Outlier Detection Technique in Data Mining: A Research Perspective

[1] M. O. Mansur, [2] Mohd. Noor Md. Sap

Faculty of Computer Science and Information Systems

Universiti Teknologi Malaysia

[1] mansurukm4@gmail.com, [2] mohdnoor@fsksm.utm.my

Abstract:

While the field of data mining has been studied extensively, most of the work has concentrated on discovery of patterns. Outlier detection as a branch of data mining has many important applications, and deserves more attention from data mining community. Most methods in the early work that detects outliers independently have been developed in field of Statistics. Finding ,removing and detecting outliers is very important in data mining, for example error in large databases can be extremely common, so an important property of a data mining algorithm is robustness with respect to outliers in the database. Most sophisticated methods in data mining address this problem to some extent, but not fully, and can be improved by addressing the problem more directly. The identification of outliers can lead to the discovery of unexpected knowledge in areas such as credit card fraud detection, calling card fraud detection, discovering criminal behaviors, discovering computer intrusion, etc. In this paper we will explain the first part of our research, which is focused on outlier identification and provide a description of why an identified outlier exceptional, based on Distance-Based outlier detection and Density-Based outlier detection.

Keywords

Outlier detection, Distance-based, Density-based, Data Mining

1. Introduction

Outlier detection is an important branch in data mining, which is the discovery of data that deviate a lot from other data patterns. D.Hawkins [1], gives definition to outlier as: An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism. There are many studies have been conducted on outlier detection for large datasets. A lot of work has been done in this area of research which is detecting outliers. The early work is based on statistics ([2], [3]), and assume that a priori knowledge of distribution is known. Most of tests depended on the distribution whether or not the distribution parameters are known. In other related area dealing with detecting outliers is clustering algorithms where outliers are objects not located in clusters of a dataset, and these algorithms generate outliers as by product. Recently, researchers have proposed distance-based, density-based and connectivity-based outlier detection methods .The advantage of these methods is that, they do not have any priori knowledge about the data distribution. In this paper we will explain these methods and their drawbacks.

2. Related work:

Most of previous studies on outlier detection have been done in field of Statistics. These two methods are:

A) Distribution – based : this method is depending on :

   i) The data distribution, where a standard distribution like Normal, Poisson, etc is used to fit the data best. In this case outliers are defined based on the probability distribution.
   ii) The parameters of the distribution (Mean and Variance) are known or not.
   iii) The number and the type of expected outliers (upper and lower outliers in an ordered sample)[2].

   This method has two main problems. First, all most of the distributions used are univariate (e.g., single attribute), they are some tests that are multivariate (e.g., multivariate normal outliers) this restriction makes them unsuitable for multidimensional datasets. Second, all of them are distribution-based. In many situations and applications, it is difficult to know whether a particular attribute follows a normal distribution, a gamma distribution, etc. The underlying distribution is unknown, so in this case we have to perform very extensive testing to find a distribution that fits that attribute. Fitting the data with standard distribution is costly.

B) Depth – based: Each data object is represented as a point in a k-d space, and is assigned as depth. With respect to outlier detection, outliers are more like to be data objects with smaller depth. There are many methods of depth have been proposed ([11], [12]). These methods avoid the serious problem of distribution fitting, and conceptually allow multidimensional data objects to be processed, however, in practice; the computation of *k*-dimensional layers relies on the computation of *k*-dimensional convex hulls. This is because the lower bound complexity of computing a *k*–dimensional convex hull is $\Omega\left(N^{k/2}\right)$. In fact, depth-based methods give acceptable performance for $k \leq 2$.

C) Clustering approaches: Clustering–based approaches such as(CLARANS [13], DBSCAN [14] ,BIRCH[15], CURE[16]), are amid to partitioning data into a number of clustering, where each data point can be assigned a degree of membership to each of the clusters. With respect to outlier detection, they consider outlier but in the way that don't interfere with the clustering process.

3. Distance-based outlier detection scheme: ([5], [9], [10])

The notion of outliers in DB (p, D) - outlier:

An object O in a dataset T is a DB (p, D)-outlier if at least fraction p of the objects in T lies greater than distance D from O.

