ANALYSIS VOL. 1, 2015

ISSN 2394–3750
EISSN 2394–3769

SCIENCE & TECHNOLOGY

Outliers

Prayank Mathur

Indian School of Mines, India; Email: prayank27mathur@gmail.com

Article History
Received: 08 January 2015
Accepted: 13 February 2015
Published: 1 April 2015

Citation
Prayank Mathur. Outliers. *Science & Technology,* 2015, 1(2), 63-66

INTRODUCTION

Data Mining has always been an area of huge interest for the scientific community because with the ever increasing data, its storage, manipulation and retrieval are of at most importance nowadays. One of the major fields under this section is outlier detection and analysis which is an important topic of research nowadays. Outlier is defined as data objects which are grossly different or inconsistent from the remaining dataset. Finding them is one of the most prominent problems in the data mining field. Depending upon the area of application outlier detection is useful in a lot of domains, mainly including intrusion detection systems, credit cards as well as transaction fraud detection systems where any unusual activity is an outlier. Even the performance of athletes can be mapped using and analyzed by this concept.

Mainly outliers may be generated due to measurement impairments, rare normal events exhibiting entirely different characteristics, deliberate action etc. A lot many methods have been put forward by various research scholars and scientists and have been mainly classified into three types.

They are:-

1. Distance methods
2. Density methods
3. Statistical methods

DISTANCE BASED OUTLIERS

Each of the above methods have a lot of algorithms which have been proposed by scientists. However a search for a better and more efficient algorithm is always a dire need of the community that has more profound results on real world data as well synthetic datasets. The above methods do have such algorithms which have a complex implementation and utilize advanced data structures as well some of the most complex algorithms existent today. Depending on the type of data present different methods are used to calculate outliers, i.e., like statistical methods are used for calculating any unusual activity in the bank account of an individual, etc.

Now moving on to explain the first two methods with an algorithm to give you an insight into this topic I introduce the first method to calculate outliers. These techniques counts the number of pattern falling under the selected threshold distance r from a particular point x in the dataset. If the count is more than a preset number then x is considered as normal otherwise an outlier. The method popularly known as N Dot[1] calculates outliers by calculating a set of certain terms and then approaching the result. Let us introduce their method; the whole method concentrates around calculating Nearest Neighbor Factor (NNF). If Nearest Neighbor Factor of the point w.r.t majority of its neighbor is more than a threshold limit then the point is declared as a potential outlier.

Now let us introduce some basic terminologies :-

1. K Nearest Neighbor (knn) Set
2. Average knn distance
3. Nearest Neighbor Factor

K NEAREST NEIGHBOR (KNN) SET

Let D be a dataset of and x be a point in D.

For a natural number k and a distance function d, a set $Nnk(x) = \{q1 \in D | d(x,q1) < d(x,q2), q2 \in D\}$ is called knn of x if the following two conditions hold.

(1) $|Nnk| > k$ if q2 is not unique in D or $|Nnk|=k$ otherwise.
(2) $|Nnk \setminus Nq2|=k-1$, where Nq2 is the set of all q2 point(s).

AVERAGE KNN DISTANCE

Let Nnk be the knn of a point x in dataset D. Average knn distance of x is the average of distances between x and q belongs to Nnk ,i.e.,

$$\text{Average knn distance}(x) = \sum q\, d(x,q) \mid q \in Nnk)/|Nnk|$$

Average knn distance of a point x is the average of distances between x and its knn. If Average knn distance of x is less as compared to other point y, it indicates that x's neighborhood is more dense compared to that of y.

NEAREST NEIGHBOR FACTOR

Let x be a point in D and Nnk(x) be the knn of x .The NNF of x with respect to $q \in NNk(x)$ is the ratio of d(x,q) and Average knn distance of q.

NNF(x,q)=d(x,q)/Average knn distance(q).

HOW IT WORKS

Given a dataset D, it calculates knn and Average knn distance for all points in D. In the next step, it computes Nearest Neighbor Factor for all points in the dataset using the previously calculated knn and Average knn Distance. N Dot decides whether x is an outlier or not based on a voting mechanism. Votes are counted based on the generated NNF w.r.t respect to all its k nearest neighbors. Values with If $NNF(x,q \mid q \in Nnk(x))$ is more than a threshold value(=1.5 in most experiments), x is considered as an outlier with respect to q. Subsequently, a vote is counted for x being an outlier point. If the number of votes is at least 2/3 of the number of nearest neighbors then x is declared as an outlier point.

ALGORITHM FOR N DOT(D,K)

For each $x \in D$ do

calculate knn set Nnk(x) of x. calculate average distance of x.

end for

for each x Є D do

Count=0 /*Count counts the number of votes for x */ for each q Є Nnk(x) do

if NNF(x,q)>= 1.5 then Count=Count +1
end if end for
if Count >=2/3*|NNk(x)| then output x as an outlier in D
end if end for

COMPLEXITY

Time and space requirement of N DoT are as follows

1. Finding knn set and average knn distance of all point takes time of O(n^2) where n is size of the dataset and space requirement of step is O(n).
2. Deciding a point x to be outlier or not take time O(|NNK(x)|)=O(k) . For whole dataset step takes time of O(n*k)=O(n) as k is a small constant.

