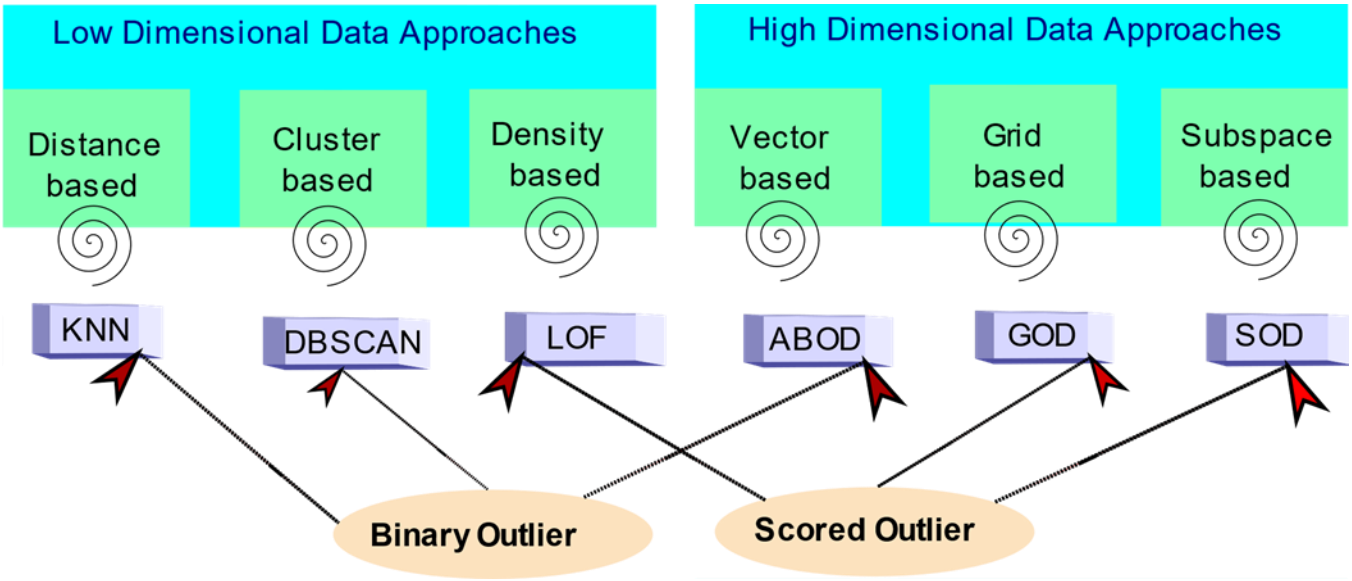

Fig. 2. Outlier Detection Techniques.

II. Problem Description

A. Motivation

Before experimenting and proving the hypothesis, we study and analyze how to deal with the curse of dimensionality when anomaly detection of high dimensional data is to be explored. As discussed earlier, subspace-based outlier detection is not a perfect solution regarding time expense and accuracy of results. We come across the following three reasons why investigating the problem is necessary, namely, i) Similarity of data points, ii) Curse of dimensionality, iii) Accuracy of outliers data.

B. Likeness

When the number of dimensions grows then at some point, distance functions cannot determine relative difference due to convergence of distance between any two data points. As shown in equation 1, when dimensionality grows to infinity then difference regarding the distance between farthest and nearest point is indistinguishable [6]. Hence there arises the importance of proximity function when most outlier detection techniques use the notion of distance.

$$\lim_{dim \to \infty} \frac{Distance.max - Distance.min}{Distance.min} = 0 \quad (1)$$

C. Curse of Dimensionality

The high number of dimensions is hard to describe, tedious to visualize and it becomes infeasible to dig out all subspaces due to exponential growth of all combinations of subspaces when each new dimension is added.

D. Accuracy of Outliers

Subspace anomaly detection techniques, as devised by many researchers, cannot explore all subspaces hidden in datasets for reasons discussed earlier. Hence outliers with accurate scores could not be retrieved which affects overall confidence in the accuracy of results.

III. Related Work

Outlier's definition proposed by Hawkins is accepted universally as it is very precise and straightforward, that is "An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism" [7]. There are many well-known domains in which outlier detection is being applied fruitfully like fault detection in the engineering field, fraud detection in the financial sector, intrusion detection in computer networks, etc. [8,9]. In a broad sense, outliers are classified/detected as either binary or scored depending upon methodology to be utilized or the requirement of stakeholders [10,11].

Amongst the class of local neighborhood-based outlier detection, the local outlier factor is an algorithm proposed by Hans-Peter Kriegel et al. in 2000 for finding abnormal data points by calculating the local deviation (outliers) of a given data point with respect to its neighbors [12, 18, 19].