Distance – Based outlier detection using parameters p and D.It is suitable for situations where the observed distribution dose not fit any standard distribution and the very important about

Distance-Based outlier, is that it is well defined for *k*-dimensional datasets for any value of *k*. There are two simple algorithms, both having a complexity of O (k N^2), *k* is dimensionality and N is the number of objects in the dataset. These algorithms readily support datasets with many more than two attributes, and they also present optimized cell-based algorithm that has a complexity that a linear wrt N, but exponential wrt *k*, as well as, for datasets are mainly disk-resident, they have another version of the cell-based algorithm.

DB(p,D)-outlier detected using parameters p and D.The user has to choose suitable values for p and D to defined the strength of the outliers requested which may involve trail and error and numerous iterations. In this case is quite difficult to choose suitable values for p and D so it will be costly and it dose not provide a ranking for the outliers. For instance a point with very few neighboring points within a distance D can be regarded in some sense as being a stronger outlier than a point with more neighboring within a distance D.Cell-based algorithm whose complexity is linear in the size of the database dose not scale for higher number of dimensions (e.g.,5). S. Ramaswamy [3] presents a new definition for outliers and developed algorithms for mining outliers which the user dose not need to specify the distance parameter D.Instead, it is based on the distance of the k^{th} nearest neighbor of a point(special case of DB(p,D)-outlier).But ,they have the same weakness which is they are not powerful enough to cope with certain process with different densities in data clusters[17].

4. Density-based local outlier detection scheme: [7]

It assigns to each object a degree to be an outlier. This degree is called the local outlier factor (LOF) of an object. It is local in that, the degree depends on how isolated the object is with respect to the surrounding neighborhood. In LOF algorithm, outliers are data objects with high LOF values whereas data objects with low LOF values are likely to be normal with respect to their neighborhood. High LOF is an indication of low-density neighborhood and hence high potential of being outlier. We briefly review the steps of computing LOF value of an object p in a dataset D:

Let D be a database, p, q, o some objects in D and k be a positive integer. The distance function (Euclidean distance) d (q, p) to denote the distance between objects p and q.

Step 1: Computing (k- *distance of p*):

The k-distance of p is the distance d (p, o) between p and o such that:

I - For at least k objects $o' \in D \setminus \{p\}$ it holds that $d(p, o') \leq d(p,o)$, and

II -For at most (k - 1) objects $o' \in D \setminus \{p\}$ it holds that $d(p, o') < d(p, o)$.

k-distance (p) provides a measure of the density around the object p, when k-distance of p is small meaning that the area around p is dense and vice versa.

Step 2: Finding (k-*distance neighborhood of p*):

The k-distance neighborhood of p contains every object whose distance for p is not greater than the k-distance.

$N_{k-distance}(p) = \{q \in D \setminus \{p\} \mid d(p, q) \leq \text{k-distance}(p)\}$.

Step 3: Computing (reachability *distance of p wrt object o*):

The reachability distance of object p with respect to object o is

Reach-dist-k (p, o) = max {k-distance (o), d (p, o)}.

Step 4: Computing (the *local reachability density of p*):

The local reachability density of an object p is the inverse of the average reachability distance from the k-nearest-neighbors of p:

$$lrd_k(p) = 1 / \left[\frac{\sum_{o \in N_{k(p)}} reach-dist_k(p,o)}{|N_{k(p)}(p)|}\right]$$

Essentially, the local reachability density of an object p is an estimation of the density at point p by analyzing the k-distance of the objects in $N_k(p)$. The local reachability density of p just the reciprocal of the average distance between p and the objects in its k-neighborhood. Based on local reachability density, the local outlier factor can be defined as follows.

Step 5: *Local outlier factor of p*

$$LOF_k(p) = \frac{\sum_{o \in N_k(p)} \frac{lrd_k(o)}{lrd_k(p)}}{|N_k(p)|}$$

The local outlier factor is a ratio that determines whether or not an object is outlier with respect to its neighborhood .$LOF_k(p)$ is the average of the ratios of the local reachability density of p and that of p's k-nearest-neighbors.

4.1 – Limitations of the LOF Algorithm [18]

The major drawback of LOF algorithm lies in computing reachability distances defined as [*reach-dist-k (p, o) = max (k-distance (o), d (p, o))*] .Computing reachability distance of p involves computing distances of all objects within p's neighborhood, and each compared with the k-distance of that neighborhood which is very expensive when MinPts is large. Secondly, LOF has to be computed for every object before the few outliers are detected. This is not a desirable exercise since outliers constitute only small fraction of the entire dataset.

