Outlier Detection Methods--- An Analysis

Pragyan Paramita Das, Maya Nayak
Asst. Professor, Dept. of I.T. Orissa Engineering College
Vice Principal and H.O.D., Dept. of I.T. Orissa Engineering College

Abstract

An outlier is an extreme observation that is considerably dissimilar from the rest of the objects. The detection of outlier is helpful in many applications such as data cleaning, network intrusion, credit card fraud detection, telecom fraud detection, customer segmentation, medical analysis etc. Outliers behave very differently from the rest of the observations in the dataset. Outliers are mostly removed to improve the accuracy of the predictions. But, the presence of an outlier can have certain meaning also. In our work we compare detection of outlier techniques based on statistical method, density based method, distance based method and deviation based.

Keywords

Outlier detection, statistical method, density based method, deviation based method, distance based method, artificial intelligence, fuzzy logic, neural network.

"1" Introduction

In literature different definitions of outlier are given:

- "An outlier is an observation that deviates so much from other observations as to arouse suspicions that is was generated by a different mechanism " (Hawkins, 1980).

- "An outlier is an observation which appear to be inconsistent with the remainder of the dataset" (Barnet & Lewis, 1994).

- "An outlier in a set of data is an observation or a point that is considerably dissimilar or inconsistent with the remainder of the data" (Ramasmawy at al., 2000).

Outlier analysis has assumed importance due to the fact that an outlier represents a critical component of information in a data set. In a credit card transaction an outlier may mean credit card theft or misuse. In public health an outlier may frequently mean a new disease in the patient. Similarly in a computer network it could mean a hacked computer and a compromised system. Therefore, detecting outlier is an important data mining task.

Some of the prominent causes for outliers are mentioned below:

1. Human error: Outliers may exist due to a faulty reporting system or human error.
2. Environmental change: A new buying pattern amongst the customers or change in the nature of the environment itself can cause the presence of outliers.
3. Error in instrument: Defects in the instruments which are used to measure and report may be a cause for outliers.

4. Malicious activity: As already mentioned above a credit card fraud, hacking into a network system may be a cause for outliers.

1.1. Statement of the Problem

As already stated an outlier is an observation that does not conform to normal behaviour. But defining a normal behaviour is very challenging. Some of the difficulties that are encountered in the process are:

1. A normal region which will encompass all possible normal behaviour is difficult.
2. A normal region which is defined at present may not be normal in the future due to evolution of data.
3. In many cases of malicious behaviour the hacker often disguises the hacking as normal behaviour causing difficulty in identification.
4. For those data which lie at the bordering area between normal and outlier region represent difficulty in classification.
5. "Noise" in data is often confused with outliers.

Thus, there are numerous methods for the detection of outliers which have been explored in disciplines like data mining, machine learning and statistics.

1.2. Organization of this paper

In Section 1 we have described the complexity of the problem as well as the type of outliers.In Section 2 we have identified the four basic methods for the detection of outliers. In Section 3 we describe other methods for the detection of outlier as well as comparison of the methods and the conclusion.

1.3. Literature Survey

Outlier detection techniques based on statistical and machine learning techniques have been attempted by Hodge and Austin [2004]. Petrovskiy [2003] presented data mining techniques for the detection of outliers. Markou and Singh [2003] used neural networks for the detection of outliers. Lazarevic et al. [2003] used network intrusion detection techniques. Forrest et al. [1999], Snyder [2001] and Dasgupta and Nino [2000] have developed techniques for system call intrusion. Tang et al. [2006] have provided analysis and unification of many distance based outlier detection techniques. All these efforts have basically focussed on a particular subset of the existing techniques. Our paper will provide a comprehensive review of four popular methods for outlier detection.

1.4. Types of outliers

Type I Outliers.

In a particular dataset an individual outlying instance is termed as a Type I outlier[1]. That single point of data is an outlier because of its attribute values which is inconsistent with values taken by normal instances. Many of the existing outlier detection scheme focus on this single outlier. These techniques analyze the relationship of this single point of data with regard to the rest of the points in the dataset. For example, in credit card data or medical data each data represents a single transaction or to a single patient.

Type II Outliers.

These outliers are caused because of the occurrence of an individual data instance in a

specific context in the given data. These outliers are also individual data instances. But these type II outliers are in the context of a particular dataset especially in relation to its structure and problem formulation.

Type III Outliers.

