Nearest Neighbour Based Outlier Detection Techniques

Dr. Shuchita Upadhyaya, Karanjit Singh
Dept. of Computer Science and Applications, Kurukshetra University
Haryana, India
HQ Base Workshop Group EME, Meerut Cantt
UP, India

***Abstract* — Outlier detection is an important research area forming part of many application domains. Specific application domains call for specific detection techniques, while the more generic ones can be applied in a large number of scenarios with good results. This survey tries to provide a structured and comprehensive overview of the research on Nearest Neighbor Based Outlier Detection listing out various techniques as applicable to our area of research. We have focused on the underlying approach adopted by each technique. We have identified key assumptions, which are used by the techniques to differentiate between normal and Outlier behavior. When applying a given technique to a particular domain, these assumptions can be used as guidelines to assess the effectiveness of the technique in that domain. We provide a basic outlier detection technique, and then show how the different existing techniques in that category are variants of this basic technique. This template provides an easier and succinct understanding of the Nearest Neighbor based techniques. Further we identify the advantages and disadvantages of various Nearest Neighbor based techniques. We also provide a discussion on the computational complexity of the techniques since it is an important issue in our application domain. We hope that this survey will provide a better understanding of the different directions in which research has been done on this topic, and how techniques developed in this area can be applied in other domains for which they were not intended to begin with.**

***Keywords* — Outlier, Outlier Detection, Nearest Neighbour Concept, Multivariate, Algorithms, Data Mining, Nearest Neighbor based Outlier Detection, K-Nearest Neighbor, LOF Neighborhood, COF Neighborhood**

I. INTRODUCTION

A. General Description and Underlying Assumptions

The concept of nearest neighbor analysis has been used in several outlier detection techniques. Such techniques are based on the key assumption that *Instances of Normal Data occur in dense neighborhoods, while outliers occur far away from their closest neighbors.*

B. General Methodology of Operation

Nearest neighbor based outlier detection techniques require a distance (or similarity measure) defined between two data instances. Distance (or similarity) between two data instances can be computed in different ways.

1) For Continuous Attributes - A popular choice is the Euclidean distance, but other measures can also be used [1].

2) For Categorical Attributes – Often a simple matching coefficient is used. More complex distance measures such as [2,3] can also be used.

3) For Multivariate Data – The distance or similarity is usually computed for each attribute and then combined. [1].

Most of the techniques that will be discussed in this paper , do not require strictly metric distance measures. The measures are typically required to be positive-definite and symmetric, but need not satisfy the triangle inequality.

II. CATEGORISATION OF NEAREST NEIGHBOR BASED OUTLIER DETECTION TECHNIQUES

Nearest neighbor based outlier detection techniques can be broadly grouped into two categories based on how they compute the outlier score:

1) Distance to Kth Nearest Neighbor Based – These techniques use the distance of a data instance to its kth nearest neighbor as the outlier score.

2) Relative Density Based – These techniques compute the relative density of each data instance to compute its outlier score.

3) Using Other Manners - Additionally there are some techniques that use the distance between data instances in a different manner to detect outliers.

III. USING DISTANCE TO KTH NEAREST NEIGHBOR

A. Basic Definition

A basic nearest neighbor outlier detection technique is based on the following definition - *The outlier score of a data instance is defined as its distance to its k^{th} nearest neighbor in a given data set.* This basic technique has been applied to detect land mines from satellite ground images [4]

and to detect shorted turns (outliers) in the DC field windings of large synchronous generators where k has been taken as 1[5]. A threshold can be applied on the outlier score to determine if a test instance is an outlier or not. On the other hand, *n* instances with the largest outlier scores can be chosen as the outliers [6].

B. Extension of the Basic Technique

The basic technique has been extended by researchers in three different ways.

1) By Modifying the Definition - The first set of variants modify the above definition to obtain the outlier score of a data instance.

2) Use of Different Measures - The second set of variants use different distance/similarity measures to handle different data types.

