Expert Systems with Applications 36 (2009) 7104–7113

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Projected outlier detection in high-dimensional mixed-attributes data set

Mao Ye [a,*], Xue Li [b], Maria E. Orlowska [c]

[a] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China
[b] School of Information Technology and Electronic Engineering, The University of Queensland, Brisbane, Qld 4072, Australia
[c] Polish-Japanese Institute of Information Technology, Faculty of Information Technology, Ul. Koszykowa 86, 02-008 Warsaw, Poland

ARTICLE INFO

Keywords:
Outlier detection
Data mining
High-dimensional spaces
Mixed-attribute data sets

ABSTRACT

Detecting outlier efficiently is an active research issue in data mining, which has important applications in the field of fraud detection, network intrusion detection, monitoring criminal activities in electronic commerce, etc. Because of the sparsity of high dimensional data, it is reasonable and meaningful to detect the outliers in suitable projected subspaces. We call such subspace and outliers in the subspace as anomaly subspace and projected outlier respectively. Many efficient algorithms have already been proposed for outlier detection based on different approaches, but there are few literatures on projected outlier detection for high dimensional data sets with mixed continuous and categorical attributes. In this paper, a novel projected outlier detection algorithm is proposed to detect projected outliers in high-dimensional mixed attribute data set. Our main contributions are: (1) combined with information entropy, a novel measure of anomaly subspace is proposed. In this anomaly subspace, meaningful outliers could be detected and explained. Unlike the previous projected outlier detection methods, the dimension of anomaly subspace is not decided beforehand; (2) theoretical analysis about this measure is presented; (3) bottom-up method is proposed to find the interesting anomaly subspaces; (4) the outlying degree of projected outlier is defined, which has good explanations; (5) the data set with mixed data type is handled; (6) experiments on synthetic and real data sets to evaluate the effectiveness of our approach are performed.

1. Introduction

Outlier detection in large data sets is the one of major research fields in data mining. This technology is to find a small group of data points automatically which are different from the rest of the large amount of data based on some measures. Outlier detection has many applications in fraud detection (Fawcett & Provost, 1997), network intrusion detection (Eskin, Arnold, Prerau, Portoy, & Stolfo, 2002; Knorr, Ng, & Tucakov, 2000), insurance fraud (Bolton & Hand, 2002), medical diagnosis (Penny & Jolliffe, 2001) and marketing. Efficient detection of such outliers can help us make good decisions on erroneous data or prevent the negative effect of malicious and faulty behavior. Outlier detection can also be used in data cleaning (Otey, Ghoting, & Srinivasan, 2006). Since the anomaly data may distort the distribution of real data badly, accurate and efficient removal of outliers in data sets may greatly enhance the performance of data mining algorithms and techniques (Gamberger, Lavac, & Groselj, 1999).

Many outlier detection algorithms have been proposed in recent years. They are based on supervised approach or unsupervised approach. Supervised approach needs to learn a classification model over a set of sample data already labeled as outlier or not, then a new sample is classified as outlier or normal by using this model (Fawcett & Provost, 1997; Lee, Stolfo, & Mok, 1998). Unsupervised method detect the outliers in the set of data without training samples. Among the unsupervised approaches, most of algorithms are either distance based or density based (Otey et al., 2006; Knorr & Ng, 1998; Knorr & Ng, 1999; Breuning, Kriegel, Ng, & Sander, 2000; Ramaswamy, Rastogi, & Shim, 2000). However, real-life data sets limit the effectiveness of these algorithms because a few real-life data sets are high-dimensional and may contain a mixture of attribute types (i.e. continuous and categorical attributes) (Otey et al., 2006). Here, high-dimension of data set means that there are at least ten attributes.

The first problem is that some of real-life data are always in high-dimensional. Most of previous outlier detection algorithms are originally aimed for detecting outliers in low-dimensional data. With the more and more high-dimensional data are found nowadays, we need specially design algorithm to deal with the "dimensionality curse" problem. As reported in Beyer, Goldstein, and Ramakrishnan (1999) recently, in high dimensional space, distances between every pair of data objects are almost the same for a wide variety of data distributions and distance functions. Thus we can hardly find any outlier pattern in high dimensional data

* Corresponding author.
E-mail addresses: yem_mei29@hotmail.com (M. Ye), xueli@itee.uq.edu.au (X. Li), maria@itee.uq.edu.au (M.E. Orlowska).

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.08.030

sets by using the concept of proximity and neighborhood (Ng, Fu, & Wong, 2005). For example, the distance based outlier detection algorithms proposed in Knorr and Ng (1998), Knorr and Ng (1999). The outlier is defined in Knorr and Ng (1998) as the following: A point p in a data set is an outlier with respect to the parameters k and λ, if no more than k points in the data set are at a distance λ or less from p. As denoted in (Aggarwal & Philip, 2005), in high dimensional data sets, since the equal distance of almost all pairs of data, most of the points are likely to lie in a thin shell (Beyer et al., 1999), it will be very difficult to choose the parameter λ. For the density based outlier detection algorithm (Breuning et al., 2000), it computes the local density of a point o by using the average smoothed distances to a certain number of points in the locality of o. In this case, the local locality is difficult to be defined, since the meaningful concept of proximity and neighborhood does not exist in sparse high-dimensional data.

Traditional method to handle the high dimensional problem is using dimensionality reduction techniques, such as principal component analysis (PCA) which projects the whole data set onto a subspace while minimizing the information loss. The major disadvantage of using dimension reduction techniques is that they may lead to significant loss of information. And in the data clustering literatures, it has been found that more meaningful cluster can be found in the particular subspace (Aggarwal & Philip, 2005). Clusters are embedded in the subspaces of the high dimensional data. Different patterns can only be found in different subspaces. This is because different localities of the data are dense with respect to different subsets of attributes (Aggarwal & Philip, 2005). These cluster are referred to as projected cluster or subspace cluster. For a review on subspace clustering, please refer to (Patrikainen & Meilla, 2006; Parsons, Haque, & Liu, 2004; Parsons, Haque, & Liu, 2004). Actually in Ramaswamy et al. (2000), it has also been found that more interesting outliers can be found on the NBA98 basket ball database by using fewer features. So it is reasonable to detect more meaningful outliers in the projected subspaces.

Consider Fig. 1, there are only thee points which are equidistance for every pair of data. traditional algorithms would fail to discover any outliers in the 2-dimensional space XY. However, if we project the whole data to subspace X and subspace Y, it is easy to see that an outlier can be detected in space Y by using density based algorithm, while this point is normal in another subspace. As pointed out in Procopiuc, Jones, Agarwal, and Murali (2002), Ng et al. (2005), for real life data, often only the original attributes are meaningful and interpretable. For example, suppose a employee database contains three attributes: age, income and the number of family member. We may uncover outliers in the subspace [age, income] and cannot find any anomaly by using all attributes. The outliers uncovered in the subspace [age, income] can be easily interpreted, i.e. who has a high salary with a more younger age. This means that the outliers should be detected in the subspace with the original attributes as the axes.

