Research Paper Engineering

DISTANCE-BASED OUTLIER DETECTION ON TIME SERIES DATA A STABILIZED AND RENOVATED TECHNIQUE

(Dr.)Maya Nayak	Professor and Head of Department, Information Technology, Orissa Engineering College, Bhubaneswar, DT- Khurda, Odisha, India
Prasannajit Dash	Assistant Professor, Department of Information Technology, Orissa Engineering College, Bhubaneswar, DT- Khurda, Odisha, India
Kalyanamayee Swain	Research Scholar, Department of Information Technology, Orissa Engineering College, Bhubaneswar, DT- Khurda, Odisha, India

ABSTRACT

Finding outlier as exceptions in a large that is multidimensional dataset is the goal of this paper. The identification of outliers discovers the truly unexpected knowledge in areas like network intrusion detection, credit card fraud, outfalls in stock market data. The existing methods for finding outliers in large datasets that can only works efficiently with two dimensions/ attributes of a test data set. From this paper the function of DB i.e. Distance – Based Outliers is covered. For usefulness of DB Outliers, we focus on the development of algorithms for computing such outliers. The outlier i.e. the distance-based one presenting two simple algorithms having each a complexity of the $O(kN^2)$, k being the dimensionality and N being the number of objects in the data set. So for this paper these algorithms readily support datasets with many more than two attributes. With the experimental results the nested loop algorithms which are best for k<=4.

KEYWORDS DB-outlier, complexity, dimension, attribute, object, tuple

1. INTRODUCTION

Discovering tasks on knowledge based are broadly classified into categories like dependency detection, class identification, class description and exception outlier detection. In correspond to the first three categories, the task applies to a large number of objects in the dataset. The task responds to the data mining (e.g. association rules, classification, data clustering etc.). The fourth category, in contrast, focuses on a very small percentage of data objects quite discarded as noise. From knowledge discovery point of view, the sample applications like the detection of credit card fraud and the monitoring of criminal activities in internet commerce applications like stock market data which is perhaps in monetary amount, type of purchase, timeframe, location – that may interest us, either for fraud detection for marketing reasons.

1.1 RELATED WORKS

While there is no formal definition of an outlier, anyway Hawkin's definition brings the spirit: "an outlier is an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism". For many years the outlier tests have been developed depending on the (i) data distribution, (ii) whether or not the distribution parameters like mean, standard deviation or variance, (iii) the number of expected outliers including the upper or lower outliers. While testing these outliers suffers from two serious problems, the first is that all the outliers are univariate (i.e. single attribute). This restriction creates havoc for multidimensional datasets. The second one is that all of them are distribution based. In some circumstances we do not know whether a particular attribute follows a normal distribution, a gamma distribution, etc., we need to perform an extensive testing to find a distribution that fits the attribute. However, in practice, the mathematical computation of k-dimensional layers heavily depends on the computation of k-dimensional convex hulls. Because the lower bound complexity of computing a k-dimensional convex hull is $\Omega(N^{\lfloor k/2 \rfloor})$. So depth-based methods are not expected to be more practical for more than 4 dimensional for large datasets.

1.2 DISTANCE-BASED OUTLIERS AND ITS CONTRIBUTION TO THE PAPER

The distance based outlier[1] is defined as an object P in a dataset R is DB (p,D) is an outlier if at least fraction p of the total objects in R lies greater than the euclidean distance D from P. After standardization, or without standardization in certain applications, the dissimilarity (or similarity) between the objects is typically computed based on the distance between each pair of objects. The most popular distance measure is Euclidean distance which is defined as

$$d(i,j)=\sqrt{(x_{i1}-x_{j1})^2 + (x_{i2}-x_{j2})^2+\ldots\ldots+(x_{in}-x_{jn})^2} \quad \text{(Eq.1)}$$

The Euclidean distance satisfies the following mathematic requirements of a distance function:

(a) $d(i,j) >= 0$: Distance is a nonnegative number

(b) $d(i,i) = 0$: Distance of an object to itself is 0

(c) $d(i,j) = d(j,i)$: Distance is a symmetric function

(d) $d(i,j) <= d(i,h) + d(h,j)$: Going directly from object I to object j in space is no more than making a detour over any other object h

Here the term DB(p,D) – outlier [4] means the shorthand notation for Distance – Based outlier while detecting using the parameters p and D. This is nothing but the Hawkin's definition. Only to those situations where the observed distributions does not fit to the standard distribution. It is very important that it is well defined for k-dimensional datasets for any value of k. Anyway, the DB-outliers[3] are not restricted computationally to small values of k. So the DB-outliers go beyond the data space and rely on the computation of distance values based on a metric distance function.