These are not individual observations but rather are an entire subset of the entire dataset which are outliers. Their occurrence together constitute an anomalous formulation. They are usually meaningful when the data has a sequential nature. These outliers are either subgraphs or subsets occurring in the data.

There are four basic methods for the detection of outliers. They are the statistical method, deviation method, density method and the distance method. Each of these methods is explained below in some detail.

"2" Methods of Outlier Detection

2.1 Statistical method

The need for outlier detection was experienced by statisticians in as early as 19th century [Edgeworth 1887]. The presence of outlying or discordant observations in the data biased the statistical analysis being performed on the data. This led to the notion of accommodation or removal of outliers in different statistical techniques. These techniques which were developed to accommodate or ignore outliers while performing some other form of statistical analysis gave rise to exact statistical methods which are meant to detect outliers in a given data set.[2]

An outlier is an observation which is suspected of being partially or wholly irrelevant because it is not generated by the stochastic model assumed [Anscombe and Guttman 1960]. This is the underlying assumption of any statistical method. Thus the statistical detection technique first determines the probability distribution function of the dataset and then tests if the particular datapoint is generated by that model or not. These statistical models are usually model based techniques i.e. they first capture the distribution of the data and then evaluate how well the data instance fits the model. If the probability of that particular data point to be generated by this model is low, then this point is deemed to be an outlier.

The methodologies used in these methods are variants of standard distribution models like normal, Poisson. Those data points which deviate from these models are flagged as outliers. The most popular amongst these methods is the Gaussian function. A Gaussian mixture model (GMM) was proposed by (Roberts & Tarassenko, 1995). This was based on based on how much a data point deviates from the model. Another model was proposed by Laurikkala et al.(Laurikkala et al., 2000). This method uses informal box plots to pinpoint outliers in both univariate and multivariate dataset. After pinpointing the outliers it produces a graphical representation and then there is a visual pinpoint of the outlying points by a human statistician. Real-valued, ordinal and categorical attributes can be handled through this method.[3]

A statistically discordancy test examines two hypothesis: a working hypothesis and an alternative hypothesis. Working hypothesis denotes a statement that the entire data set on n objects comes from an initial distribution model, F.

H: oi €F, where i= 1,2,……n

The discordancy test verifies whether the object i.e. oi is significantly large or small in proportion to the distribution "F". Assuming that some statistics T has been chosen for discordancy testing, and the value of the statistic for object oi is vi,, then the distribution of T is constructed. Significance probability, SP(vi)= Prob(T> vi) is evaluated. If SP(vi) is sufficiently small, then oi is discordant and the working hypothesis is rejected. An alternative hypothesis, H, which states that oi comes from another distribution model G, is adopted. The result is dependent on which F is chosen because oi may be an outlier under one model and a normal data point in another model.

Another method was proposed by Grubbs which calculates a Z value as the difference between the mean value for the attribute and the query value divided by the standard deviation for the attribute, where the mean and the standard deviation are calculated from all attribute values including the query value. The Z value for the query is compared with a 1% or 5% significance level. This technique requires no pre-defined parameter since the parameters are derived from the dataset. But for this method to be successful, the higher the number of records, the more statistically representative the sample is likely to be. (Grubbs, 1969).

Distribution based methods have many advantages. They are mathematically justified and give fast evaluation once they are built. They are generally suited to quantitative real-valued data sets or quantitative ordinal data distributions, where the ordinal data can be transformed into suitable numerical values through statistical processing.

However, most distribution models typically apply directly to the future space and are univariate i.e. have very few degrees of freedom. Thus, they are unsuitable even for moderately high dimensional data sets. If there are data sets where they are no prior distribution of points, then expensive tests are required to determine which model best fits the data. This fact limits their applicability and increases the processing time if complex data transformations are necessary before processing

2.2 Distance-based method

The notion of distance-based (DB) outlier as defined by Knorr and Ng (1988): An object O in a dataset T is a DB (p,D)-outlier if at least fraction p of the objects in T lie greater than distance D from O. The concept of distance based outlier is well defined for any dimensional dataset. The parameter p is the minimum fraction of objects in a data space that must be outside an outlier D-neighbourhood (Li & Hiroyuki, 2007). For many discordancy test, it can be shown that if an object, O, is an outlier according to the given test, then O is also a DB(pct,dmin) outlier for some suitable defined "pct" and "dmin". This distance based method generalizes many concepts from distribution-based approach. It also works well in cases of computational complexity. [4]

Some of the popular algorithms for mining distance based outliers are as follows:

Index based algorithm: Given a data set, the index based algorithm used multidimensional indexing structure, such as R-trees or k-d trees to search for neighbourhood of each object O within radius dmin around that object.