5 - LSC-Mine: Algorithm for Mining Local Outliers:[18]

It based on the distance of an object and those of its k-nearest neighbors without computing reachability distances and local reachability densities and pruned the data objects which they are not possible to be outliers to reduce the number of computations.

Definition 1 Compute the Local sparsity ratio of an object p

$$\text{lsr}_k(p) = \frac{|N_k(p)|}{\sum_{o \in N_k(p)} distofN_k(p)}$$

where:

distfN (p): consists of actual distances of objects in k-distance neighborhood of p.

$\text{lsr}_k(p)$: measures the concentration of objects around an object p.

Definition 2 Computing the pruning factor (Pf)

$$\text{Pf} = \frac{\sum |N_k(p)|}{\sum \sum_{o \in N_k(p)} distofN_k(p)}$$

Any object with a local sparsity ratio less than Pf is removed because it cannot be as outlier candidate.

Definition 3 Computing the Local sparsity Coefficient of p

$$\text{LSC}_k(p) = \frac{\sum_{o \in N_k(p)} \frac{lsr_k(o)}{lsr_k(p)}}{|N_k(p)|}$$

Where a high $\text{LSC}_k(p)$ means that the neighborhood around an object crowded hence it has higher potential to be an outlier.

LSC-Mine algorithm will compute for each object:

1) k-distance.
2) k-distance neighborhood.
3) local sparsity ratio.
4) pruning factor.
5) candidate set that is not to be pruned.
6) LSC of objects.
7) rank objects with the highest LSC as strongest outliers.

6 – Connectivity-Based outlier detection scheme: [19]

The connectivity based outlier detection scheme is based on the idea of differentiating "low density" from "isolativity". While low density normally refers to the fact that the numbers of points in the "close" neighborhood of an object is "connected" to other objects. As a result, isolation can imply low density, but the other direction is not always true.

Definition 1 : Let D is a data set. An interpretation of D is a partition $D = D_o + D_n$, where D_o and D_n are the outlier set and non outlier set, respectively.

Definition 2 : Let $P,Q \subseteq D$, $P \cap Q = \emptyset$ and $P,Q \neq \emptyset$. We define

dist (P, Q) = {dist(x,y) : $x \in P$ & $y \in Q$}, and call dist(P,Q) the distance between P and Q . For any given $q \in Q$, we say that q is the nearest neighbor of P and Q if there is a $p \in P$ such that dist (p, q) = dist (P,Q).

Definition 3 : Let G = $\{p_1, p_2, \ldots\ldots, p_r\}$ be a set of D. A set-based nearest path , or SBN-path, from p_1 on G is a sequence ‹ $p_1, p_2, \ldots\ldots, p_r$ › such that for all $1 \leq i \leq r - 1$, p_{i+1} is a nearest neighbor of a set

$\{p_1, p_2, \ldots\ldots\ldots, p_i\}$ in $\{p_{i+1}, \ldots\ldots\ldots, p_r\}$.

In the above , if the nearest neighbor is not unique, we can impose a predefined order among the neighbors to break the tie . Thus an SBN-path is uniquely determined.

Definition 4 : Let s = ‹ $p_1, p_2, \ldots\ldots\ldots, p_r$ › be an SBN-path. A set-based nearest trail, denoted as SBN-trail, with respect to s is a sequence ‹ $e_1, e_2, \ldots\ldots\ldots, e_{r-1}$ › such that for all $1 \leq$ r-1 $\leq$ r-1, $e_i = (o_i, p_{i+1})$ where $o_i \in \{p_1, p_2, \ldots\ldots\ldots, p_i\}$ and dist (e_i) = dist (o_i, p_{i+1}) = dist($\{p_1, p_2, \ldots\ldots\ldots, p_i\}$, $\{p_{i+1}, \ldots\ldots\ldots, p_r\}$).

Again if o_i is not uniquely determined. We should break the tie by a predefined order. Thus the SBN-trail is unique for any sbn-path.