Local Outlier Probability (LoOP) [13] is a method derived from local outlier factor but using inexpensive local statistics to become less sensitive to the choice of the parameter k. Besides, the resulting values are scaled to a value range of 0 to 1.

A novel method, Local Subspace Classifier (LSC) is used in [14] that is based on the feature vector extraction method. LSC determines outlier measure based on time increment for distance applied on the model. This method was improved in terms of computation in [15] by proposing method Fast LSC. In this approach, clustering is used to reduce the amount of data and hence proves ten times faster as compared to the LSC method.

Bo Tang [16] detects outliers based on distance function utilizing a density-based approach. He utilizes three types of measures to determine density estimation which are classic k nearest neighbors, reverse nearest neighbors and shared nearest neighbors.

IV. Experimental Work

The proposed research is evaluated and tested in RapidMiner and ELKI tools which are specialized ones for data mining and outlier detection tasks. Artificial data is generated to test and compare results with other algorithms. Public/Real data having a different number of dimensions and records present on research database websites like KDD and UCI machine learning laboratory is used for experimentation of algorithms.

As dimensionality grows towards infinity (number of dimensions large enough), the distance between any two data points approaches to zero (small enough to differentiate). That's why the Local Outlier Factor (LOF) of all data points gives similar points that exhibit that all data points are equally dispersed. As the value of Euclidean distance is different than Manhattan distance, so we get different results for LOF applied to the same dataset as shown in Table I. We can observe that the difference of LOF for Manhattan distance is higher than that of Euclidean distance. A Manhattan distance replaces Euclidean geometry with Taxicab geometry in which the distance between two data points is the sum of the absolute differences of their cartesian coordinates.

A dataset named "Concrete Data" extracted from UCI Machine Learning Repository is chosen for research experimentation. It is real data having 9 attributes and 1031 instances. Amongst the class of local neighborhood algorithms, LOF is selected to test and compare results on different proximity functions. Value of K (minimum points) is also varied to judge its effect on net results. The density of points is similar when a score of LOF is approximately equal to one. Outlier points are those which possess LOF score greater than one whereas inlier points show score less than one. Different proximity functions applied to density-based outlier detection (LOF) are discussed below. Euclidean distance is also known as the Pythagorean metric (shown in equation 2). It calculates straight line distance between any two data points.

$$d\ (x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (2)$$

Manhattan distance also known as the taxicab metric (shown in equation 3) finds the rectilinear distance between any two data points. This taxicab geometry has been used in regression analysis since the eighteenth century.

$$d\ (x, y) = \sqrt{\sum_{i=1}^{n} |x_i - y_i|} \quad (3)$$

Squared Euclidean distance is also extensively used in regression analysis. Optimization problems are relatively more easily solvable using this metric. It is determined using equation 4.

$$d^2\ (x, y) = (x_1 - y_1)^2 + (x_2 - y_2)^2 + - - + (x_n - y_n)^2 \quad (4)$$

A density based outlier detection algorithm, i.e. LOF is measured using equation 5 where LrdK represents local reachability of a point amongst k min_pts and NeighK denotes the neighborhood of a point for k min_pts.

$$\text{LOFk(p)} = \frac{\sum_{p' \varepsilon\, Neighk(p)} \frac{Lrdk(p')}{Lrdk(p)}}{Neighk(p)} \quad (5)$$

In Fig. 3, the run time of the LOF is shown while being applied on different proximity functions along with variation in several neighboring points also known as min_pts (k).

This graph clearly shows that run time for Squared Euclidean distance is minimum as compared to other proximity functions. It is also authenticated for different values of k (5, 10, 15) that other distance functions are relatively time expensive.

A comparison of outlier-ness and inlier-ness is shown in Fig. 4, where outlier score (outlier factor) and several outliers are compared for different proximity functions. Results reveal that the Squared Euclidean function gives much better results as both score and number get inclined.

Fig. 5 reveals an effect on the outlier score as the dimensionality of data is increased. The different number of dimensions to be used are 2, 4, 6 and 9. From this graph, we can conform to two very important things. First is that distance between points diminishes as the number of dimensions is increased, it is confirmed as the outlier score decreases by an increasing number of dimensions. Second is that the outlier score for points has reasonable differences for Squared Euclidean function as compared to other distance functions.

It could be concluded that high dimensional data requires to choose proximity function carefully while detecting outliers. In our work, Squared Euclidean proves to be very efficient for high dimensional data as its run time and outlier score are far better than that of other proximity functions.