3) Improvement of Efficiency – These focus on improving the efficiency of the basic technique in different ways. (the complexity of the basic technique is $O(N^2)$, where N is the data size)

C. Examples and Illustrations – Continuous Attributes Based

1) Outlier Score as Sum of Distances - Some techniques compute the outlier score of a data instance as the sum of its distances from its k nearest neighbors. [7,8,9] A similar technique called Peer Group Analysis has been applied to detect credit card frauds [10].

2) Outlier Score As Count of Number of Nearest Neighbours that are Not More Than distanced Apart - Another way to compute the outlier score of a data instance is to count the number of nearest neighbors (n) that are not more than d distance apart from the given data instance [11,12,13,14]. This method can also be viewed as estimating the global density for each data instance since it involves counting the number of neighbors in a hyper-sphere of radius d. For example, in a Two dimensional data set, the density of a data instance = $n/\pi d2$. The inverse of the density is the outlier score for the data instance. Instead of computing the actual density, several techniques fix the radius d and use 1/n as the outlier score, while several techniques fix n and use 1/d as the outlier score.

3) Examples and Illustrations – Other Data Types

While most of these techniques discussed in the Kth Nearest Neighbor category have been proposed to handle continuous attributes, variants have been proposed to handle other data types. A hyper-graph based Outlier Test Technique for Continuous Data (HOT) has also been proposed [15] where the categorical values are modeled using a hyper-graph, and distance between two data instances is measured by analyzing the graph connectivity.

A distance measure for data containing a mix of categorical and continuous attributes has been proposed for outlier detection [16]. The links between two instances are defined by adding distance for continuous and categorical attributes separately. For categorical attributes, the distance between them is the number of attributes for which the two instances have same value. For continuous attributes, to capture the dependencies between the continuous values a covariance matrix is maintained. Techniques proposed in [13] have been adapted for continuous sequences as well [17] while techniques proposed in [6] have been extended to spatial data [18].

IV. Variants with an Aim to Improve Efficiency

These basic techniques have also been varied to improve the efficiency.

D. Pruning the Search Space

This may be done by either ignoring instances which cannot be outliers or by concentrating on instances which are most likely outliers. A simple pruning step could result in the average complexity of nearest neighbor search to be almost linear for a sufficiently randomized data.

1) By Setting the Pruning Threshold to the weakest outlier - One such algorithm proposed in [19], first calculates the nearest neighbours for a data instance and then sets the outlier threshold for any data instance to the score of the weakest outlier found. This discards instances that are close, and hence not interesting.

2) Pruning Partitions not able to Contain Top K Outliers - This partition based technique [20] first clusters the instances and then for instances in each partition, it computes lower and upper bounds on distance of an instance from its Kth nearest neighbour. By using this information partitions that cannot possibly contain the top K outliers are identified and pruned. In the final phase from the remaining instances (belonging to unpruned partitions) outliers are computed.

3) Sampling – Here instead of computing over entire dataset, the nearest neighbour of every instance is computed within a smaller sample from the data set [21]. Sampling improves the efficiency of the nearest neighbour based technique and the complexity of the proposed technique is reduced to O(MN) where M is the sample size chosen.

4) Hypergrid based Partitioning – In this technique [22] the attribute space is partitioned into a hyper-grid consisting of hypercubes of fixed sizes. The thought process behind such techniques is that if a hypercube contains many instances, these hypercubes are likely to be normal. Also, if for a given instance, the hypercube that contains the instance and its adjoining hypercubes contain very few instances, the given instance is likely to be an outlier. This approach can be extended by linearizing the search space through the Hilbert space filling curve [8]. Here an n-dimensional data set is fitted in a hypercube N = [0; 1]n. This hypercube is then mapped to the interval I = [0; 1] using the Hilbert Space Filling Curve and the k-

nearest neighbors of a data instance are obtained by examining its successors and predecessors in I.

V. USING RELATIVE DENSITY

Density based outlier detection techniques estimate the density of the neighborhood of every data instance. An instance lying in a low density neighborhood is considered to be an outlier while those lying in a dense neighborhood is deemed normal.