Nested Loop algorithm: It divides the memory buffer space into two halves, and the data set into several logical blocks. By carefully choosing the order in which blocks are loaded into each half, I/O efficiency can be achieved.

Cell Based algorithm: To avoid O (n2) a cell based algorithm, was developed for memory resident data sets. Its complexity is O(ck+n) where c is constant depending on the number of cells and k is the dimensionality. The data is partitioned into cells with a side length equal to dmin/2 √k. Each cell has two layers surrounding it. The first layer is one cell thick, while the second is[2 √k-1]. The algorithms counts outliers on a cell by cell basis. For a given cell it accumulates three counts i.e. the number of objects in the cell, in the cell and the first layer together, and in the cell and both layers together. An object O is considered an outlier if cell+1 layer count is less than or equal to M. If cell + 2 layers count is less than or equal to M, then all of the objects may be outliers. To detect these outliers, object by object processing are used, where for each object O , in the cell, objects in the second layer of O are examined. Only those objects having no more than M points in their dmin-neighbourhood are outliers. The dmin-neighbourhood of an oj=bject consists of the object's cell, all of its first layer and some of its second layer.

This distance based method is further extended based on the distance of a point from its kth nearest neighbour (Ramasmamy et al., 2000). After ranking points by the distance to its k-th nearest neighbor, the top k points are identified as outliers. Alternatively, in the algorithm proposed by Angiulli and Pizzuti (Angiulli & Pizzuti,2000), the outlier factor of each data point is computed as the sum of distances from its k nearest neighbors.

Both the method proposed in (Matsumoto et al., 2007) and the Mahalanobis outlier analysis (MOA) (Marquez et al., 2002) are distance-based approaches which exploit Mahalanobis distance as outlying degree of each data point. In 1936 P.C. Mahalanobis introduced a distance measure (Mahalanobis, 1936) which is based on correlations between variables by which different patterns can be identified and analyzed and provide a useful way of determining similarity of an unknown sample set to a known one. It takes into account the correlations of the data set and is scale-invariant, i.e. it is not dependent on the scale of measurements. Mahalanobis distance is computed on the basis of the variance of data points. It describes the distance between each data point and the center of mass relating to that particular data set. When one data point is on the center of mass, its Mahalanobis distance is zero, and when one data point is distant from the center of mass, its Mahalanobis distance is more than zero. Therefore, datapoints that are located far away from the center of mass are considered outliers.[5]

2.3 Density-based method

In the density based method, each object is assigned a degree to be an outlier. This degree is called the local outlier factor (LOF) of an object.

The degree depends on how isolated the object is with respect to the surrounding neighbourhood. 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 neighbourhood for that particular data set. High LOF is an indication of low-density neighbourhood and hence high potential of being outlier. In a typical density-based clustering algorithm, there are two parameters that define the notion of density: a parameter MinPts specifying a minimum number of objects and a parameter specifying a volume. The density threshold is determined by these two parameters for the clustering algorithms to operate. Objects or regions are connected if their neighbourhood densities exceed the given density threshold. For the detection of density based outliers, it is necessary to compare the densities of different sets of objects. This effectively means that the density of sets of objects must be dynamically determined. An idea is to keep MinPts as the only parameter and use the values reach-dist MinPts(p,o), for o ЄNMinPts(p).[6]

2.4 Deviation based outlier techniques

This technique identifies outliers by examining the main characteristic of objects in a group. Those objects that deviate from this description are considered to be outliers. There are primarily two techniques for deviation based methods i.e. sequential exception technique and OLAP Data Cube technique.

Sequential exception technique

This stimulates the way in which humans can distinguish unusual objects from a series of objects. It uses the implicit redundancy of the data. Given a data set D of n objects, it builds a sequence of subsets i.e. {D1, D2, Dm} with 2<=m<=n such that Dj-1 C DJ , where Dj —C D. This technique works through selection of a sequence of subsets from the set for analysis. For every subset, it determines the dissimilarity difference of the subset with respect to the preceding subset in the sequence.