TABLE I. LOF Comparison for Euclidean and Manhattan Distance

Dataset	Euclidean Distance, k=2	Manhattan Distance, k=2
ID=1: 8.0 0.0 1.0 2.0 2.0 8.0	lof=1.018	lof=1.024
ID=2: 10.0 9.0 1.0 2.0 2.0 11.0	lof=0.982	lof=0.986
ID=3: 4.0 8.0 1.0 2.0 2.0 7.0	lof=1.018	lof=1.013
ID=4: 3.0 1.0 1.0 2.0 2.0 2.0	lof=1.097	lof=1.307
ID=5: 0.0 4.0 1.0 2.0 2.0 14.0	lof=0.980	lof=0.988
	Min-LOF=0.982 Max-LOF=1.097	Min-LOF=0.986 Max-LOF=1.307
	Difference-LOF=0.116	Difference-LOF=0.318

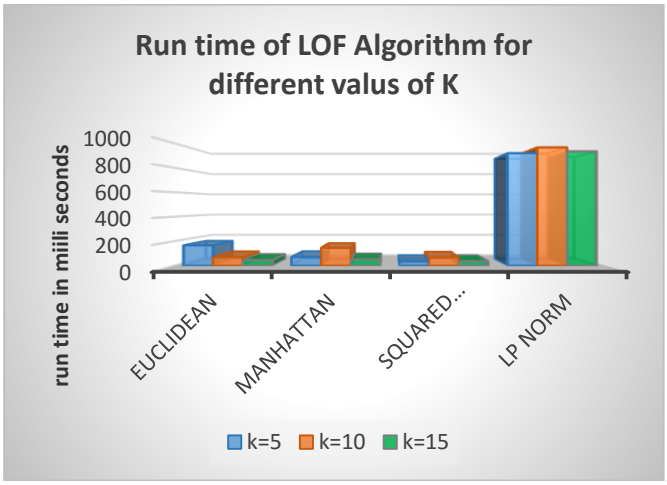

Fig. 3. LOF Comparison for different Proximity Functions.

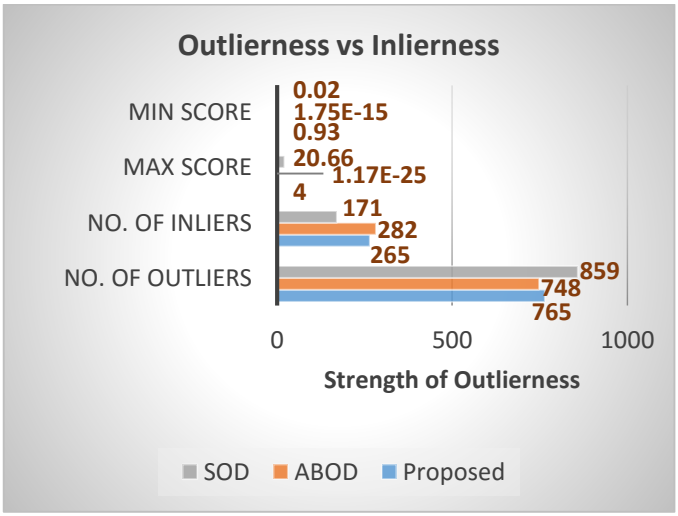

Fig. 4. A Comparison of Outlier and Inlier Scores.

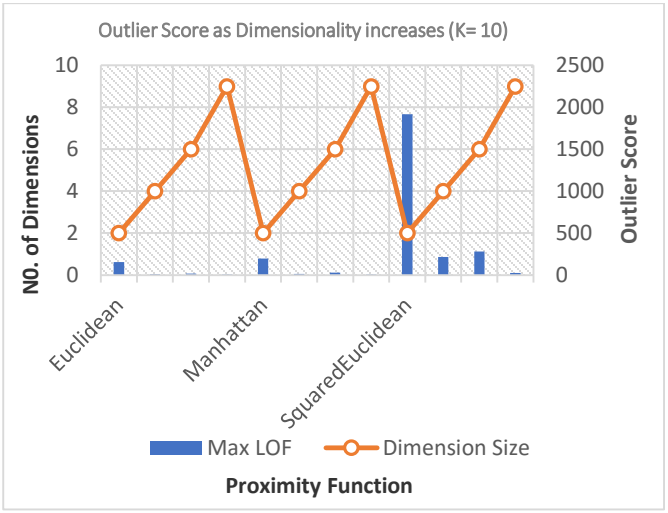

Fig. 5. Effect of Dimensionality on Outlier Score.

V. Limitation

The above experimentation works on numerical or continuous data only but it could be adapted for other data types if the distance between data points is quantifiable. For example, the edit distance metric calculates the distance between words containing alphabetical letters.