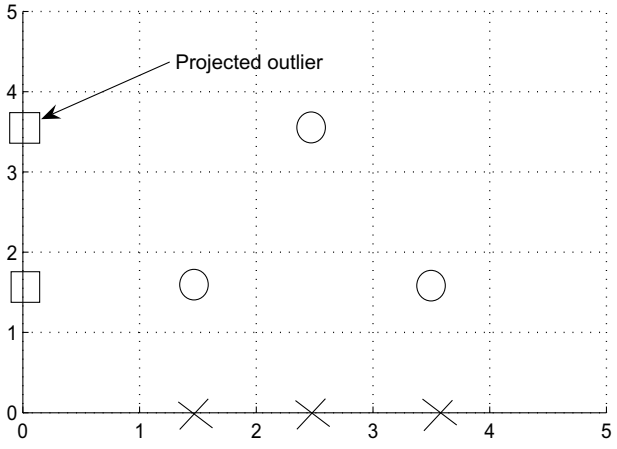

Fig. 1. Illustration of projected outlier.

The second problem is that the attributes in a data set may be a mixture of categorical and continuous types (Otey et al., 2006). In real applications, data are usually mix-typed (Otey et al., 2006). Since most current definition on outliers are based on distance (Knorr & Ng, 1998; Knorr & Ng, 1999), categorical data cannot be handled efficiently (Li, Qian, Zhou, Jin, & Yu, 2003). To detect outliers in categorical data, the "curse of dimensionality" problem will also be faced, which will bring similar difficulties as the continuous type data. Consider a modified example from (Li et al., 2003), the custom database has three categorical attributes: Age-range, car-type and salary-level. In the database, there are two instances of ('Middle','Sedan','High'), two of ('Young','Sports','High'), and three of ('Middle','Sports','High'). We can see that the occurrences are very close to each other. The outliers cannot be detected based occurrence. However, if this data set is projected into the subspace [Age-range, Salary-Level], the two instances ('Young','Sports', 'High') can be considered as outliers in the subspace [Age-range, Salary-Level].

1.1. Our contributions

In this paper, we will address the above problems. Our method, which is referred to as Projected Outlier Detection in High Dimensional Mixed-Attributes Data Sets (PODM), is focus on uncovering projected outliers in varying dimensionality mixed type attributes subspace without the users to input the dimensionality of anomaly space. We adapt the projected outlier definition in Aggarwal and Philip (2005). This definition considers a point to be an outlier if in some lower-dimensional projection it is present in a local region of abnormally low density. And we call such point as projected outlier and the lower-dimensional subspace as anomaly subspace. Specifically, the contributions can be summarized as follows:

- Combined with information entropy, a novel measure of anomaly subspace is proposed. In this anomaly subspace, meaningful outliers could be detected and explained. Unlike the previous projected outlier detection method, the dimensionality of projected outlier subspace is not decided beforehand.
- Theoretical analysis about this measure is presented. We prove a few properties of the measure of anomaly space, which is very useful for practical computations.
- Bottom-up method is proposed to find interesting anomaly subspace.
- The outlying degree of projected outliers is defined. And we can report the regions or anomaly subspace where the projected outliers are located, which is interpretable and very useful to users.
- The anomaly subspace of data sets with mixed data type is handled. The idea of frequent items is used.
- Experiments on several real and synthetic data sets are performed, which confirm the utility of the projected outliers uncovered.

The remainder of this paper is organized as follows. In Section 2, previous outlier detection algorithms have been discussed. In Section 3, the measure of goodness of anomaly was defined and some properties have been proved. The outlying degree of projected outliers is also defined. The algorithms to find the anomaly subspace and projected outliers are presented in Section 4. We present experiment results in Section 5. Finally, our conclusions are given in Section 6 and possible future works are also addressed.

2. Related works

Outlier detection is firstly studied in the field of statistics, where the data is assumed to follow a parametric distribution. Recent results show that these methods are not suitable for data mining applications in even moderate high-dimensional spaces. In this section, an overview of unsupervised outlier detection methods on low dimensional data set is first given, which include distance-based methods (Knorr & Ng, 1998; Knorr & Ng, 1998), clustering-based approaches (Guha, Rastogi, & Shim, 2000; Sequeira & Zaki, 2002 and density-based algorithms (Breuning et al., 2000; Papadimitriou, Kitawaga, Gibbons, & Faloutsos, 2003). Then, we will give more descriptions about two algorithms in Otey et al. (2006), Aggarwal & Philip (2005), which have close relationship with our work.

Distance-based outlier detection approach was first presented in Knorr & Ng (1998), Knorr & Ng (1999). A point is an outlier if the data are sparse within the hyper-sphere for given radius. Since they define outlier by using the distribution of distance of the other points to a given point. They may fail when processing categorical data and be susceptible to the dimensionality curse. Clustering-based algorithms define outliers as points that do not lie in cluster (Guha et al., 2000; Sequeira & Zaki, 2002). They implicitly define the outliers as the background noise. However, recent studies show that more useful outliers are neither a part of cluster nor a part of the background noise; rather they are specifically points that behave very differently from the norm (Aggarwal & Philip, 2005). Density-based methods, for example (LOF method) (Breuning et al., 2000), study the outlying characteristics in local area, in which data points usually share similar density distribution property. However, this method also face the "curse of dimensionality" problem since the data in high dimensional space are sparse. It is hard to find a suitable density measure.

Following that, to overcome the high-dimensional difficulties, an efficient and effective algorithm has been proposed in Aggarwal & Philip (2005). For similicity, we refer it to as Projected *O*utlier *D*ectection in High Dimensional Data Set by Using *E*volutionary Computation (PODE).

The method PODE defines outliers as such patterns in those abnormal projections of the data which have abnormally low presence that cannot be justified by randomness. An abnormal lower-dimensional projection is one in which the density of the data is exceptionally lower than average. In order to find such projections, grid discretization of data is performed. Each attributed of the data is divided into ϕ ranges based on equidepth basis. Each range contains $f = 1/\phi$ of the records. Consider a k-dimensional cube that is created by picking grid ranges from k different dimensions. Assume that there are a total of N points, and the dimensionality is d. If the data were uniformly distributed, then the presence of absence of any point in a k-dimensional cube will be probability f^k. Thus the expected fraction and standard deviation of the points in a k-dimensional cube are $N \cdot f^k$ and $\sqrt{N \cdot f^k \cdot (1-f)^k}$. Let $n(D)$ be the real number of points in a k-dimensional cube D. The sparsity coefficient $S(D)$ of the cube D is defined as $S(D) = (n(D) - N \cdot f^k)/\sqrt{N \cdot f^k \cdot (1-f)^k}$. Then, the outlier detection problem is changed to find the most sparse k-dimensional cubes in the data.

An efficient and effective algorithm base on evolutionary computation has been proposed. Assume the grid range for the ith dimension is denoted by m_i, the value of m_i can take on any of vales 1 through ϕ, or it can take on the value $*$, which denotes a 'don't care'. Consider a four dimensional example with $\phi = 10$ to use evolutionary computation techniques, then the gene expression of a possible solution can be $*3*9$. By using such expression of chromosome, evolutionary computation techniques are used. The algorithm PODE has two parameters k and ϕ, which can only find k-dimensional abnormally projections. However, more interesting and useful outlier patterns may exist in different dimensional subspace. And PODE cannot handle data set with categorical attributes.

3. Problem statements

Given an un-normalized data set D with N data objects, let Q be the set of dimensionality of D, i.e. $Q = \{1,2,\ldots,d\}$, where d is the dimensionality of this set. The dimensions measure different things. Without loss of generality, assume the first p dimension (or attributes) have continuous values and the other attributes have categorical values, let the index sets $N = \{1,\ldots,p\}$ and $C = \{p+1,\ldots,d\}$. For the first p dimensions, each dimension is discretized by its corresponding equi-width parameter δ. How to choose this parameter will be clear in the next section. For the categorical data dimension, the data block corresponds to a value of the corresponding attribute. Thus the high-dimensional data space is partitioned to a few data cells.