Definition 5: Let G = $\{p_1, p_2, \ldots\ldots\ldots, p_r\}$ be a subset of D. Let s = ‹ $p_1, p_2, \ldots\ldots\ldots, p_r$ › be an SBN-path from p_1 and e = ‹ $e_1, e_2, \ldots, e_{r-1}$ › be a SBN-trail with respect to s. The average chaining distance from p_1 to G – $\{p_1\}$, denoted by ac-dist $_G(p_1)$,is defined as

$$\text{ac-dist}_G(p_1) = \sum_{i=1}^{r-1} \frac{2(r-i)}{r(r-1)} . \text{dist}(e_i)$$

Definition 6 : Let $p \in D$ and k be a positive integer. The connectivity-based outlier factor (COF) at p with respect to its k-neighborhood is defined as

$$\text{COF}_k(p) = \frac{|N_k(p)|.ac - dist_{n_{k(p)}}(p)}{\sum_{o \in n_k(p)} ac - dist_{n_{k(o)}}(o)}$$

A threshold on COF can be set to define outliers.

The connectivity-based outlier factor at p is the ratio of the average chaining distance from p to $N_k(P)$ and the average of the average chaining distance from p's k-distance neighbors to their own k-distance neighbors. It indicates how far away a point shifts from a pattern.

7 – Outlier Detection Integrating Semantic Knowledge (SOF) [4] :

The basic idea is that, the records with the same class label should be similar with each other from the semantic knowledge that the people in the same group should have similar ideas.

Semantic outlier: A semantic outlier is a data point, which behaves differently with other points in the same class.

7.1 - Formal Definition of Semantic Outlier:

Definition 1 Let $A_1, \ldots\ldots, A_m$ be a set of attributes with domains $D_1, \ldots\ldots, D_m$ respectively. Let the set D be a set of records where each record t: t $\in$ $D_1 x \ldots\ldots x D_m$. The results of clustering algorithm executed on D is denoted as: C =

$\{C_1,\ldots,C_k\}$,where k is the number of cluster.

Definition 2 Let CL is an additional attribute for D, which distinguishes the class of records and has the set of different attributes values $\{cl_1,\ldots\ldots,cl_p\}$. The output C = $\{C_1,\ldots, C_k\}$,just as described in Definition 1 will be produced if a clustering algorithm is executed on D, and define $Pr(cl_i \mid D)$ and $Pr(cl_i \mid C_j)$ as the frequency of cl_i in D and frequency of cl_i in C_j

$$Pr(cl_i \mid D) = \frac{\left|\{t \mid t.CL = cl_i, t \in D\}\right|}{|D|}$$

$$Pr(cl_i \mid C_j) = \frac{\left|\{t \mid t.CL = cl_i, t \in C_j\}\right|}{|C_j|}$$

Based on running a clustering algorithm on the data set D ,it is expected that the records in every output cluster should be identified with the same class label, those records whose class labels are different from that of the majority of the cluster as outliers. If the value of $Pr(cl_i \mid C_j)$ small, it indicates that records with label cl_i are consider to be outliers.

Definition 3 *Similarity*

The similarity between R(set of records) and t (record) is

$$sim(t,R) = \frac{\sum_{i=1}^{|R|} similarity(t,T_i)}{|R|}$$

where $\forall T_i \in R$.

Definition 4 *Semantic outlier factor of a record t*

Supposes the clustering algorithm assign t to C_k and the class value of t is cl_i. And R is the subset of D with class value cl_i. The semantic outlier factor of a record t is :

$$SOF(t) = \frac{Pr(cl_i \mid C_k) * sim(t,R)}{Pr(cl_i \mid D)}$$

The measure sim (t, R) describes how records differ from others in the same class.

The Algorithm FindSOF has three parts:

I-Clustering the dataset using (Squeezer algorithm (5) which good for categorical data).

II - Updating counters.

III- Computing the values of SOF.

The value of SOF computed as

$$SOF(t) = \left(\frac{(cl_i,C_j)\_counter}{|C_j|}\right) * sim(t,(cl_i,D)) * \left(\frac{|D|}{(cl_i,D)\_counter}\right)$$

The concept is very interesting and useful but it works mostly on categorical data.

8- Discussion andConclusions

Finding outliers is an important task in data mining. Outlier detection as a branch of data mining has many important applications and deserves more attention from data mining community. The recently researches work on outlier detection is either distance-based, density-based or connectivity-based outlier detection .The advantage is, they do not have any priori knowledge about the data distribution.

We address some weakness from our review that:

I - Statistical-based depend on the parameters of the distribution (Mean and Variance) which are unknown.