OLAP Data Cube technique

The OLAP approach uses data cubes to identify regions of anomalies or outliers in large multidimensional data. This process is also overlapped with cube computation. In this approach, precomputed measures indicating data exceptions are used to guide the user in data analysis. A cell value in the cube is considered to be an exception if it is significantly different for the expected value which is based on the statistical model. This method uses the background colour to reflect the degree of exception of each cell. The measure value of a cell can reflect exceptions occurring at more detailed or lower levels of the cube, where these exceptions are not visible from the current level.

“3” Other Methods to detect outliers

New forms of outlier detection methods have been developed in recent times. Some of the methods are given below:

3.1 Neural network method

In 1943, McCulloch and Pitts introduced the idea of an artificial neuron to process data. Further, this work was advanced by arranging neurons in layers. The Multi-Layer Perceptron (MLP) is based on the rules to cope with multiple layers of perceptrons5. The Radial Basis Function (RBF) exploits gaussian activation functions in the first layer. First the inputs and outputs are defined, and then the data points which violate these input patterns are found. An outlier is frequently one which is furthest from the median value. [7]

Some of the other neural network methods for the detection of outliers are Principle Component Analysis (PCA) and Partial Least Squares (PLS). Outliers can be found by investigating points at the edges of the previously created clusters. Liu and Gader indicated that including outlier samples in training data and using more hidden nodes than required for classification for MLP and BRF networks and proceeding an RBF with principal Component decomposition can achieve outlier rejection. They further analyzed that further addition of a regularization term to the PCA-RBF can achieve an outlier rejection performance which will be equivalent or better than that of other networks without training on outliers .

Replicator Neural Network (RNN) has also been used for outlier detection. RNN are multi-layer perceptron neural networks with three hidden layers and the same number of output neurons and input neurons to model the data. The input variables are also the output variables so that the RNN forms compressed model of data during training. A measure of outlyingness of individuals is developed as the reconstruction error of individual data points. But an disadvantage of this method is that RNN degrades with datasets containing radial outliers and also in general neural network methods do not work well with small databases. However RNN performs satisfactory for small and large datasets.

Minimum Message Length (MLL) is another method for the detection of outlier. MLL clustering works well for scattered outlier while The self-organizing map (SOM) is an artificial neural networks, which is trained by using unsupervised learning in order to produce a low dimensional representation of the training samples while preserving the topological properties of the input space. The SOM based outlier detection method is non-parametric and can be used to detect outliers from large multidimensional datasets. This method has the advantage that it does not require any a priori assumption on the variable & is easy to implement and does not have problems with dimensionality of data. [8]

3.2 Fuzzy logic

Fuzzy Logic (FL) is linked with the theory of fuzzy sets, a theory which relates to classes of objects with unsharp boundaries in which membership is a matter of degree. From the beginning fuzzy logic has been applied to many applications from consumer products to industrial systems and transportations. Fuzzy logic has also

been used in soft computing to combine genetic algorithms or neural networks theory. . Since fuzzy logic is built atop the structures of qualitative description used in everyday language, fuzzy logic is easy to use.

Fuzzy inference system (FIS) (Ross, 2004) is the process of formulating the mapping from a given input to an output using fuzzy logic. The mapping then provides a basis from which patterns can be discerned and subsequent decisions can be made. The process of fuzzy inference involves: membership functions (MF). In this membership function, a curve defines how each point in the input space is mapped to a membership value or degree of membership between 0 and 1; fuzzy logic operators (and, or, not); if then rules. Since decisions are based on the testing of all of the rules in an FIS, the rules must be combined in some manner in order to make decision. Aggregation is the process by which the fuzzy sets that represents the outputs of each rule are combined into a single fuzzy set. This aggregation only occurs once for each output variable, just prior to the final step, which is defuzzification. Given below is an example which combines several outliers detection strategies by jointly evaluating four features of a particular datum within a series of measurements.

For each pattern the following four features are extracted and fed as input of the FIS:

- Distance between each element and the centroid of the overall distribution normalized with respect to the average value. (dist)

- Fraction of the total number of elements that are near to the pattern itself. (n-points)

- Mean distance between the considered pattern and the remaining patterns normalized with respect to the maximum value. (memb-deg)

- Degree of membership of the patterns to the cluster to which it has been assigned by the preliminary fuzzy c-means clustering stage. (mean-dist)

In geometry the centroid of an object X in n-dimensional space is the intersection of all hyperplanes that divide X into two parts of equal moment about the hyperplane. Roughly speaking, the centroid is a sort of “average” of all points of X.