Here the equi-width method is used to discretize the continuous attribute instead of equi-depth method in Aggarwal & Philip (2005). Although the equi-depth method can avoid the existence of empty (or sparse) data cells, the outliers may be descritized into a normal cell. For example, in one dimension space, there are two clusters asides and an outlier in the center. After the equi-depth method is applied, three cells with same number of points are created. The outliers cannot be inspected. Moreover, by using equi-width method, the exact regions of outliers can be localized.

Now, let us first give some definitions.

Definition 1. A data object x is a single data item. It is represented by a vector of d attribute values in d-dimensional space: $x = (x_1,\ldots,x_d)$, where x_j is the attribute value.

Definition 2. A subspace S with dimensionality d_i is defined by a set of d_i attributes, where $d_i < d$.

Definition 3. A subspace X is called anomaly subspace, in which the interesting projected outliers can be found.

Definition 4. The data cell c is referred to as anomaly cell in the anomaly subspace X if the density of cell c is lower than a user defined threshold value.

Definition 5. A data object x is a projected outlier if it lies in an anomaly cell. Based on the above definitions, our projected outlier detection problem can be stated as the following: Given a data set D with N data objects, the equi-width parameter δ, and integer number k, find the anomaly subspaces, and identify the anomaly cells of each anomaly subspace, and sort all of these anomaly cells by scoring their outlying degrees, finally report the cells which have the first k outlying degrees, i.e. the set $C = \{c_1,\ldots,c_k\}$. The data objects in the cells of set C will be the projected outliers uncovered.

This problem consists of several small problems. The first problem is to define a suitable measure to decide which subspace is the anomaly subspace. The second problem is how to define the outlying degree of anomaly cells (projected outliers) in different anomaly subspace. The third problem is to derive an algorithm which can find the top-k anomaly cells efficiently.

4. Anomaly subspace and outlying degree

In this section, we give some notations about the anomaly subspace and the outliers in such subspaces. There are two questions

to be answered. The first question is what subspace is an anomaly subspace. The second question is how to compare the outlying degrees of anomaly cells (projected outliers) in different anomaly subspace. In our algorithm, the outlying degree of an anomaly cell is the same as that of projected outliers in it.

4.1. Anomaly subspace

In a mixed data type subspace X, we treat the continuous attributes as the categorical attributes by discretizing them using equi-width method, the data space was partitioned into a set of cells. Let π be the set of all cells which contain data, and $p(c)$ be the density of a cell c in terms of the percentage of data contained in the cell c. Entropy was first used in Cheng et al. (1999) as a measure to find the suitable subspace to cluster numerical data. It was extended to categorical data case in Simovici, Cristofor, & Critofor (2005). And more general entropy functions are formed. Here, we adopt the Gini-entropy definition. As illustrated in Simovici et al. (2005), this entropy measures the average distance between any two data objects of a data set.

Definition 6. For the data set D in subspace X, the Gini-entropy is defined as follows:

$$H(X) = 1 - \sum_{c\in\pi} |p(c)|^2. \tag{1}$$

If the distance $d(x,x')$ of two data objects x and x' in subspace X is defined as

$$d(x,x') = \begin{cases} 0 & \text{if } x = x'; \\ 1 & \text{otherwise}, \end{cases} \tag{2}$$

the average distance of the data objects in the subspace X will equal to $H(X)$ (Simovici et al., 2005). It means the smaller $H(X)$ is, the more compactness of the projected data in subspace X is. By using similar method in Cheng et al. (1999), we can show the relationships between Gini-entropy and the density of cells. For simplicity, assume there are k dense cells with the density equals to α and $n-k$ non-dense cells with density β. For more complex case, theoretical analysis cannot be performed. However, theoretical analysis of simple case at least can illustrate some phenomena partially. Then we have

$$\begin{aligned} H(X) =& 1 - \sum_{c\in\pi} |p(c)|^2 \\ =& 1 - k|\alpha|^2 - (n-k)|\beta|^2. \end{aligned}$$

Theorem 1. $\frac{dH(x)}{d\alpha} \leqslant 0$ *if and only if* $\alpha \geqslant \beta$.

Proof. Since

$$k\alpha + (n-k)\beta = 1,$$

it means that

$$\frac{d\beta}{d\alpha} = \frac{k}{k-n}.$$

Directly,

$$\begin{aligned} \frac{dH(x)}{d\alpha} =& -2k\alpha - 2(n-k)\frac{d\beta}{d\alpha} \\ =& -2k\alpha + 2k\beta \\ =& -2k(\alpha-\beta). \end{aligned}$$

It means that $\frac{dH(x)}{d\alpha} \leqslant 0$ if and only if $\alpha \geqslant \beta$. □

From Theorem 1, Gini-entropy will decrease if the density of the density cells is increase. Thus, as the tradition defined entropy used in Cheng et al. (1999), Gini-entropy can also be used as a measure to find the good subspaces which have interesting clusters in mixed type data set. However, the entropy measure is not very suitable to find the anomaly subspace since it is not sensitive to the outliers.

Example 1. Suppose 6000 data objects are projected into two subspaces. A subspace has four data cells which have 1500 data objects respectively. Another subspace has five data cells which have three 1500, 1499 and 1 data objects respectively. The Gini-entropy of first subspace $H_1 = 1 - 4*(1500/6000)^2 = 0.75$. And the Gini-entropy of the second subspace $H_2 = 1 - 3*(1500/6000)^2 - (1499/6000)^2 - (1/6000)^2 = 0.75008$. The first subspace has good clusters, and the second subspace has one outlier. The Gini-entropy measures of both subspace are almost the same. The anomaly subspace cannot be found accurately. The other general entropy defined in Simovici et al. (2005) have the similar phenomena as that of Gini-entropy.

Moreover, we can give more theoretical analysis about this phenomena. For a data set D, assume there are two subspace, one subspace X has k dense cells with density α, another subspace Y has $k-1$ dense cells with density α, a dense cell with density $\bar{\alpha}$ and a non-dense cell with only one data object.

Theorem 2. $H(X) - H(Y) \sim O\left(\frac{1}{|D|}\right)$, *where* $|D|$ *is the cardinality of the data set D.*

Proof. By the definition of Geni-entropy,

$$H(X) = 1 - k|\alpha|^2,$$

$$H(Y) = 1 - (k-1)|\alpha|^2 - |\bar{\alpha}|^2 - \left|\frac{1}{|D|}\right|^2.$$

Since $(k-1)\alpha + \bar{\alpha} + 1/|D| = 1$ and $k\alpha = 1$, we have

$$\bar{\alpha} = \alpha - \frac{1}{|D|}.$$

It follows that:

$$\begin{aligned} H(X) - H(Y) =& -|\alpha|^2 + |\bar{\alpha}|^2 + \left|\frac{1}{|D|}\right|^2 \\ =& -|\alpha|^2 + \left|\alpha - \frac{1}{|D|}\right|^2 + \left|\frac{1}{|D|}\right|^2 \\ =& \frac{1}{|D|^2} - \frac{2\alpha}{|D|}. \end{aligned}$$

Since $\alpha < 1$, the theorem follows. □

Assume X is the anomaly subspace of data set D, and the cell set is π. $p(c)$ is the density of a cell $c \in \pi$. To characterize the outlying degree of the anomaly subspace, we define a new measure as follows:

$$O(X) = \frac{1}{|D|}\sum_{c\in\pi}\frac{1}{p(c)}. \tag{3}$$

Suppose that there are k dense cells with the density equals to α and $n-k$ non-dense cells with density β. We have the following theorem.