2. PROPERTIES OF DB-OUTLIERS

Definition 1 : DB(p,D) unifies or generalizes another definition Def for outliers, if there exist specific values p0, D0 such that

the object O is an outlier according to Def if O is a DB(p0,D0) –outlier. Outliers can be considered in a normal distribution[2] to be the observations that lie 3 or more standard deviations (i.e. >= 3σ) from the mean μ.

Definition 2 : Define Defnormal as 't' is an outlier in a normal distribution with mean μ and standard deviation σ if |(t- μ) / σ| >= 3. DB(p,D) unifies Defnormal with p0 = 0.9988 and D0 = 0.13 σ so that, t is an outlier according to Defnormal if t is a DB(0.9988, 0.13 σ) as an outlier. If the value 3σ is changed to some other value , such as 4 σ , the above definition should be modified with the related p0 and D0 to show that DB(p,D) still unifies the new definition of Defnormal. The principle of using a tail to indentify outliers can also be applied to a Poisson distribution.

Definition 3: Define Defnormal shows that t is an outlier in a Poisson distribution with parameter μ=3.0 if t>=8. DB(p,D) unifies DefPoisson with p0 = 0.9892 and D0 =1.

3. A NESTED LOOP ALGORITHM FOR FINDING ALL DB(p,D)-OUTLIERS

The nested loop algorithm i.e. algorithm NL shown below uses having nested loop design in a block oriented nature. Let us assume that the total buffer size of B% of the total dataset size, the algorithm divides the full buffer space into equally two halves as they are mentioned as the first array and second arrays. The whole dataset got divided into arrays and directly computes the distance between each pair of objects or tuples. For each object t in the first array, a count of its D-neighbors is maintained. The process of counting gets stopped for a particular tuple whenever the number of D-neighbors exceeds M.

PSEUDO-CODE FOR ALGORITHM NL (NESTED LOOP)

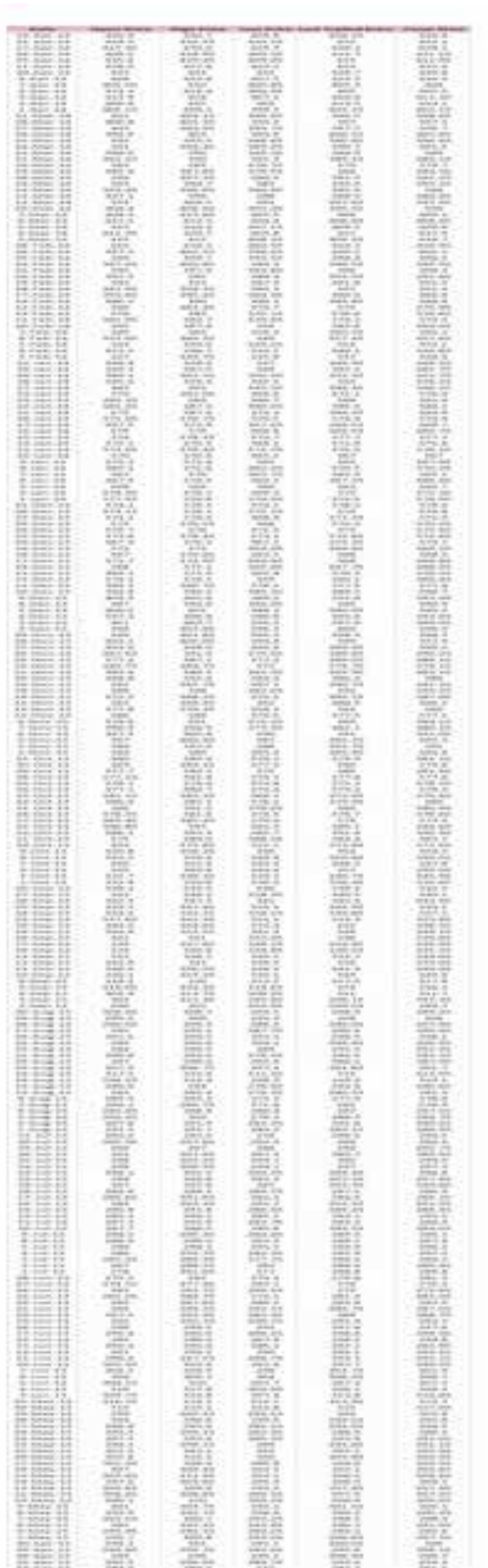
1. The size that is B/2% of the dataset will be filled with a block having tuples from T.
2. In the first array, for each tuple ti , do the following :

