International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 2 | ISSN : 2456-3307

A Survey on Anomalies Detection using Density Based - Rank Based Outlier Detection Methods

Nehal Patel*, Jayna Shah
Computer Engineering, Sardar Vallabhbhai Patel Institute of Technology, Vasad, Gujarat, India

ABSTRACT

Outlier Analysis is important research area in data mining. Outlier detection is the process of finding an outlying pattern from a given dataset. Outlier detection became an important subject in different knowledge domains. The aim of this paper is to present various Density and Rank based techniques of outlier detection. So a researcher can get direction with these approaches and they can be integrated with any kind of general Applications.

Keywords : Outlier Analysis, Anomaly Detection, Density Based, Rank Based.

I. INTRODUCTION

Detection of anomalies in data is defined as finding patterns in data that do not confirm to normal behavior or data that do not confirmed to expected behavior, such a data is called as outliers, anomalies, exceptions. Anomaly and Outlier have similar meaning. The analysts have strong interest in outliers because they may represent critical and actionable information in various domains, such as cybersecurity, finance, health, defence, home safety, industry, science and many KDD applications. An Outlier is an observation in data instances, which is different from the others in dataset. There are many reasons due to outliers arise like poor data quality, malfunctioning of equipment. Outlier detection is very much popular in data mining field and it is an active research area due to its various applications like fraud detection, network sensor, email spam, stock market analysis, intrusion detection and also in data cleaning.

II. RELATED WORKS

Outlier Analysis and Anomaly Detection Approach Layout

An outlier/anomaly is an observation point that is distant from other observations. An anomaly is a "variation from the norm".

- Density-based: Points that are in relatively low-density regions are considered more anomalous.
- Rank-based: A data point is "more anomalous" if it is not the nearest neighbor of its nearest neighbors.

For each of these approaches, it can be supervised, semi-supervised, or unsupervised.

1) In the supervised case, classification labels are known for a set of "training" data, and all comparisons and distances are with respect to such training data.
2) In the unsupervised case, no such labels are known, so distances and comparisons are with respect to the entire data set.
3) In semi-supervised problems, labels are known for some data, but not for most others. For instance, a few cases of intrusion of a certain new category may be available, and a semi-supervised learning algorithm may attempt to determine which other suspected cases of intrusion belong to the same category. Algorithms often proceed in multiple phases, with the early phase assigning tentative labels to unlabeled data.

An unsupervised anomaly detection algorithm should meet the following characteristics:

1. Normal behaviors have to be dynamically defined. No prior training data set or reference data set for normal behavior is needed.
2. Outliers must be detected effectively even if the distribution of data is unknown.
3. The algorithm should be adaptable to different domain characteristics; it should be applicable or modifiable for outlier detection in different domains, without requiring substantial domain knowledge.

a. Density based Approaches

In density based approaches the main idea is to consider the behaviours of a point with respect to its neighbours' density values. The neighbourhood is conceptualized by considering k nearest neighbours, where k is either iteratively estimated or is a preassigned integer. The underlying assumption is that if the density at a point p is 'smaller' than the densities of its neighbours, it must be an anomaly. It examines the k-neighbourhood of a data point, has many good features. For instance, it is independent of the distribution of the data and is capable of detecting isolated objects.

Three well-known density-based algorithms are as following:

1) LOF [1]

Breunig et al. proposed the following approach to find anomalies in a given dataset. As the name of the algorithm suggests, the Local Outlier Factor (LOF) measures the local deviation of a data point p ∈ D with respect to its k nearest neighbors. A point p is declared anomalous if its LOF is 'large.'

The LOF of a point is obtained as described in the following steps:

1. Find the distance, $d_k(p)$ between p and its k^{th} nearest neighbor. Denote the set of k nearest neighbors of p by $N_k(p) = \{q \in D - \{p\} : d(p,q) \leq d_k(p)\}$.

2. Define the reachability distance of a point q from p, as $R_k(p,q)$ = max $\{d_k(q), d_k(p,q)\}$.

3. The local reachability density of a point is defined as the inverse of the average reachability distance. Specifically, it is

$$l_k(p) = [\frac{\sum_{q \in N_k(p)} R_k(p,q)}{|N_k(p)|}]^{-1}.$$

4. LOF (local outlier factor) of a point p is defined as:

$$L_k(p) = \left[\frac{\sum_{o \in N_k(p)} \frac{l_k(o)}{l_k(p)}}{|N_k(p)|}\right]$$

5. The LOF of each point is calculated, and points are sorted in decreasing order of $L_k(p)$. If the LOF values are `large', the corresponding points are declared as outliers.

6. To account for k, the final decision is taken as follows: $L_k(p)$ is calculated for selected values of k in a pre-specified range, max $L_k(p)$ is retained, and a p with large LOF is declared an outlier.