Theorem 3. $\frac{dO(X)}{d\alpha} > 0$ *if* $\alpha > \beta$.

Proof. Differential the outlying function of anomaly subspace $O(X)$ with respect to the density α, we have

$$\frac{dO(X)}{d\alpha} = \frac{1}{|D|}\left(-\frac{k}{\alpha^2} - \frac{n-k}{\beta^2}\frac{d\beta}{d\alpha}\right).$$

Since $\frac{d\beta}{d\alpha} = \frac{k}{k-n}$, it follows that:

$$\frac{dO(X)}{d\alpha} = \frac{1}{|D|}\left(\frac{k}{\beta^2} - \frac{k}{\alpha^2}\right).$$

If $\alpha > \beta$, the theorem follows. □

This theorem tells us that the density of the dense cells increases, the outlying degree of anomaly space is increase. Because the density of the $n - k$ non-density cells are decrease, the data objects in these cells are more outlying, which means that the outlying degree of anomaly space is increased. From Theorem 3, we have the following interesting proposition.

Proposition 1. $\frac{dO(X)}{d\alpha} < 0$ *if* $\alpha < \beta$.

This says that the density of the dense cells decrease, the outlying degree of anomaly subspace is decreased until the density of the dense cells is less than that of the non-dense cells. Then, the outlying degree will increase when the density of the dense cells are decreased. This is because the non-dense cells are changed to dense cells when the density of dense cells are decreased. And when the density of dense cells are small enough, the dense cells will be changed to non-dense cells. Thus, the outlying degree of anomaly subspace is increased again.

For the data set D, assume there are two subspace, a subspace X has k dense cells with density α, another subspace Y has $k-1$ dense cells with density α, a dense cell with density $\bar{\alpha}$ and a non-density cell with only one data object. We can show that this outlying degree function is sensitive to the outlier.

Theorem 4. $O(\overline{X}) - O(X) \sim O(1)$.

Proof. By the definition of outlying degree function in (3), we have that

$$O(X) = \frac{1}{|D|}\frac{k}{\alpha},$$

$$O(\overline{X}) = \frac{1}{|D|}\left(\frac{k-1}{\alpha} + \frac{1}{\bar{\alpha}} + |D|\right).$$

It follows that:

$$O(\overline{X}) - O(X) = \frac{1}{|D|}\left(\frac{1}{\bar{\alpha}} - \frac{1}{\alpha}\right) + 1.$$

Since $\alpha > \bar{\alpha}$, the theorem follows. □

By Theorems 3, 4 and Proposition 1, it looks that the outlying degree function can be used to find the anomaly subspace. However, this measure has its default, i.e. it is not very sensitive to the compactness of the projected data set in subspace.

Example 2. Suppose 10,000 data objects are projected into two subspaces. One subspace has two data cells which have 9999 data objects and one data object respectively. Another subspace has three data cells which have 5000, 4999 and 1 data objects respectively. The outlying degree of the first subspace O_1 = 1/9999 + 1 = 1.001. And the outlying degree of the second subspace O_2 = 1/5000 + 1/4999 + 1 = 1.004. The outlier in the first subspace is considered to be more interesting than that in the second subspace. The outlying measures of both subspace are almost the same. On the other hand, the Gini-entropy of the above subspace are 0.0002 and 0.5001 respectively. If we have conditions that the Gini-entropy is small and choose only one outlier, the first subspace will be considered as an anomaly subspace. However, by using the outlying degree measure of subspace, the second subspace will be considered as an anomaly subspace.

Thus, by combining the outlying degree measure with the Gini-entropy, the anomaly subspace can be found. The formal definition of anomaly subspace can stated as follows.

Definition 7. Give two threshold values ϵ and ξ, the subspace X is refereed to as an anomaly subspace, if $H(X) < \epsilon$ and $O(X) > \xi$.

4.2. The outlying degree of anomaly cell

As stated in Definition 4, the projected outlier is the data object in the low density cell of anomaly subspace. And our problem is to find the k most outlying anomaly cells in these anomaly subspace. To characterize the outlying degree of anomaly cells(projected outliers) in different subspace, we fit the definition of anomaly scores in Otey et al. (2006) to our cases. Before more details of our outlying degree definition, similarly as the definition in Ng et al. (2005), we first give a signature to each anomaly cell.

Definition 8. A signature s of a cell c in subspace X is an order list of d entries. If we index the categorical attribute values by number, specifically, $s = [s_1, s_2, \ldots, s_d]$, where

The projected outliers in data cell c have the same signatures as that of the cell.

For example, if the signature of cell c is $s = [*, *, 2, 3, *]$, it means that the subspace which this anomaly cell lies in consists of the third and fourth attributes. And the values of the third and fourth attributes of this cell equal to the second and the third categorical (or numerical interval) values respectively. A signature of data cell can specify a unique data cell. Thus the signature can be used to represent the cell. We say the signature $\theta \subseteq s$ means that the signature θ copies s except that some values of components are changed to $*$. For the previous example, $\theta = [*, *, *, 3, *] \subseteq s$.

By using the idea of link analysis in Ghoting, Otey, & Partheasarathy (2004), Otey et al. (2006), the outlying degree of the anomaly cell c (or projected outliers in this cell) is defined as follows.

Definition 9. Assume a data cell c is the anomaly cell with signature s and $p(\theta)$ is the density of cell α with signature θ, and the entropy of subspace which contains the cell θ is less than the threshold value ϵ, the outlying degree of the data cell c is defined as:

$$\mathrm{PO}(s) = \sum_{\theta \subseteq s}\left(\frac{|\theta|}{Np(\theta)}\middle| p(\theta) \leqslant \zeta\right), \tag{4}$$

where $Np(\theta)$ actually is the number of data objects in the cell θ, ζ is the user defined threshold value. $|\theta|$ is the number of components of signature θ which do not equal to $*$. It actually means the dimensionality of subspace the cell θ belongs to.

Remark 1. The outlying degree of an projected outlier is consider to be more high, if the dimensionality of associated anomaly subspace is high and more sub-signatures are infrequent. Based on these principles, the outlying function is designed.

Remark 2. In the definition, corresponding to the cell sub-signatures, only the frequent number of those signatures, which belong to a subspace whose entropy is less than a threshold value ϵ, is considered. If the entropy of subspace is big, almost all cells are sparse. The frequent number of such signature has not contribution to characterize the outlying degree.

Remark 3. The anomaly score function of data object P for categorical data was defined in Ghoting et al. (2004) as follows:

$$\mathrm{Score}(P) = \sum_{\theta \subseteq S}\left(\frac{1}{|\theta|}\middle| p(\theta) \leqslant \zeta\right),$$

where S is the signature of data object p in whole space. Since we need compare the outlying degree of projected outliers in different subspace, this definition is not very suitable to our problem. For example, consider two anomaly cells with signatures $s_1 = [*, *, 2, 3, 3]$ and $s_2 = [*, *, 1, 3, *]$, assume all sub-signatures are frequent and

both cells have only one data object respectively, then Score(s_1) = 1/3 and Score(s_2) = 1/2. However, in our problem, the outlier with high dimensionality is considered to be more outlying. Such definition will give us opposite answer. On the other hand, by using our definition, PO(s_1) = 3 and PO(s_2) = 2. The cell s_1 considered to be more anomaly.