For an instance under test, its distance from its k^{th} nearest neighbor may be viewed as the radius of a hyper-sphere, centered at the given test instance, which contains k other instances. Hence, distance to the k^{th} nearest neighbor for the test instance can be viewed as an estimate of the inverse of the density of the instance in the data set. Thus the basic nearest neighbor based hyper-grid based partitioning described in the previous subsection can be considered as a density based outlier detection technique.

If the data has regions of varying densities as in case of a 2 dimensional data set shown in Figure 1 then density based techniques perform poorly.

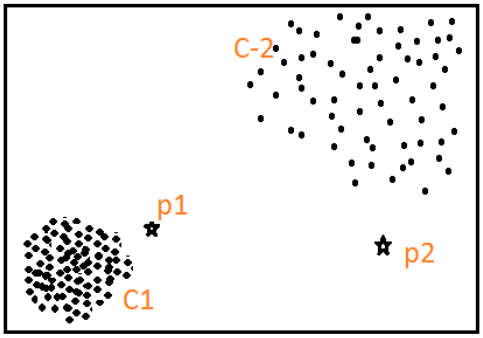

Figure 1: Advantage of Local Density Based Techniques over Global Density Based Techniques

Due to the low density of the cluster C_2 it is apparent that for every instance p in cluster C_2, the distance between the instance p and its nearest neighbor is greater than the distance between the instance p_1 and the nearest neighbor from the cluster C_1, and the instance p_1 will not be considered as outlier. Hence, the basic technique will fail to distinguish between p_1 and instances in C_2. However, the instance p_2 may be detected.

This issue of varying densities in the data set may be handled by a set of techniques which compute density of instances in relation to the density of their neighbors. In this, an outlier score known as *Local Outlier Factor* (LOF) is assigned to a given data instance [23,24]. The LOF score for any given data instance, is equal to the ratio of average local density of the k nearest neighbors of the instance and the local density of the data instance itself.

To find the local density for a data instance, the radius of the smallest hyper-sphere centered at the data instance, that contains its k nearest neighbors, is found. This k divided by the volume of this hyper-sphere gives the local density. For a normal instance lying in a dense region, its local density will be similar to that of its neighbors, while for an outlier instance, the local density will be lower than that of its nearest neighbors. Hence the outlier instances will get a higher LOF score. In Figure 1, LOF will be able to capture both outliers (p_1 and p_2) due to the fact that it considers the density around the data instances.

VI. VARIANTS OF LOF TECHNIQUE.

Variants estimate the local density of an instance through a variety of methods which are either different or adaptations of original technique to more complex data types. Since the complexity of original LOF technique is $O(N^2)$ (N is the data size), newer techniques propose to improve the efficiency of LOF.

A. Connectivity-Based Outlier Factor (COF)

COF is a variation of the LOF, the difference being in the manner in which the k neighborhood for an instance is computed. In COF, the neighborhood is computed in an incremental mode. To begin, the closest instance to the given instance is added to the neighborhood set. The next instance added is the one with its distance to the existing neighborhood to be minimum among all remanant instances. The distance between an instance and a set of instances is defined as the minimum distance between the given instance and any instance belonging to the given set. In this fashion the neighborhood is grown until it reaches size k. Once the neighborhood is computed, the outlier score (COF) is computed in the same manner as LOF.