2) COF [2]

LOF performs well in many application domains, but its effectiveness will diminish if the density of an outlier is close to densities of its neighbours. To solve such a deficiency of LOF, Tang et al. suggest a new method to calculate the density as described below.

The COF of a point is obtained as described in the following steps:

1. Define the distance between two non-empty sets P and Q as d (P, Q) = min {d (p, q) : p ∈ P; q ∈ Q}. This can be used to find the minimum distance between a point and a set by treating one of the set as a singleton.

2. Given a point p, define set-based path (SBN) of length k as a path < p ≡ $p_1, p_2, \ldots, p_k$ > such that for all 1≤ i ≤ k-1, p_{i+1} is the nearest neighbour of the set $\{p_1, p_2, \ldots, p_i\}$. In other words, the SBN-path represents the order in which nearest neighbours of p are successively obtained. The set $N_k(p) = \{p_1, p_2, \ldots, p_k\}$ is the set of k nearest neighbours of p.

3. The Set-based trail (SBT) is an ordered collection of k-1 edges associated with a given SBN path < p $\equiv p_1, p_2, \ldots, p_k$ > .The i^{th} edge $e_i$connects a point o ∈ {p_1,p_2,….., p_i}to p_{i+1} and is of minimum distance; i.e., length of e_i is equal to d(o, p_{i+1}) = d({p_1, p_2,….., p_i}, {p_{i+1},….. p_k}). Denote the length of edge $e_i$as l (e_i).

4. Given p, the associated SBN path < p ≡ p_1 , p_2 ,…..,p_k >, and the SBT < e_i ; e , …..,e_{k-1} >, the average-chaining distance (A) of p is weighted sum of the lengths of the edges, with larger weights assigned to nearest edges, that is:

$$A_{N_k(p)}(p) = \frac{2}{k}\sum_{i=1}^{k-1}\frac{k-i}{k-1}\, l(e_1).$$

5. Finally, the connectivity-based outlier factor (COF) of a point p is defined as

$$COF_k(p) = [A_{N_k(p)}(p)]\, [\frac{\sum_{o\in N_k(p)} A_{N_k(p)}(o)}{|N_k(p)|}]^{-1}.$$

6. As in COF, larger values of $COF_k(p)$ denote higher possibility that p is an outlier.

3) INFLO [3]

Jin et al. assigned to each object the degree of being INFLuenced Outlierness (INFLO) and introduce a new idea called 'reverse neighbors' of a data point when estimating its density distribution.

The INFLO of a point is obtained as described in the following steps:

INFLO the k nearest neighbors and reverse nearest neighbors of an object p are used to obtain a measure of outlierness. Recall that given an object p

1. Reverse Nearest Neighborhood (RNN) of p is defined as

$$RN_k(p) = \{q : q \in D \; and \; p \in N_k(q)\}.$$

Note that $N_k(p)$ has exactly k objects but $RN_k(p)$ may not have k objects. In some instances, it may be empty, because for all q ∈ (p) , p may not be in any of the set of $N_k(q)$.

2. The k-influential space for p, denoted as $IS_k(p)$ = $N_k(p) \cup RN_k(p)$.

3. The influenced outlierness of a point p is defined as

$$INFLO_k(p) = \frac{1}{den(p)} \frac{\sum_{o\in IS_{k(p)}} den(0)}{|(IS_k(p))|}$$

The common theme among these algorithms is that they all assign outlierness to each object in the data set and an object will be considered as an outlier if its outlierness is greater than a pre-defined threshold (usually the threshold is determined by users or domain experts).

b. Rank Based Approaches

Density based Approaches has the following shortcomings:

• If some neighbors of the point are located in one cluster, and the other neighbors are located in another cluster, and the two clusters have different densities, then comparing the density of the data point with all of its neighbors may lead to a wrong conclusion and the recognition of real outliers may fail.

• The notion of density does not work well for sparse data sets such as a cluster of points on a single straight line. Even if each point in the set has equal distances to its closest neighbors, its density may vary depending on its position in the dataset.

In such situations, to find anomalous observations, the ideal solution is to transform the data so that all regions in the transformed space have similar local distributions. The rank-based approach attempts to achieve this goal.
Here Some Approaches are as following:

1) RBDA [4]

RBDA is a rank-based outlier detection approach that identifies outliers based on the mutual closeness of a data point and its neighbors. This is a new approach to identify outliers based on the mutual closeness of a data point and its neighbors.

Description of Rank-based Detection Algorithm (RBDA) Algorithm:

1. For p ∈ D let q ∈ $N_k(p)$. calculate the rank of p among all neighbors of q; i.e., calculate the set of d (q, o) for all o ∈ D − {q} and find the rank of d (q, p) in this set. Let this be $r_q(p)$.