5. Algorithms

In this section, algorithm PODM will be presented in details. Similarly as the method in Cheng et al. (1999), an apriori-like algorithm is proposed to find the top-k outlying degree anomaly cells.

5.1. The PODM algorithm

The PODM algorithm can be summarized as the following three steps:

1. Find out the anomaly subspaces by using apriori method based on our two measures of anomaly subspace.
2. Compute the outlying degree of anomaly cells of each anomaly subspace recursively.
3. Output anomaly cells with the k highest outlying degree and their associated signatures, i.e. the corresponding anomaly subspace and attribute ranges.

In step 1, bottom up method is used to find the anomaly subspace by using apriori method, which is similar to the method in Cheng et al. (1999). The procedure of apriori method is as follows. The algorithm first finds one-dimensional subspaces whose Geni-entropies H are less than the threshold value ϵ. Then, based on these subspaces, we construct the candidate two-dimensional subspaces and check them whether the Geni-entropy is less than the threshold value ϵ. Keeping these subspaces, and repeat the process again until the more subspace can be found whose Geni-entropy is less than ϵ. We refer to all of these subspaces as candidate anomaly subspaces. To use the notation of Geni-entropy, in the beginning, these subspaces are divided to cells as stated in Section 4. Downward property of Geni-entropy is required to use apriori method. It is known that traditional defined entropy satisfy the downward property (Cheng et al., 1999). Actually, we can observe that the Geni-entropy also satisfies this downward property.

Theorem 5. *Suppose K and L are the two subspaces of data set D such that $K \subset L$, then $H(K) \leqslant H(L)$.*

As pointed in Section 4, these candidate anomaly subspaces may not be satisfied, since some of them are good for clustering but not for outlier detection. And the number of these subspaces may be large. So we use the outlying degree function $O(X)$ for any candidate anomaly subspace X to decide whether it is an anomaly subspace. By our definition, if $O(X) < \xi$, the candidate anomaly subspace X will be the anomaly subspace we wanted. Next, we will show the outlying degree function $O(X)$ satisfies downward property, which can simplify computations to find all anomaly subspaces.

Theorem 6. *Suppose K and L are two subspaces of data set D such that $K \subset L$, then $O(K) < O(L)$.*

Proof. Without loss of generality, assume c_{ij} $(1 \leqslant i \leqslant H, 1 \leqslant j \leqslant N_i)$ are the cells in the subspace L with number equals to $\sum_{i=1}^{H} N_i$. And $\bar{c}_i, i \leqslant H$ are the cells in subspace K, where the data objects in $\bar{c}_i$ consists of the data projections of cells C_{ij} in subspace L for $1 \leqslant j \leqslant N_i$. By definition, we have

$$O(L) = \sum_{i=1}^{H} \sum_{j=1}^{N_i} \frac{1}{|D|p(c_{ij})},$$

and

$$O(K) = \sum_{i=1}^{H} \frac{1}{|D|p(\bar{c}_i)},$$

where $|D|p(\bar{c}_i) = |D|\sum_{j=1}^{N_i} p(c_{ij})$ for $1 \leqslant i \leqslant H$. It follows that:

$$\begin{aligned} O(K) &< \sum_{i=1}^{H} \frac{1}{|D|\sum_{j=1}^{N_i} p(c_{ij})} \\ &< \sum_{i=1}^{H} \sum_{j=1}^{N_i} \frac{1}{|D|p(c_{ij})} = O(L). \end{aligned}$$

For any two subspace K and L such that $K \subset L$, this property says that $O(L) > \xi$ if $O(K) > \xi$. If K is anomaly subspace, then all candidate anomaly subspace, which K belongs to, will be anomaly subspace. Thus, we need not compute the outlying degree of every candidate anomaly subspace. Then the anomaly subspaces are between two borders of subspaces, the upper border is formed by geni-entropy condition $H < \epsilon$ and the lower border is formed by outlying degree condition $O > \xi$. In our paper, we assume the outliers in higher dimensional subspace are more interesting. It means that the anomaly subspaces consist of the upper border are more interesting. High dimensionality means that the associated object will be an outlier in multiple independent subspaces and detection error rate will be smaller since more attributes are used. Then, it is reasonable to pay more attention to this object. It is worth for the extra computational expense to find this upper border. For example, in the network intrusion detection, each network connection has more than thirty attributes. If all attributes are used to detect abnormal detection, many normal network connection are detected as illegal intrusion. On the other hand, if a small set of attributes are used, many illegal connections will be reported as normal. So we find the outliers in the upper border subspaces with the entropy constrains. By using this idea, the error rate of detection will decrease. In this example, traditional outlier detection algorithm will find many outliers in high dimensional space such that these outliers cannot be used. Our algorithm only finds the outliers in the subspace which satisfies the entropy condition. In this way, there are not too much outliers, which can be analyzed more efficiently. Consider an example for a data set with four attributes $[A,B,C,D,E]$, these two borders are drawn in Fig. 2. It can be observed that subspaces AC, BC and ABD consists of upper border of all anomaly subspaces. In the first step of our algorithm, our goal is to find such upper border anomaly subspaces. We refer to these subspaces as interesting anomaly subspaces.

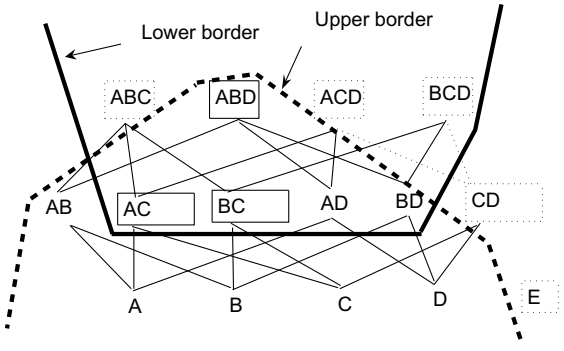

Fig. 2. Illustrations of anomaly subspaces.

After the interesting anomaly subspaces are found, we need to compute the outlying degree of anomaly cells in each anomaly subspaces. By definition, this outlying degree can be computed at the same time recursively. Consider an example for a data set with four attributes $[A,B,C,D]$, the subspace with attributes $[A,B,C]$ is anomaly subspace, we want to compute the outlying degree of outliers in this subspace. By using apriori method, we first calculate the density of all cells in one dimensional subspace. Suppose the anomaly cell signatures corresponding to subspace A are $[A_i,*,*,*]$, similarly $[*,B_i,*,*]$ for subspace B, and $[*,*,C_i,*]$ for subspace C. If all of these subspaces satisfy the entropy condition, then let all of those anomaly cell signatures consist of set S. For these anomaly cells, we can compute their outlying degree as

$$\mathrm{PO}(\theta) = \frac{|\theta|}{Np(\theta)},$$

where θ represents any one of these cell signatures. Then, the density of all cells in two dimensional anomaly subspace are calculated. Suppose the anomaly cell signatures corresponding to subspace AB are $[\overline{A}_i,\overline{B}_i,*,*]$, similarly $[*,\overline{B}_i,\overline{C}_i,*]$ for subspace BC, and $[\overline{A}_i,*,\overline{C}_i,*]$ for subspace AC. Again, assume these subspaces satisfy entropy condition. Then, the outlying degree of any cells noted as $\bar{\theta}$ can be computed as

$$\mathrm{PO}(\bar{\theta}) = \frac{|\bar{\theta}|}{Np(\bar{\theta})} + \sum_{\theta\subset\bar{\theta},\theta\in S} \mathrm{PO}(\theta).$$

Reconstruct the set S by using those two dimensional anomaly cell signatures. We can compute the outlying degree of any anomaly cell θ in subspace ABC as

$$\mathrm{PO}(\theta) = \frac{|\theta|}{Np(\theta)} + \sum_{\bar{\theta}\subset\theta,\bar{\theta}\in S} \mathrm{PO}(\bar{\theta}).$$

By using this method, we can compute the outlying degree of any high-dimensional anomaly cells recursively at the same time to find interesting anomaly subspaces.