II - Distance-base, DB (p, D) depends on parameters p (which always close to 1), and the value of D the user has to choose (try by error).

III -Density-base local outlier depends on the value of MinPts, if the value of MinPts is large; LOF has to be computed for every object before the few outliers are detected. This is not a desirable exercise since outliers constitute only small fraction of the entire dataset.

IV- Outlier Detection Integrating Semantic Knowledge (SOF)
It works mostly on categorical data.

On the basis of various issues addressed in this paper, the following conclusions can draw:

A)-Most of the methods still distance-based.

B)-Most of them have high time complexity.

C)-Most of them depend on clustering methods to detect the outliers.

We conclude from our review of existing outlier detection schemes and clustering methods that they all suffer from the fact that they either depend on prespecified values for the scale parameters or the fraction of inliers. These two main issues make them very sensitive to initialization; or they have to perform a quasi-exhaustive search on these parameters, which makes them require a very high computational cost.

References :

[1] D.Hawkins: "Identification of outliers". Chapman and Hall, London. 1980.

[2] V. Barnett. Lewis. "Outliers in Statistical Data". John Wiley & Sons, 1994.

[3] S. Ramaswamy,R. Rastogi,K.Shim "Efficient algorithms for mining outliers from large data sets". Proceedings of the International Conference on Management of Data, Dallas, Texas.2000.

[4] Z. He, S. Deng, X. Xu, "Outlier Detection Integrating Semantic Knowledge". WAIM2002,LNCS 2419, pp 126-131.2002

[5] E.M. Knorr,R.T.Ng, "Algorithms for Mining Distance-Based Outliers in Large Dataset". In Proc. of the 24th VLDB Conference, New York,USA,1988.

[6] E.M. Knorr,R.T.Ng, "A unified approach for mining outliers".In proc. 7th CASCON, pp 236-248,1997

[7] M. Breunig,HP. Kriegel, R.T. Ng,J. Sander, "LOF: Identifying Density-Based Local Outliers".Proc. of ACM SIGMOD 2000 Int. Conf. on Management of Data,Dallas, TX 2000

[8] E.M. Knorr,R.T.Ng, "Finding intensional knowledge of distance-based outliers". In Proc. VLDB, pp. 211-222,1999.

[9] E.M. Knorr,R.T.Ng and V. Tueakov, " Diatance-Based outliers:Algorithms and Applications". In Proc.VLDB Journal,8(3):237-253.2000

[10] E.M. Knorr,R.T.Ng, and R.H.Zamar, "Robust space transformations for distance-based operations". In Proc. ACM SIGKDD, pp 126-135 , 2001

[11] T. Johnson, I. Kwok. R. Ng. "Fast computation of 2-Dimensional Depth Contours", Proc. 4th Int. Conf. on Knowledge Discovery and Data Mining, New York, NY, AAAI Press, pp. 224-228,1998

[12] I. Ruts, P. Rousseeuw, "Computing Depth Contours of Bivariate Point Clouds". Journal of Computational Statistics and Data Analysis,23, pp. 153-168, 1996

[13] T. Raymond, Ng, Han Jiawei. "Efficient and effective clustering methods for spatial data mining". In Proc. VLDB, Chile,1994

[14] M.Ester, H. Peter, X.Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Proceeding of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR, pp. 226-231, 1996.

[15] T.Zhang, R. Raghu, M. Livny, "BIRCH: An efficient data clustering method for very large databases".In Proc. ACM SIGMOD, Conf. Management of Data. Pp.103-114. Montreal,Canada,Tune,1996.

[16] S.Guba,R.Rajeev,K.Shim "CURE: An efficient clustering algorithm for large databases". In Proc. ACM SIGMOD Conf. Management of Data, June, 1998.

[17] Z.Chen,a.Fu,J.Tang, "On Complementarity of Cluster and Outlier Detection Schemes".Springer-Verlag,LNCS 2737, pp.234-243,2003 .

[18] M.Agyemang., Ezeife, C.I. "LSC-Mine: Algorithm for Mining Local Outliers".15th Information Resources Management Association (IRMA) International Conference, New Orleans, Louisiana, USA, May 23-26, 2004

[19] J. Tang, Z.Chen, A.Fu, D.Cheung, "A Robust Outlier Detection Scheme in Large Data Sets", PAKDD, 2002.