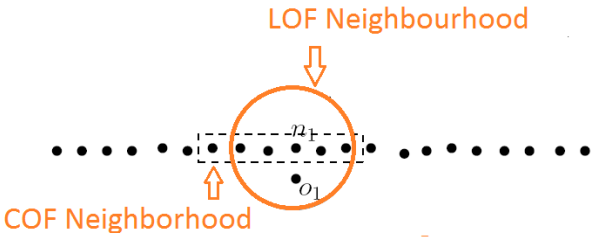

Figure 2: Illustration of difference between the two computations. Note that COF is able to capture regions such as Straight Lines

B. Outlier Detection using In-degree Number (ODIN)

This is a simpler version of LOF which calculates a quantity called ODIN for each data instance which equals the number of *k nearest neighbors of the data instance which have the given data instance in their k nearest neighbor list [26]. The inverse of ODIN is the outlier score for the data instance.*

C. Multi-granularity Deviation Factor (MDEF)

Another variation of LOF is a measure called MDEF [27] which for a given data instance is equal to the standard deviation of the local densities of the nearest neighbors of the given data instance (including the data instance itself). The inverse of the standard deviation is the outlier score for the

data instance. This detection technique not only finds outlier instances but also outlier micro-clusters.

D. Variants of LOF to handle Different Data Types

Some Datatypes where variants of LOF have been proposed are tabulated below:

TABLE I

VARIANTS OF LOF TO SUIT DIFFERENT DATA TYPES

Data Types	References
Spatial Outliers in Climate Data	[28]
Categorical Attributes	[29]
Sequential Outliers in Protein Sequences usings Probabilistic Suffix Trees (PST)	[30]
Video Sensor Data	[31]

E. LOF Technique Variants to Improve its Efficiency.

1) Finding Only Top n Outliers - This variant, in which only the top n outliers are found instead of finding LOF score for every data instance, includes finding micro-clusters in the data and then finding upper and lower bound on LOF for each of the micro-clusters [32].

2) Prune all clusters which do not contain instances to figure in Top n Outlier List - There are three variants of LOF proposed [33] which enhance its performance by making certain assumptions about the problem to prune all those clusters which definitely do not contain instances which will figure in the top n \outlier list". detailed analysis to find the LOF score for each instance is carried out for the remaining clusters.

VII. COMPUTATIONAL COMPLEXITY

The basic drawback of nearest neighbor and LOF techniques is the $O(N^2)$ complexity. Since these techniques find nearest neighbors for each instance, they need efficient data structures such as k-d trees [34] and R-trees [35]. But such techniques do not scale well with increase in the number attributes. Some techniques therefore directly optimize the detection process with the assumption that only top few outliers are interesting. But then they are inapplicable if an outlier score is required for every test instance. Partitioning the attribute space into a hyper-grid renders linearity in data size but gets exponential in the number of attributes, and therefore not well suited for large number of attributes. Although sampling techniques try to address the $O(N^2)$ complexity issue by determining the nearest neighbors within a small sample of the data set but then sampling itself might result in incorrect outlier scores if the sample size is very small.

VIII. ADVANTAGES AND DISADVANTAGES OF NEAREST NEIGHBOR BASED TECHNIQUES

A. Advantages

The advantages of nearest neighbor based techniques are as follows:

1) Purely Data Driven - They are unsupervised in nature and do not make any assumptions regarding the generative distribution for the data. This purely data driven approach is a key advantage of this technique.

2) Better Performance of Semi-supervised Techniques - In terms of missed outlier, in this case, semi-supervised techniques perform better than unsupervised techniques, since the likelihood of an outlier to form a close neighborhood in the training data set is very low.

3) Easy Adaptation to Different Data Types – It is straight forward to apply nearest neighbor based techniques to a different data type since it primarily requires defining an appropriate distance measure for the given data.

B. Disadvantages

The disadvantages of nearest neighbor based techniques are as follows:

4) For Unsupervised Techniques - If the data has normal instances that do not have enough close neighbors or if the data has outliers that have enough close neighbors, nearest neighbor technique fails to label them correctly, resulting in missed outliers.

5) For Semi-supervised Techniques - If normal instances in test data do not have enough similar normal instances in the training data, false positive rate is high.

6) Computational complexity of testing phase – This is a great challenge involving computation of distance of each test instance with all instances belonging to either the test data itself, or to the training data, to compute the nearest neighbors.