We can save the statistical results of a cell θ by using a cell summary $\mathrm{CS} = \{\theta, \mathrm{PO}(\theta), k\}$, where θ is its signature, $\mathrm{PO}(\theta)$ is outlying degree and k the number of data objects in the cell. $\mathrm{PO}(\theta)$ has value if and only if the associated subspace satisfy the entropy condition. Let the anomaly cell summaries in the interesting anomaly subspaces construct set IC (interesting cells). Finally, the first k outlying anomaly cells in set IC and their corresponding signatures are reported.

By using the definition of cell signature, the data structure hash index can be used to improve the efficiency of our algorithm. When we construct the cells of subspace and count the number of data objects in each cell, the hash index can be used to organize these cells by using their signatures. And the hash index can also be used in computing the outlying degree of anomaly cells.

The procedures of algorithm PODM can be summarized as follows.

Algorithm PODM (ϵ,ξ,ζ,k,d)

```
t = 1;
Form the set C_1 of one dimensional candidate subspaces;
Build cells of all 1-dimensional candidate subspace;
For t = 1 to d
  Scan database once;
    For each subspace s ∈ C_t
      Compute the density of every cells;
      Build B+ tree of non-empty cells;
      Compute geni-entropy H(s);
      Compute outlying degree O(s);
      If H(s) < ε then
        The t-dimensional possible anomaly subspace
        A_t = A_t ∪ s;
        Compute the outlying degree PO(θ) of cell θ
        where p(θ) < ζ by using set S;
        Keep the cell θ in the set NewS;
        Build B+ tree of set NewS;
      End
      If no more possible anomaly subspaces are found;
      break;
      If H(s) < ε and O(s) > ξ then
        IA=FormIA(s), IC=FromIC(c), c ⊂ IA;
      End
    End
    S=NewS;
    Generate candidate subspaces set C_{t+1} by using A_t;
  End
  Output the k most outlying degree outliers and their
  signatures which belong to the set IC.
```

The procedure FormIA(s) is used to construct the interesting anomaly subspace set IA. It can be summarized as follows: For any subspace $c \in IA$, if subspace s is an interesting anomaly subspace and $s \subset c$, discard s; Otherwise, discard all the subspaces c which $c \subset s$ and insert s to the set IA.

The procedure FormIC(c) is used to form the interesting anomaly cell set IC in which the cells belong to the interesting anomaly subspace set IA.

In our algorithm, we use a few parameters h, ϵ, ξ, ζ and k. The parameter k is given by user. The parameter h is used to discretize the continuous value attribute for constructing cells. The choosing of parameter h is critical. If h is large, we cannot find any outliers. However, if h is small, too many outliers may be found. We use the entropy condition to choose suitable h, i.e. for an one dimensional subspace X, choosing h such that $H(X) < \epsilon$. If we cannot find such h in the subspace X, by our algorithm, this attribute will not be considered again. The parameters ϵ and ξ are entropy and outlying degree thresholds of subspace respectively, which should be adjusted and decided in each iterations of practical computation such that the k outliers could be found. The parameter ζ is used to decide the non-dense cells if the associated subspace satisfies the entropy condition. For simplicity, the ζ is always selected such that approximate k cells are specified to be non-dense.

6. Simulations and discussion

6.1. Setup

We evaluate our theory and algorithms by using a PC with 1.7 GHz Pentium IV processor and 512 MB of memory, running Windows XP, and programmed by MATLAB. Three data sets in UCI depository and one synthetic data set are used to evaluate the accuracy and efficiency of our algorithm. The first data set has both categorical and continuous attributes. The second data set has only continuous attributes. The third data set has only categorical attributes. The data sets are as the following data.

6.1.1. Credit approval data

Credit approval data set contains 690 credit card applications. All attribute names and values have been changed to meaningless symbols to protect confidentiality of the data. This dataset has a good mix of attributes – continuous, nominal with small numbers of values, and nominal with larger numbers of values. There are also a few missing values. The data are classified as two classes. One is marked as $'+'$ with 307 instances, another is marked as $'-'$ with 383 instances. To test our projected outlier detection

algorithm, we randomly select 10 instances in ′+′ class and combine them with the ′−′ class to form a new data set. In this data set, the data in ′+′ class is about 2.5%.

6.1.2. Wisconsin diagnostic breast cancer data

Wisconsin diagnostic breast cancer data set contains 569 diagnostic instances. Attributes are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. The data have two classes. One class is diagnosed as benign with 357 instances, another class is diagnosed as malignant with 212 instances. The data has 30 real-valued attributes. We randomly select 10 instances in class malignant and combine class benign to form a new data set. The instances diagnosed as malignant are about 2.7%.

6.1.3. Mushroom data

This data set includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family. Each species is identified as definitely edible, definitely poisonous. This data set contains 8124 samples, which are classified to two classes. One class is edible with 4088 samples, another class is poisonous with 3916 samples. All 22 attributes are nominally valued. Since the edible samples are distributed sparsely, we randomly select 25 edible samples and combine class poisonous to form a new data set. The edible samples are about 0.6%. We use our algorithm to detect the edible samples in this new data set.

6.1.4. Synthetic data

There are very few publicly mixed-attribute data set with labeled projected outliers. We write a program to produce such data set which has mixed-attribute data. The user supply the cardinality and dimensionality of data set, subspace clusters and projected outliers. In our paper, we will produce a few synthetic data sets which have different dimensionality, cardinality and different dimensionality of subspace clusters. In all of cases, the projected outliers are sparsely embedded in three dimensional subspace in which the data are well clustered.

6.2. Simulation results

6.2.1. Performance

We run our algorithm on these four data sets. In the credit approval data set, we set the initial entropy of anomaly subspace less than 2, initial outlying degree of that is less than 60 and bigger than 2. The outlying degree of interesting anomaly subspace is greater than 1. For the Wisconsin diagnostic breast cancer data, we set the initial entropy of anomaly subspace is less than 2, and outlying degree of that is less than 15 and greater than 0.65. The outlying degree of interesting anomaly subspace is greater than 4. For the mushroom data set, the entropy threshold is 1.5 and the outlying threshold is greater than 1 and less than 20. The threshold of interesting anomaly is equal to 1. For synthetic data set, we produce 10,000 samples, with about 0.4% outliers embedded in three dimensional subspace. The entropy threshold is equal to 4, and the outlying degree is greater than 1. The upper threshold of outlying are not set. The first k outlying degree instances are required to be reported, where the k is the parameter supplied by user. However, since the degree of different anomaly cells may be the same, in practice, more than k projected outliers will be reported. The detection rates and false positive rates are shown in Table 1.