The FIS is of Mandami type (Mandami & Assilian, 1975) and the FIS output variable, named outlier-index (outindx), is defined in the range [0;1]. The output function provides an indication on the risk that the considered pattern is an outlier.The inference rules relating the output variable to the four inputs is formulated through a set of 6 fuzzy rules, that are listed below:

1. IF (dist is very high) AND (n-points is very small) AND (memb-deg is low) AND (mean-dist is big) THEN (outindx is very high).

2. IF (dist is medium) AND (n-points is small) AND (memb-deg is quite low) AND (mean-dist is small) THEN (outindx is quite high).

3. IF (dist is low) AND (n-points is medium) AND (memb-deg is quite low) AND (mean-dist is very small) THEN (outindx is low).

4. IF (dist is medium) AND (n-points is very small) AND (memb-deg is quite low) AND (mean-dist is small) THEN (outindx is quite high).

5. IF (dist is low) AND (n-points is small) AND (memb-deg is high) AND (mean-dist is quite big) THEN (outindx is low).

6. IF (dist is low) AND (n-points is medium) AND (memb-deg is high) AND (mean-dist is small) THEN (outindx is low).

3.3 Artificial intelligence techniques

When dealing with industrial automation, where data coming from the production field are collected with different and heterogeneous means, the occurrence of outliers is more the rule than the exception. Standard outlier detection methods fail to detect outliers in industrial data because of the high dimensionality of the data. In these cases, the use of artificial intelligence techniques has received increasing attention in the scientific and industrial community, as the application of these techniques shows the advantage of requiring poor or no a priori theoretical assumption on the considered data. Moreover their implementation is relatively simple and with no apparent limitation on the dimensionality of the data.[9]

"4" Evaluation of outlier detection techniques

Evaluation of an outlier detection technique is very important to establish its usefulness in detecting outliers in a given data set. However there are several techniques that requires a number of parameters that need to be determined empirically. An evaluation metric is required in such cases to determine the best values for the involved parameters. Outliers can be detected in a given dataset as well as in an application domain.

4.1 Detecting outliers in a given data set

The objective of any outlier detection technique is to detect outliers in a data set. For this purpose, involves running the procedure on a validation set and seeing how well the technique detects the outliers. A labelling type of outlier detection technique is typically evaluated using any of the evaluation techniques from 2-class classification literature.[10] First of all a benchmark data set is chosen. One of the primary requirements is that the outliers should be labelled in the data set. The evaluation data is then split into training and testing data sets by using techniques such as hold-out, cross-validation, jack-knife estimation. The outlier detection technique is then applied to the test part of the validation data and the instances are labelled as outlier data points or normal data points. Then the predicted labels are compared with the actual labels to construct a confusion matrix as given below:

		Actual	
Predicted		Outlier	Normal instances
	Outlier	Ot	Of
	Normal instances	Nf	Nt

A confusion matrix generated after running an outlier detection technique on validation data is constructed using the above quantities, represented by f (Ot , Of , Nf , Nt , Θ, C). Θ represents the parameters associated with the outlier detection technique and C is a cost matrix that assigns weights of each of the four quantities.

For scoring type of outlier detection techniques, there is typically a threshold parameter to determine the cut-off above which the instances are treated as outliers. The threshold parameter is either determined using the outlier scores or left for the users to choose. After applying this cut-off, the confusion matrix is constructed and the evaluation metric is computed for the given value of threshold (incorporated in Θ).

Parametric statistical techniques are typically evaluated on artificially generated data sets from known distributions. Outliers are artificially injected in the data ensuring that they do not belong to the distribution from which rest of the samples are generated. The data sets containing rare class are chosen for this purpose and the instances belonging to the rare class are treated as outliers in the data. Another technique is to take a labelled data set and remove instances of any one class. This reduced subset then forms the normal data. All or few instances from the removed class are injected in the normal data as outliers.