IX. PERFORMANCE

Performance of a nearest neighbor based technique relies heavily on a distance measure, defined between a pair of data instances, that can distinguish between normal and anomalous instances effectively. Defining distance measures between instances can be challenging when data is complex, e.g. sequences, graphs, etc

X. CONCLUDING REMARKS AND FUTURE WORK

In this survey we have discussed different ways in which the problem of Nearest Neighbour based outlier detection has been formulated in literature, and we attempted to provide an overview of the huge literature on different techniques. For each subcategory of Nearest Neighbour Based technique, we could identify a unique assumption regarding the notion of normal data and outliers. When applying a given technique to a particular domain, these assumptions can be used as guidelines to assess the effectiveness of the technique in that domain. We understand that ideally, a comprehensive survey

should not only allow a reader to understand the motivation behind using a particular technique, but also provide a comparative analysis of various techniques. But the current research done in an unstructured fashion, without relying on a unified notion of outliers, makes a theoretical understanding of the outlier detection problem a difficult task. A possible future work would be to unify the assumptions made by different techniques regarding the normal and outlier behaviour into a statistical or machine learning framework.

XI. Acknowledgments

Dr Shuchita Upadhyaya and Karanjit Singh and thanks the staff of Computer Science and Applications department of Kurukshetra University for their wholehearted support in referencing the study material. The authors sincerely thank the library staff for the late duty hours at times.

References

[1] Tan, P.-N., Steinbach, M., and Kumar, V. 2005. Introduction to Data Mining. Addison-Wesley.

[2] Boriah, S., Chandola, V., and Kumar, V. 2008. Similarity measures for categorical data: A comparative evaluation. In Proceedings of the eighth SIAM International Conference on Data Mining. 243 - 254.

[3] Chandola, V., Eilertson, E., Ertoz, L., Simon, G., and Kumar, V. 2006. Data mining for cyber security. In Data Warehousing and Data Mining Techniques for Computer Security, A. Singhal, Ed. Springer.

[4] Byers, S. D. and Raftery, A. E. 1998. Nearest neighbor clutter removal for estimating features in spatial point processes. Journal of the American Statistical Association 93, 577 - 584.

[5] Guttormsson, S., II, R. M., and El-Sharkawi, M. 1999. Elliptical novelty grouping for on-line short-turn detection of excited running rotors. IEEE Transactions on Energy Conversion 14, 1 (March).

[6] Ramaswamy, S., Rastogi, R., and Shim, K. 2000. Efficient algorithms for mining outliers from large data sets. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data. ACM Press, 427 - 438.

[7] Eskin, E., Arnold, A., Prerau, M., Portnoy, L., and Stolfo, S. 2002. A geometric frame-work for unsupervised outlier detection. In Proceedings of Applications of Data Mining in Computer Security. Kluwer Academics, 78 - 100.

[8] Angiulli, F. and Pizzuti, C. 2002. Fast outlier detection in high dimensional spaces. In Proceed-ings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery. Springer-Verlag, 15 - 26.

[9] Zhang, J. and Wang, H. 2006. Detecting outlying subspaces for high-dimensional data: the new task, algorithms, and performance. Knowledge and Information Systems 10, 3, 333 - 355.

[10] Bolton, R. and Hand, D. 1999. Unsupervised profiling methods for fraud detection. In Credit Scoring and Credit Control VII.

[11] Knorr, E. M. and Ng, R. T. 1997. A unified approach for mining outliers. In Proceedings of the 1997 conference of the Centre for Advanced Studies on Collaborative research. IBM Press, 11.

[12] Knorr, E. M. and Ng, R. T. 1998. Algorithms for mining distance-based outliers in large datasets. In Proceedings of the 24rd International Conference on Very Large Data Bases. Morgan Kaufmann Publishers Inc., 392 - 403.

[13] Knorr, E. M. and Ng, R. T. 1999. Finding intensional knowledge of distance-based outliers. In The VLDB Journal. 211 - 222.

[14] Knorr, E. M., Ng, R. T., and Tucakov, V. 2000. Distance-based outliers: algorithms and applications. The VLDB Journal 8, 3-4, 237 - 253.