VI. Conclusion

Knowledge discovery has been utilized through outlier detection, a subfield of data mining. Data science and data mining help businessmen while taking crucial decisions for an organization. Local neighborhood-based outlier detection has been accepted as a state of art methodology while detecting outliers amongst different densities of clusters. High dimensional data pose serious challenges to data miners due to its inherent problems that result in the failure of traditional techniques. Another solution being tried is to find outliers within subspaces which compromises accuracy and also proves expensive in terms of its time complexity. In this study, we have compared results in terms of outlier-ness, inlier-ness, run time, dimensionality variation and different values of minimum points (k) when applied for different proximity functions to be utilized in density-based techniques. We have concluded that the Squared Euclidean function proves to be a very efficient proximity function while detecting outliers amongst high dimensional data.

Acknowledgment

Authors acknowledge the Department of Computer Science and IT, The Islamia University of Bahawalpur Pakistan and Department of Computer Science, Khawaja Fareed University Rahim Yar Khan Pakistan, for facilitating a suitable environment for the successful completion of this research work.

References

[1] C. C. Aggarwal and P. S. Yu, “An effective and efficient algorithm for high-dimensional outlier detection,” The VLDB Journal, vol. 14, no. 2, pp. 211–221, Apr. 2005, doi: 10.1007/s00778-004-0125-5.

[2] C. C. Aggarwal, Outlier Analysis. New York: Springer-Verlag, 2013.

[3] E. M. Knorr and R. T. Ng, “Algorithms for mining distance-based outliers in large datasets,” in VLDB, 1998, vol. 98, pp. 392–403.

[4] H. V. Nguyen, V. Gopalkrishnan, and I. Assent, “An unbiased distance-based outlier detection approach for high-dimensional data,” in International Conference on Database Systems for Advanced Applications, 2011, pp. 138–152.

[5] M. Ye, X. Li, and M. E. Orlowska, “Projected outlier detection in high-dimensional mixed-attributes data set,” Expert Systems with Applications, vol. 36, no. 3, Part 2, pp. 7104–7113, Apr. 2009, doi: 10.1016/j.eswa.2008.08.030.

[6] C. C. Aggarwal and P. S. Yu, “Outlier detection with uncertain data,” in Proceedings of the 2008 SIAM International Conference on Data Mining, 2008, pp. 483–493.

[7] D. M. Hawkins, Identification of outliers, vol. 11. Springer, 1980.

[8] R. Domingues, M. Filippone, P. Michiardi, and J. Zouaoui, “A comparative evaluation of outlier detection algorithms: Experiments and analyses,” Pattern Recognition, vol. 74, pp. 406–421, 2018.

[9] M. Bai, X. Wang, J. Xin, and G. Wang, “An efficient algorithm for distributed density-based outlier detection on big data,” Neurocomputing, vol. 181, pp. 19–28, 2016.

[10] J. Leng, “A novel subspace outlier detection approach in high dimensional data sets,” 2010.

[11] A. Agrawal, “Local subspace based outlier detection,” in International Conference on Contemporary Computing, 2009, pp. 149–157.

[12] B. M, KriegelHans-Peter, N. T, and SanderJörg, “LOF,” ACM SIGMOD Record, May 2000.

[13] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek, “LoOP: local outlier probabilities,” in Proceedings of the 18th ACM conference on Information and knowledge management, 2009, pp. 1649–1652.

[14] S. Hotta, “Local subspace classifier with transform-invariance for image classification,” IEICE TRANSACTIONS on Information and Systems, vol. 91, no. 6, pp. 1756–1763, 2008.

[15] H. Shibuya and S. Maeda, “Anomaly Detection Method Based on Fast Local Subspace Classifier,” Electronics and Communications in Japan, vol. 99, no. 1, pp. 32–41, 2016, doi: 10.1002/ecj.11770.

[16] B. Tang and H. He, “A local density-based approach for outlier detection,” Neurocomputing, vol. 241, pp. 171–180, 2011.

[17] Aggarwal C “Outlier analysis”, 2nd edn. Springer, Berlin, 2017.

[18] Yang, Ping, Dan Wang, Zhuojun Wei, Xiaolin Du, and Tong Li. "An Outlier Detection Approach Based on Improved Self-Organizing Feature Map Clustering Algorithm." IEEE Access 7 (2019): 115914-115925.

[19] Boddy, Aaron J., William Hurst, Michael Mackay, and Abdennour El Rhalibi. "Density-based outlier detection for safeguarding electronic patient record systems." IEEE Access 7 (2019): 40285-40294.