From Table 1, we can see that our algorithm can detect the known outliers efficiently. Because some of normal data in data sets $1-3$ show outlying property in some subspaces, compared with other outlier detection algorithm in full dimensional space, the false positive rate may be high if we still use the known class label to decide whether such instances are outliers. Actually, the normal instances, which are considered to be projected outliers, also can give us some knowledge in some sense (Aggarwal & Philip, 2005). For the synthetic data set, the normal data are not projected outliers in any subspace. So the false positive detection rate is low.

Table 1
The performance of different data sets

Accuracy	Credit approval	Breast cancer	Mushroom	Synthetic
Detection rate	0.7	0.8	0.96	0.97
False positive rate	0.3	0.2	0.29	0.03
Execution time (s)	18.9	244.4	2256.6	879.8

6.2.2. The performance with different lattice scales

The performance of our algorithm not only depends on the entropy and outlying thresholds, but also the equiwidth parameter to discretize the continuous attribute value of data, i.e. the different lattice scales. With the different lattice scales, under the same entropy and outlying threshold values as that in last section, the execution time, detection rate and false positive rate curves are drawn in Figs. 3–5, respectively. Since the mushroom data set has only categorical attributes, only three curves are drawn in these figures.

From Fig. 3, with the increase of lattice level, we could observe that the execution time on credit and breast cancer data set are first increasing, then it will decrease. The reason of execution time increasing is that the cells are increasing when the lattice level increases. However, when the lattice level is large enough, the entropy of some subspaces will increase and may be larger than the entropy threshold, so some subspaces will be considered to be non-intersecting which will not be handled. Thus the execution time is decreased. The execution time on synthetic data set has similar phenomena, from Fig. 3, we can see that the execution time first is decreasing, then is increasing because more cells need to be calculated while the entropy and outlying of subspace is still less than the threshold. And when the lattice level is increasing more higher, the execution time will decrease again with very bad detection rate, which is not drawn here.

From Figs. 4 and 5, we could observe that the detection rate first increase then decrease and the false positive rate first decrease then increase. When the lattice level is low, the projected outliers lies in the normal cells, thus the projected outliers cannot be identified. And some normal instances are consider to be outliers by

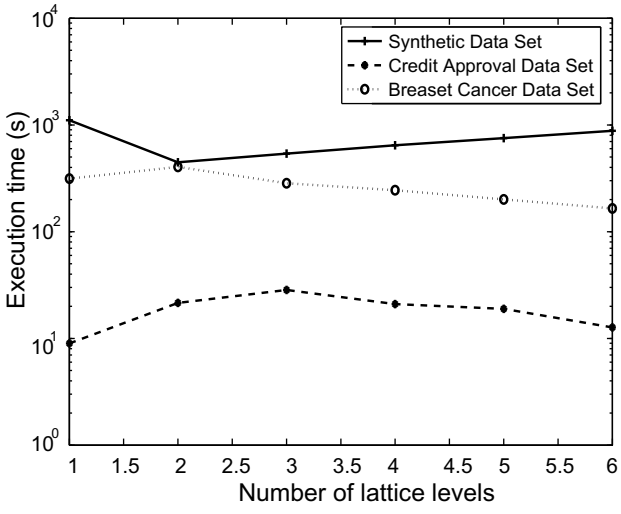

Fig. 3. Execution time variation with increasing lattice levels.

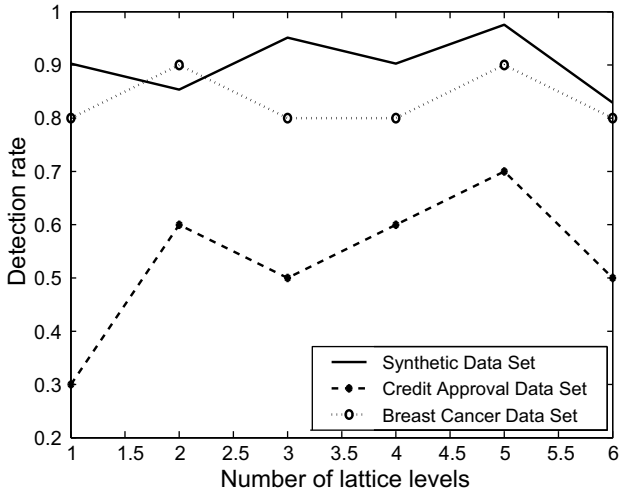

Fig. 4. Detection rate variation with increasing lattice levels.

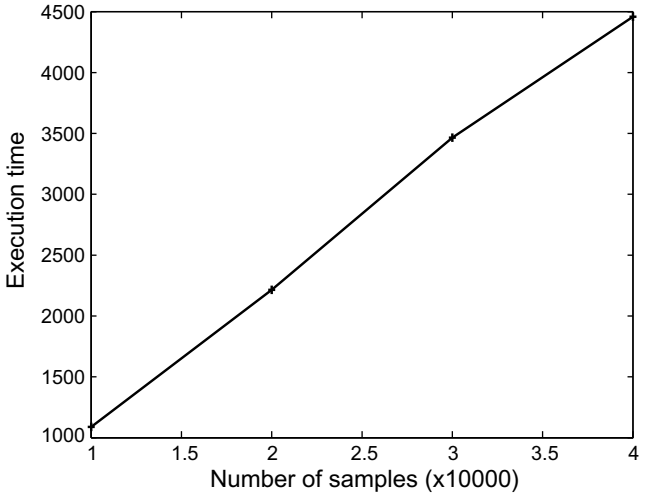

Fig. 6. Execution time with the size of synthetic data set.

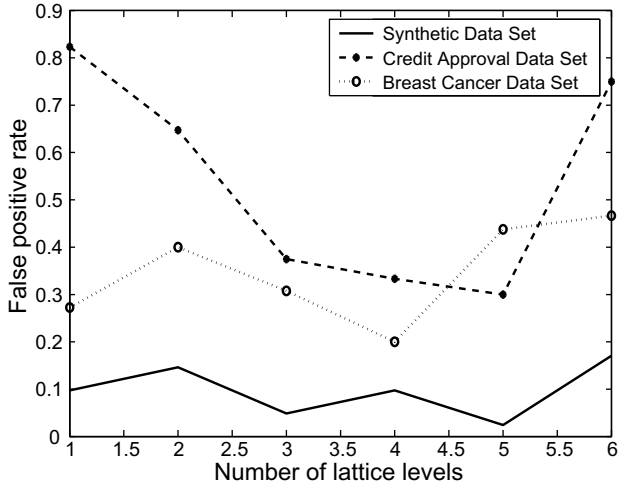

Fig. 5. False positive rate variation with increasing lattice levels.