[15] Wei, L., Qian, W., Zhou, A., and Jin, W. 2003. Hot: Hypergraph-based outlier test for categori-cal data. In Proceedings of the 7th Pacific-Asia Conference on Knowledge and Data Discovery. 399 - 410.

[16] Otey, M. E., Ghoting, A., and Parthasarathy, S. 2006. Fast distributed outlier detection in mixed-attribute data sets. Data Mining and Knowledge Discovery 12, 2-3, 203 - 228.

[17] Palshikar, G. K. 2005. Distance-based outliers in sequences. Lecture Notes in Computer Science 3816, 547 - 552.

[18] Kou, Y., Lu, C.-T., and Chen, D. 2006. Spatial weighted outlier detection. In Proceedings of SIAM Conference on Data Mining.

[19] Bay, S. D. and Schwabacher, M. 2003. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM Press, 29 - 38.

[20] Ramaswamy, S., Rastogi, R., and Shim, K. 2000. Efficient algorithms for mining outliers from large data sets. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data. ACM Press, 427 - 438.

[21] Wu, M. and Jermaine, C. 2006. Outlier detection by sampling with accuracy guarantees. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, New York, NY, USA, 767 - 772.

[22] Knorr, E. M. and Ng, R. T. 1998. Algorithms for mining distance-based outliers in large datasets. In Proceedings of the 24rd International Conference on Very Large Data Bases. Morgan Kaufmann Publishers Inc., 392 - 403.

[23] Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. 1999. Optics-of: Identifying local outliers. In Proceedings of the Third European Conference on Principles of Data Mining and Knowledge Discovery. Springer-Verlag, 262 - 270.

[24] Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. 2000. Lof: identifying density-based local outliers. In Proceedings of 2000 ACM SIGMOD International Conference on Management of Data. ACM Press, 93 - 104.

[25] Tang, J., Chen, Z., chee Fu, A. W., and W.Cheung, D. 2002. Enhancing effectiveness of outlier detections for low density patterns. In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining. 535 - 548.

[26] Hautamaki, V., Karkkainen, I., and Franti, P. 2004. Outlier detection using k-nearest neigh-bour graph. In Proceedings of 17th International Conference on Pattern Recognition. Vol. 3. IEEE Computer Society, Washington, DC, USA, 430 - 433.

[27] Papadimitriou, S., Kitagawa, H., Gibbons, P. B., and Faloutsos, C. 2002. Loci: Fast out-lier detection using the local correlation integral. Tech. Rep. IRP-TR-02-09, Intel Research Laboratory, Pittsburgh, PA. July.

[28] Sun, P. and Chawla, S. 2004. On local spatial outliers. In Proceedings of 4th IEEE International Conference on Data Mining. 209 - 216.

[29] Yu, J. X., Qian, W., Lu, H., and Zhou, A. 2006. Finding centric local outliers in categorical/ numerical spaces. Knowledge and Information Systems 9, 3, 309 - 338.

[30] Sun, P. and Chawla, S. 2006. Slom: a new measure for local spatial outliers. Knowledge and Information Systems 9, 4, 412 - 429.

[31] Pokrajac, D., Lazarevic, A., and Latecki, L. J. 2007. Incremental local outlier detection for data streams. In Proceedings of IEEE Symposium on Computational Intelligence and Data Mining.

[32] Jin, W., Tung, A. K. H., and Han, J. 2001. Mining top-n local outliers in large databases. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. ACM Press, 293 - 298.

[33] Chiu, A. and chee Fu, A. W. 2003. Enhancements on local outlier detection. In Proceedings of 7th International Database Engineering and Applications Symposium. 298 - 307.

[34] Bentley, J. L. 1975. Multidimensional binary search trees used for associative searching. Com-munications of the ACM 18, 9, 509 - 517.

[35] Roussopoulos, N., Kelley, S., and Vincent, F. 1995. Nearest neighbor queries. In Proceedings of ACM-SIGMOD International Conference on Management of Data.