2. 'Outlierness' of p, denoted by $O_k(p)$, is defined as:

$$O_k(p) = \frac{\sum_{q \in N_{k(p)}} r_q(p)}{|(N_k(p))|}.$$

If $O_k(p)$ is 'large' then p is considered an outlier.

3. To determine a criterion for `largeness', let D_O= {p $\in D \mid O_k(p) \leq O_{max}$ }where O_{max} is chosen such that the size of D_O is 75% of the size of D. normalize $O_k(p)$ as below:

$$Z_k(p) = \frac{1}{S_k}\,(O_k(p) - \bar{O}_k)$$

Where,

$\bar{O}_k = \frac{1}{|D_o|} \sum_{p\,\in D} O_k(p)$ and

$S_k^2 = \frac{1}{|D_o|-1} \sum_{p\,\in D} (O_k(p) - \bar{O}_k)^2$ and if the normalized value $Z_k(p) \geq 2.5$, then declare that p is an outlier.

2) RADA [5]

Rank-based approach ignores the useful information contained in the distance of the object from other neighbouring objects. To overcome this weakness of RBDA due to "cluster density effect", adjust the value of RBDA by the average distance of p from its k–neighbours.

Step by step description of this rank and distance based detection algorithm is given below:

1. Choose three positive integers k, l, m^*.

2. Find the clusters in D by NC(l, m^*) method.

3. Declare an object O a potential-outlier if it is does not belong to any cluster.

4. Calculate a measure of outlierness:

$$W_k(p) = O_k(p) \times \frac{\sum_{q \in N_{k(p)}} d\,(q,p)}{|(N_k(p))|}$$

5. If p is a potential-outlier and $W_k(p)$) is large, declare p is an outlier.

3) ODMR [6]

For a point near a dense and large cluster, all of the k nearest neighbors of p may find their neighbors in a close vicinity and p may not be their neighbour; a point near a dense and large cluster may be declared an anomaly, although it may not be so. ODMR modifies the rank of an observation by assigning a weight to overcome the cluster density effect. In ODMR all clusters (including isolated points viewed as a cluster of size 1) are assigned weight 1, i.e., all |C| observations of the cluster C are assigned equal weights = 1/|C|. The "modified-rank" of p with respect to q is defined as the sum of weights associated with all observations within the circle of radius d(q, p) centered at q, and the measure of outlierness is given by Equation (1) with $r_q(p)$ replaced by the modified rank of p.

In this calculate "modified-rank" of p, which is defined as the sum of weights associated with all observations within the circle of radius d (q, p) centered at q; that is modified-rank of p from q = mr_q(p) = $\sum_{s\,\in\{d(q,s)\,\leq\,d\,(q,p)\,\}} Weight\ (s)$ and sum the "modified-ranks" in q ∈ $N_k(p)$.

Step by step description of the proposed method is as follows:

1. Choose three positive integers k, l, m^*.

2. Find clusters in D by NC(l,m^*). All objects not belonging to any cluster are declared as potential-outliers.

3. If C is a cluster and p ∈ C, then the weight of p is b(p) = 1/|C|.

4. For p ∈ D and q ∈$N_k(p)$, Q denotes the set of points within a circle of radius d (q, p), i.e., Q = {s ∈ D | d(q, s) ≤ d(q, p)}. Then the modified-rank of p with respect to q, denoted as mr_q(p), is computed as mr_q(p) = $\sum_{s\,\in Q} b(s)$.

5. For a potential outlier p, its ODMR-outlierness, denoted as $ODMR_k(p)$, is defined as:

$$ODMR_k(p) = \sum_{q \in N_k(p)} mr_q(p).$$

6. If p is a potential outlier and $ODMR_k(p)$, is large, declare p is an outlier.

The performance of an outlier detection algorithm based on rank alone is highly influenced by cluster density variations. Furthermore, by definition, ranks use the relative distances and ignore the 'true' distances between the observations. The overall performance of the ODMR better than previously known algorithms.

c. Evaluation Metrics

To evaluate the performance of the algorithms, three metrics were selected Precision, Recall, and Rank-Power[6-9].

Suppose, using a given outlier detection algorithm, we identify m most suspicious instances in D which contains d_t true outliers and let m_t be the number of true outliers among m instances. Then Precision which measures the proportion of true outliers in top m suspicious instances,

$$\text{Precision} = \frac{m_t}{m}$$

and Recall which measure the accuracy of an algorithm is:

$$\text{Recall} = \frac{m_t}{d_t}$$

Precision and recall don't capture the effectiveness completely. One algorithm may identify an outlier as the most suspicious while another algorithm may identify it as the least suspicious. Yet the above two measures remain the same. Ideally, an algorithm will be considered more effective if it true outliers occupy top positions and non-outliers in the bottom of the m suspicious instances. Rank-Power was proposed by Tang et al. [10] to capture this notion. Let n denote the number of outliers found within top m instances and R_i denote the rank of the ith true outlier. Then,

$$\text{RankPower} = \frac{n(n+1)}{2\sum_{i=1}^{n} R_i}$$

Rank-Power takes maximum value 1 when all n true outliers are in top n positions.