But, most of the available benchmark data sets are adapted to be used for evaluating Type I outlier detection techniques. No data sets are available to evaluate Type II outlier detection techniques. For Type III outlier detection techniques there are no publicly available benchmark data sets, that can be used to evaluate any such technique. Several outlier detection techniques are evaluated for other objective functions such as scalability, ability to work with higher dimensional data sets, ability to handle noise. Those metrics which are described above are used for such evaluation. The only difference is the choice of data sets that can capture the complexity being evaluated. [11]

4.2 Evaluation in application domain

Such evaluation is done by choosing a validation data set that represents a sample belonging to the target application domain. The validation data should have labelled entities that are considered to be outliers in the domain. A key observation here is that such evaluation measures the performance of the entire outlier detection setting which also includes the technique, the features chosen and other related parameters /assumptions. Some of the popular application domains have benchmark data sets. Such benchmark data sets allow a standardized comparative evaluation of outlier detection techniques and hence are very useful. But often times the lack of such benchmark data sets have forced researchers to evaluate their techniques on proprietary or confidential data sets. Such data sets are not available publicly. Sometimes labelled validation data is often not available at all. In such cases a qualitative analysis is performed, which typically involves a domain expert.[12]

"5" Conclusion

Outlier detection methods can be divided between univariate methods, which have been proposed in the earlier works in this field, and multivariate

methods that usually form most of the current body of research. Another fundamental taxonomy of outlier detection methods is between parametric (statistical) methods and nonparametric methods that are model-free. Statistical parametric methods either assume a known underlying distribution of the observations or, at least, they are based on statistical estimates of unknown distribution parameters. These methods flag as outliers those observations that deviate from the model assumptions. But they are not effective in the case of high dimensional data sets or for arbitrary data sets where there is no prior intimation of the data distribution. Within the class of non-parametric outlier detection methods one can set apart the data-mining methods, also called distance-based methods. These methods are usually based on local distance measures and are capable of handling large databases. Another class of outlier detection methods is founded on clustering techniques, where a cluster of small sizes can be considered as clustered outliers which identifies both high and low density pattern clustering, further partition this class to hard classifiers and soft classifiers.

But the recent techniques like the artificial intelligence techniques present the advantage of requiring no a priori assumption on the considered data. Newer methods like Fuzzy Logic-based method outperforms the most widely adopted the traditional methods. Future work on the FIS-based outliers detection strategy will concern the algorithm optimization in order to improve its efficiency and its on-line implementation. Moreover further tests have to be performed on different applications. This will improve the detection capability of outliers.

References

1. Outlier Detection: A survey by Varun Chandola, Arindam Banerjee, and Vipin Kumar of the University of Minnesota.
2. Outlier analysis by Charu C. Aggarwal, IBM T. J. Watson Research Center, Yorktown Heights, NY, USA
3. A Meta analysis study of outlier detection methods in classification by Edgar Acuna and Caroline Rodriguez Department of Mathematics, University of Puerto Rico at Mayaguez Mayaguez, Puerto Rico 00680
4. Outlier Detection by Irad Ben-Gal, Department of Industrial Engineering, Tel Aviv University, Ramat- Aviv, Tel Aviv, Israel, 69978.
5. Outlier Detection Methods for Industrial Applications by Silvia Cateni, Valentina Colla and Marco Vannucci Scuola Superiore Sant Anna, Pisa, Italy.
6. Balance Sheet Outlier Detection Using a Graph Similarity Algorithm, University of Virginia; Stevens Institute of Technology, University of Virginia - Systems Engineering (2013)
7. Hartwig, J. and J.-E. Sturm (2012): An outlier-robust extreme bounds analysis of the determinants of health-care expenditure growth, KOF Working Papers No. 307, June, Zurich.
8. T. de Vries, S. Chawla, M. E. Houle Density-preserving projections for large-scale local anomaly detection Knowledge and Information Systems (KAIS), 32(1): 25–52, 2012.
9. H.-P. Kriegel, P. Kröger, E. Schubert, A. Zimek Outlier Detection in Arbitrarily Oriented Subspaces In Proceedings of the 12th IEEE International Conference on Data Mining (ICDM), Brussels, Belgium: 379–388, 2012.
10. T. de Vries, S. Chawla, M. E. Houle Finding Local Anomalies in Very High Dimensional Space In Proceedings of the 10th IEEE International Conference on Data Mining (ICDM), Sydney, Australia: 128–137, 2010.
11. H.-P. Kriegel, P. Kröger, E. Schubert, A. Zimek Interpreting and Unifying Outlier Scores In Proceedings of the 11th SIAM International Conference on Data Mining (SDM), Mesa, AZ: 13–24, 2011.
12. H. V. Nguyen, V. Gopalkrishnan, I. Assent An Unbiased Distance-based Outlier Detection Approach for High-dimensional Data In Proceedings of the 16th International Conference on Database Systems for Advanced Applications (DASFAA), Hong Kong, China: 138–152, 2011.