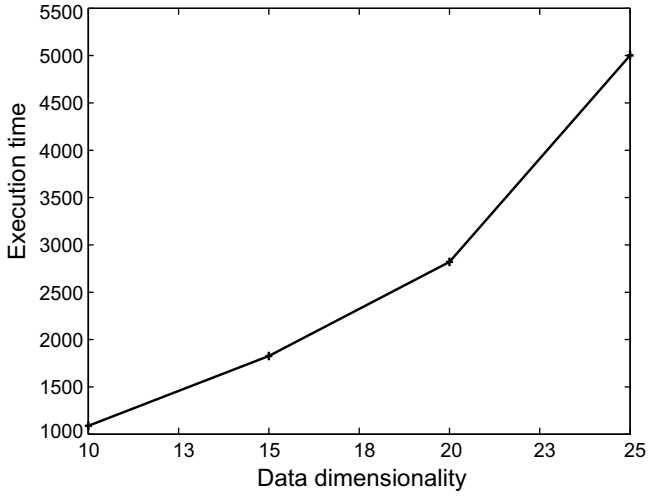

Fig. 7. Execution time with the dimensionality of synthetic data set.

calculating the outlying degree. So, the detection rate will be low and false positive rate will be high. When the lattice level is high, the number data cells are large, which means that the data cells are more sparse. In this case, some normal instances are also be consider to outliers and their outlying degree may be high. Thus, the detection rate is decreased and false positive rate is increased. The lattice level affects the execution time and accuracy very much. By using the method introduced in last section, the optimal grid size can be decided approximately.

6.2.3. The scalability test on synthetic data set

By using the synthetic data set, we also perform a few scalability test. For the test of scalability with different data set size, we produce a few different size synthetic data sets, whose dimensionality equals to 10. Four clusters embedded in subspace with dimensionality 6 and projected outliers are in three dimensional subspace. With the same entropy and outlying thresholds, Fig. 6 shows the scalability of our algorithm when the data set size is increasing. The scalability of the algorithm with different dimensionality is illustrated in Fig. 7. A few synthetic data sets for illustration of scalability with different dimensionality are produced. These data sets have different dimensionality of data, embedded clusters and projected outliers. Higher the dimensionality of data sets are, the larger the dimensionality of embedded clusters and projected outliers are. From Figs. 6 and 7, we could observe that the computation complexity increases as that of aprori algorithm. It is acceptable since many subspaces are pruned by using our method.

7. Conclusions

In this paper, combined with information entropy, a novel measure of anomaly subspace is proposed. In anomaly subspace, the meaningful outliers can be detected and explained efficiently. Theoretical analysis about this measure is presented. The outlying degree of projected outlier is defined, which has good explanations. By using this definition, the projected outliers in subspaces with different sizes of dimensionality can be detected and reported efficiently. A bottom-up method is proposed to find the interesting anomaly subspaces and calculate the outlying degree of projected outliers in high-dimensional mixed attribute data set. Thus, the projected outliers in different subspaces can be detected. Unlike the previous methods, the dimensionality of projected outliers need not to be decided beforehand. Experiments on synthetic

and small real data sets show the efficiency of our approach. However, when the size and dimensionality of data set are high, the computation complexity is high because of multiple scans of database. Our future work is to derive a fast algorithm which can reduce the computation burden efficiently.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (60702071), Program for New Century Excellent Talents in University (NCET-06-0811), Application Research Foundation of Science and Technology Bureau of Sichuan Province of China, Grant No. 2006J13-065 and the Key Program of the Youth Science Foundation of UESTC (JX0745).

References

Aggarwal, C. C., & Philip, S. Y. (2005). An effective and efficient algorithm for high-dimensional outlier detection. *The VLDB Journal, 14*, 211–221.

Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft U. (1999). When is nearest neigbor meaningful. In *Proceedings of ICDT conference*.

Cheng, C. H., Fu, A. W. -C, Zhang Y. (1999). Entropy-based subspace clustering for mining numerical data. In *Proceedings of the ACM SIGKDD international conference knowledge discovery and data mining* (pp. 84-93).

Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science, 17*, 235–255.

Breuning, M. M., Kriegel, H. P., Ng, R. T., Sander, J. (2000). LOF: Identifying density-based local outliers. In *Proceedings of the ACM SIGMOD (pp. 93–104)*.

Eskin, E., Arnold, A., Prerau, M., Portoy, L., & Stolfo, S. (2002). *A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data*. *Applications of data mining in computer security*. Kluwer.

Fawcett, T., & Provost, F. (1997). Adaptive fraud detection. *Data Mining and Knowledge Discovery, 1*, 291–316.

Gamberger, D., Lavac, N., Groselj, C. (1999). Experiments with noise filtering in a medical domain. In *Proceedings of the international conference on machine learning*.

Ghoting, A., Otey, M. E., Partheasarathy, S. (2004). Loaded: Link-based outlier and anomaly detection in evolving data sets. In *Proceedings of the international conference on machine learning*.

Guha, S., Rastogi, R., & Shim, K. (2000). ROCK:A robust clustering algorithm for categorical attributes. *Informations Systems, 25*, 345–366.

Knorr, E., Ng, R. (1998). Algorithms for mining distance-based outliers in large data sets. In *Proceedings of the VLDB conference (pp. 392-403)*.

Knorr, E., Ng, R. (1999). Finding intensional knowledge of distance-base outliers. In *Proceedings of the VLDB conference* (pp. 211-222).

Knorr, E., Ng, R., & Tucakov, V. (2000). Distance-based outlier: Algorithms and applications. *VLDB Journal, 8*(3–4), 237–253.

Lee, W., Stolfo, S. J., Mok, K. W. (1998). Mining audit data to build intrusion detection models. In *Proceedings of the international conference on knowledge discovery and data mining (KDD-98) (pp. 66-72)*.

Li, W., Qian, W.N., Zhou, A.Y., Jin, W., Yu, J.X. (2003). HOT: Hypergraph-based outlier test for categorical data. In *Proceedings of the 7th Pacific-Asia Conference*, *PAKDD 2003* (pp. 399–410).

Ng, K. K., Fu, W. C., & Wong, C. W. (2005). Projective clustering by histograms. *IEEE Transactions on Knowledge and Data Engineering, 17*(3), 369–383.

Otey, M. E., Ghoting, A., & Srinivasan, P. (2006). Fast distributed outlier detection in mixed-attribute data sets. *Data Mining and Knowledge Discovery, 12*(3), 203–228.

Papadimitriou, S., Kitawaga, H., Gibbons, P. B., Faloutsos, C. (2003). LOCI: Fast outlier detection using the local correlation integral. In *Proceedings of the international conference on data engineering*.

Parsons, L., Haque, E., Liu, H. (2004). Evaluating subspace clustering algorithms. In *Proceedings of the fourth SIAM international conference data mining, workshop clustering high dimenisional data and its applications*.

Parsons, L., Haque, E., & Liu, H. (2004). Subspace clustering for high dimensional data: A review. *ACM SIGKDD Explorations News-letter, 6*(1).

Patrikainen, A., & Meilla, M. (2006). Comparing subspace clustering. *IEEE Transactions on Knowledge and Data Engineering, 16*(7), 902–916.

Penny, K. I., & Jolliffe, IT. (2001). A comparison of multivariate outlier detection methods for clinical laboratory safety data. *The Statistician, Journal of the Royal Statistical Society, 50*, 295–308.

Procopiuc, M., Jones, M., Agarwal, P. K., Murali, T. M. (2002). A Monte-Carlo algorithm for fast projective clustering. In *Proceedings of the ACM SIGMOD*.

Ramaswamy, S., Rastogi, R., & Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. *Proceedings of the ACM SIGMOD*, 427–438.

Sequeira, K., Zaki, M. (2002). ADMIT: Anomaly-based data mining for intrusions. In *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining*.

Simovici, D. A., Cristofor, D., Critofor, L. (2005). Generalized entropy and projection clustering of categorical data. In *Proceedings of the 4th European conference on principles of data mining and knowledge discovery* (pp. 619–625).