For a fixed value of m, larger values of these metrics imply better performance.

TABLE I
COMPARATIVE ANALYSIS

Sr. No.	Author Name	Technique	Based on	Useful Points
1	Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, Jörg Sander	Local Outlier Factor(LOF)[1]	Density	-Find meaningful outlier. -Assume that patterns have high (relative) density. -Do not work well where the patterns are of low densities
2	Jian Tang, Zhixiang Chen, Ada Wai-chee Fu, David W. Cheung	Connectivity-based outlier factor (COF)[2]	Density	-Detect outlier independently of the patterns from which they deviate -More effective when a cluster and a neighboring outlier have similar neighborhood densities. -COF was designed to work better than LOF in data sets with sparse neighbourhoods, but its computation cost larger than LOF.
3	Wen Jin, Anthony K. H. Tung, Jiawei Han,	INFLuential measure of Outlierness (INFLO)[3]	Density	-Use symmetric neighborhood relationship. (consider both neighbors and reverse neighbors at the time of density distribution). -but performance is poor if p's neighborhood

	and Wei Wang			includes data objects from groups of different densitities.
4	H. Huang, K. Mehrotra, C.K. Mohan	Rank-based Detection Algorithm (RBDA)[4]	Rank	-Outlier is find, based on the mutual closeness of a point and its neighbors. -It eliminates the problem of density calculation in the neighborhood of the point that improves performance. -Performs better than several density based methods. -It overcomes the weakness of density based approach. -Doesn’t perform well as COF when dataset consisting of clusters with special shapes such as lines or circles. -The effectiveness of ranking is not good. -It suffers from cluster density effect.
5	Huaming Huang, Kishan Mehrotra, Chilukuri K. Mohan	Rank with Averaged Distance Algorithm (RADA)[5]	Rank and Distance	-Removes deficiencies of Rank based algorithms and it overcomes cluster density effect with averaged distance.
6	Huaming Huang, Kishan Mehrotra, Chilukuri K. Mohan	Outlier Detection using Modified-Ranks(ODMR)[5]	Modified Rank	-Removes deficiencies of Rank based algorithms and it overcomes cluster density effect by using modified rank.

III. CONCLUSION

By analyzing various kinds of Density based and Rank based outlier detection Algorithms it is found that they have their own advantages and disadvantages. And when working with unsupervised method, these algorithms works well for Anomaly Detection. These Algorithms can be applied to any suitable applications without prior knowledge and with unknown data distribution.

IV. REFERENCES

[1]. Breunig, M.M., Kriegel, H.-P., Ng, R.T., Sander, J.,” Lof: identifying density-based local outliers”, ACM Sigmod Record, vol. 29, pp. 93–104. ACM (2000).

[2]. Tang, J., Chen, Z., Fu, A.W., Cheung, D.W.,” Enhancing effectiveness of outlier detections for low density patterns”, vol. 2336, pp. 535–548. Springer, Heidelberg (2002).

[3]. Jin, W., Tung, A.K.H., Han, J., Wang, W.,” Ranking outliers using symmetric neighborhood relationship”, vol. 3918, pp. 577–593. Springer, Heidelberg (2006).

[4]. Huang, H., Mehrotra, K., Mohan, C.K.,”Rank-based outlier detection”, 83(3), pp. 518–531. Journal of Statistical Computation and Simulation (2013).

[5]. Huang, Huaming, Kishan Mehrotra, and Chilukuri Mohan.,"Algorithms for detecting outliers via clustering and ranks.", pp. 20-29. Advanced Research in Applied Artificial Intelligence (2012).

[6]. R. Baeza-Yates and B. Ribeiro-Neto,“Modern information retrieval”, Addison-Wesley Longman Publishing Co. Inc., Boston (1999).

[7]. X. Meng and Z. Chen,“On user-oriented measurements of effectiveness of web information retrieval systems,", pp. 527-533, In Proceeding of the international conference on internet computing (2004).

[8]. G. Salton,”Automated text processing: The transformation, analysis, and retrieval of information by computer.”, Addison-Wesley Longman Publishing Co. Inc., Boston (1998).

[9]. H. Cao, G. Si, Y. Zhang, and L. Jia, ,”Enhancing effectiveness of density-based outlier mining scheme with density-similarity-neighbor-based outlier factor", Expert Systems with Applications: An International Journal, vol. 37, December (2010).

[10]. J. Tang, Z. Chen, A. W. Fu, and D. W. Cheung,“Capabilities of outlier detection schemes in large datasets, framework and methodologies.," Knowledge and Information Systems, vol. 11, no. 1, pp. 45-84,(2006).