(a) Initialize the counter i.e. counti to 0
(b) In first array, check for each tuple i.e. tj whether the dist(ti, tj) <= D.
(c) If it happens, then increment the counti>M, where we could mark ti as a non-outlier and further proceed to the next ti
(d) While the blocks remain to be compared to the first array, then fill the second array with another block. For each unmarked tuple ti in the first array where for each tuple tj in the second array if dist(ti,tj) <= D.
(e) If it happens, then increment the counti by 1. So if counti > M, mark ti as a non outlier and proceed to next ti.
(f) For each unmarked tuple ti in the first array, report ti as an outlier.
(g) In the second array if it has served as the first array anytime before, stop; otherwise , swap the names of the first and second arrays and go to step(b).

4. APPLICATION OF STOCK MARKET DATA AS TIME SERIES

Sources:http://nseindia.com/live_market/dynaContent/live_watch/get_qoute/GetCoutejsp?symbol=TATAMOTORS

Figure1: stock market data in national stock exchange, Mumbai for tata motors



As the above test dataset for the National Stock Exchange data of Tata Motors for a quarter. These data are having the date of transaction, opening Price of the day, high and low price for a day and the closing price for a day. As the data is collected for three months, it is the presentation of time-series data as the outliers are to be observed for multidimensional data i.e. attributes around five in the data set.

The opening price and closing price are measured in terms of euclidean distance row wise as it produces the euclidean distance measurement for data in columns like open price and close price. The sum of euclidean distances (row-wise) between the data in open price and close price. So the average euclidean distance is calculated for the total objects means rows or records(tuples). Here the fraction of object is calculated as taking the value of 'p' as 0.9. As the total rows or tuples are 249 in the dataset, hence the fraction of the object i.e. 'f' is calculated as (249×(1-p)) which produces 24.9. Here the one dimensional matrix i.e. C1 of rank 249×1 where it contains all euclidean distances (row or tuple wise) between Open and Close Prices. While iterating through the records from 1 to 249, if the the value in C1 is less than the average euclidean distance then if the count of the object is greater than the fraction of the object, then it is not an outlier otherwise it is an outlier as shown in Figure2 and Figure 3 respectively. If the distance is greater than the average euclidean distance, then also it is counted as outlier. Then the Nested Loop(NL) is applied to calculate the non-outliers and outliers in respect to the date.