Now lets move on to the next classification of the outlier calculation methods, ie, density based methods. This has been a recent method in the study of outliers. Let me simply introduce a basic concept behind this method of basically how it works. The algorithm was proposed by Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng and Jörg Sander in the year 2000.

DENSITY BASED METHODS

In anomaly detection, the local outlier factor (LOF) is an algorithm for finding anomalous data points by measuring the local deviation of a given data point with respect to its neighbours.[1] As indicated the *local outlier factor* is based on the concept of a local density, where locality is given by k nearest neighbors, whose distance is used to estimate the density. By comparing the local density of an object to the local densities of its neighbors, one can identify regions of similar density, and points that have a substantially lower density than their neighbors. These are considered to be outliers. The local density is estimated by the typical distance at which a point can be "reached" from its neighbors. The definition of "reachability distance" used in LOF is an additional measure to produce more stable results within clusters. Now let me introduce this method in a little bit more detail. Let k-distance(A) be the distance of the object A to the k-th nearest neighbor. Note that the set of the k nearest neighbors includes all objects at this distance, which can in the case of a "tie" be more than k objects. We denote the set of k nearest neighbors as Nk(A).This distance is used to define what is called reachability distance

Reachability - distance k (A,B) = max {k-distance(B) , d(A,B)}

In words, the reachability distance of an object A from B is the true distance of the two objects, but at least the k-distance of B. Objects that belong to the k nearest neighbors of B are considered to be equally distant. The reason for this distance is to get more stable results. Note that this is not a distance in the mathematical definition, since it is not symmetric. (While it is a common mistake to always use the k-distance, this yields a slightly different method, referred to as Simplified-LOF) [3]

The local reachability density of an object A is defined by.

$$\mathrm{lrd}(A) := 1/\left(\frac{\sum_{B\in N_k(A)} \text{reachability-distance}_k(A,B)}{|N_k(A)|}\right)$$

Which is the quotient of the average reachability distance of the object A from its neighbors. Note that it is not the average reachability of the neighbors from A (which by definition would be the k- distance (A)), but the distance at which it can be "reached" from its neighbors. With duplicate points, this value can become infinite. The local reachability densities are then compared with

those of the neighbors using

$$\mathrm{LOF}_k(A) := \frac{\sum_{B \in N_k(A)} \frac{\mathrm{lrd}(B)}{\mathrm{lrd}(A)}}{|N_k(A)|} = \frac{\sum_{B \in N_k(A)} \mathrm{lrd}(B)}{|N_k(A)|} / \mathrm{lrd}(A)$$

Which is the average local reachability density of the neighbors divided by the objects own local reachability density. A value of approximately 1 indicates that the object is comparable to its neighbors (and thus not an outlier). A value below 1 indicates a denser region (which would be an inlier), while values significantly larger than 1 indicate outliers.

ADVANTAGES AND DISADVANTAGES

Due to the local approach, LOF is able to identify outliers in a data set that would not be outliers in another area of the data set. For example, a point at a "small" distance to a very dense cluster is an outlier, while a point within a sparse cluster might exhibit similar distances to its neighbors. While the geometric intuition of LOF is only applicable to low-dimensional vector spaces, the algorithm can be applied in any context dissimilarity function can be defined. It has experimentally been shown to work very well in numerous setups, often outperforming the competitors, for example in network intrusion detection.[4] The LOF family of methods can be easily generalized and then applied to various other problems, such as detecting outliers in geographic data, video streams or authorship networks.[3] However the resulting values are quotient-values and hard to interpret. A value of 1 or even less indicates a clear inlier, but there is no clear rule for when a point is an outlier. In one data set, a value of 1.1 may already be an outlier; in another dataset and parameterization (with strong local fluctuations) a value of 2 could still be an inlier. These differences can also occur within a dataset due to the locality of the method

CONCLUSION

In this paper the concept of outliers was introduced and its basic applications are mentioned. Then various categories for finding outliers were enumerated and two of them were explained in detail. Thus a light was thrown on the topic to highlight its advantages and importance to the data mining field.

REFERENCE

1. Breunig, M. M.; Kriegel, H.-P.; Ng, R. T.; Sander, J. (2000). "LOF: Identifying Density-based Local Outliers".
2. Neminath Hubbali; Bidyut Kr. Patra; Sukumar Nandi; NDoT: Nearest Neighbor Distance Based Outlier Detection Technique
3. Schubert, E.; Zimek, A.; Kriegel, H. -P. (2012). "Local outlier detection reconsidered: A generalized view on locality with applications to spatial, video, and network outlier detection"
4. Ar Lazarevic, Aysel Ozgur, Levent Ertoz, Jaideep Srivastava, Vipin Kumar (2003). "A comparative study of anomaly detection schemes in network intrusion detection".