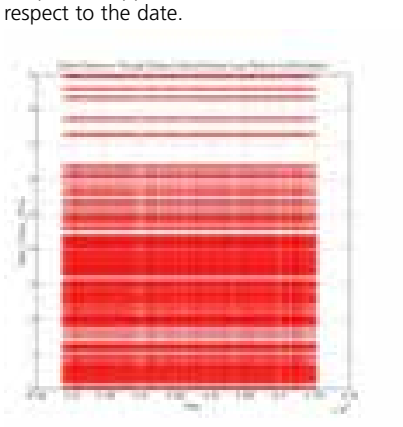


Figure 2 : Open_Close Non Outliers in respect to date

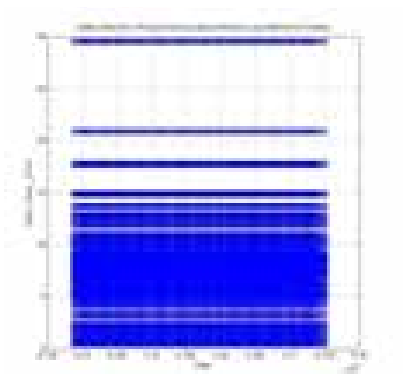


Figure 3 : Open_Close Outliers in respect to date

Simmilarly, The High price and Low price are measured in terms of euclidean distance row wise as it produces the euclidean distance measurement for columns like high price and low price. The sum of euclidean distances (row-wise) between the data in open price and close price. So the average euclidean distance is calculated for the total objects means rows or records(tuples). Here the fraction of object is calculated as taking the value of 'p' as 0.9. As the total rows or tuples are 249 in the dataset, hence the fraction of the object i.e. 'f' is calculated as (249×(1-p)) which produces 24.9. Here the one dimensional matrix i.e. C2 of rank 249×1 where it contains all euclidean distances (row or tuple wise) between High and Low Prices. While iterting through the records from 1 to 249, if the the value in C2 is less than the average euclidean distance then if the count of the object is greater than the fraction of the object, then it is not an outlier otherwise it is an outlier as shown in Figure4 and Figure 5 respectively. If the distance is greater than the average euclidean distance, then also it is counted as outlierThen the Nested Loop(NL) is applied to calculate the non-outliers and outliers in respect to the date.

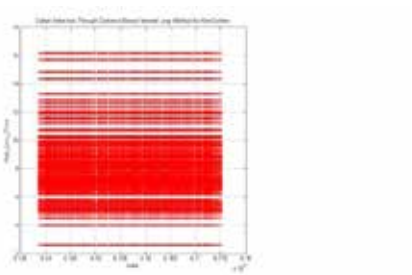


Figure 4: High_Low Non Outliers in respect to date

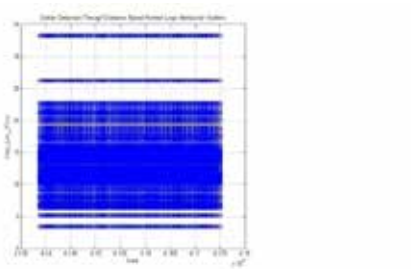


Figure 5 : High_Low Outleirs in respect to date

In Figure 2 and 4, the red colour points are the non-outliers as per the date against the equiclidean distance difference between Open and Close Price column data. The x-axis is the dates and the y is the distance differences. Simmlilarly in figure 3 and 5, the blue colour points are the outliers.

5. CONCLUSION

Algorithm Nested Loop(NL) avoids the explicit construction of any indexing structure, and its complexity is $O(kN^2)$. Compared to a tuple-by-tuple brute force algorithm that pays no attention to I/O's, Algorithm Nested Loop(NL) is always superior because it tries to minimize I/O's.

REFERENCES

[1]E. Knorr and et al. (2000), "Distance-based outliers: Algorithms and applications". VLDB Journal | [2] E. Knorr and R. Ng.(1997), "A unified notion of outliers Properties and computation". In ACM SIGKDD. | [3] E. Knorr and R. Ng.(1999), "Finding intentional knowledge of distance-based outliers". In VLDB. | [4] E. Knorr and R. T. Ng.(1998), "Algorithms for mining distance-based outliers in large datasets". In Proc. Int'lConf. on VLDB | [5] V. Barnett and T. Lewis (1994), "Outliers in Statistical Data". John Wiley. | [6] Han and Kamber (2007), "Data Mining: Concepts and Techniques Morgan Kaufmann publications" | [7] George Marakas, Data Warehousing, Data Mining and Visualisation, Pearson publications | [8] Zuriana A. B., Rosmayati M., Akbar A., Mustafa M. D.,(2006) "A Comparative Study for Outlier Detection Techniques in Data Mining" CIS. | [9] Aggarwal, C. C., Yu, S. P.,(2005) "An effective and efficient algorithm for high-dimensional outlier detection", The VLDB Journal, vol. 14, pp. 211–221.