适用于关联属性的样本自适应参数孤立点检测法*

刘胜宗[1a], 樊晓平[1a,2], 廖志芳[1b]

(1. 中南大学 a. 信息科学与工程学院; b. 软件学院 长沙 410075; 2. 湖南财政经济学院 网络化系统研究所, 长沙 410205)

摘　要: 为解决数据集中关联属性之间的干扰问题,通过引进 Mahalanobis 距离,并对传统的 k 近邻孤立点检测方法进行了改进,提出了一种新的基于样本的参数选取方法。该方法通过训练数据集中的正常数据和孤立点数据,以获得最优的 k 距离值和阈值。实验仿真结果表明,提出的算法有更高的准确率,同时降低了误检率。

关键词: 孤立点检测; 关联属性; 样本自适应; Mahalanobis 距离

中图分类号: TP301.6　　**文献标志码**: A　　**文章编号**: 1001-3695(2012)09-3259-04

doi:10.3969/j.issn.1001-3695.2012.09.015

Sample-adaptive-parameters outlier detection method for associated-attributes

LIU Sheng-zong[1a], FAN Xiao-ping[1a,2], LIAO Zhi-fang[1b]

(1. *a. School of Information Science & Engineering*, *b. School of Software*, *Central South University*, *Changsha* 410075, *China*; 2. *Laboratory of Networked Systems*, *Hunan University of Finance & Economics*, *Changsha* 410205, *China*)

Abstract: In order to solve the interfering problem of associated-attributes in datasets, this paper improved the traditional k-nearest neighbor outlier detection method by the introduction of Mahalanobis distance, and proposed a new sample-based parameters selection method which gained the optimization k-distance value and threshold by training the normal and outlier data in the sample dataset. Simulation results illustrate the proposed algorithm has higher accuracy, lower false detection rate.

Key words: outlier detection; associated-attributes; sample-adaptive; Mahalanobis distance

0　引言

孤立点检测技术在许多领域都有着重要的应用,如欺诈检测、网络入侵检测、安全欺骗行为检测、医疗诊断、水质污染源检测、审计证据获取等。在这些应用中,孤立点数据往往比正常数据更加有吸引力,因为孤立点数据集中可能含有大量潜在的有用信息,因此孤立点检测和预测的问题依然吸引着很多的关注。

一般地,将行为和特征与数据集中其他数据有显著异常的点称为孤立点。大部分既有的相关研究通过计算并比较各数据对象之间的距离或者是密度来检测孤立点。目前孤立点检测方法主要有基于聚类的方法、基于数据分布统计的方法、基于距离的方法、基于密度的方法、基于偏离的方法以及基于支持向量机的方法。其中基于距离和基于密度的方法吸引了人们更多的研究目光。

在基于密度的孤立点检测方法中,由 Breunig 等人提出的局部孤立因子方法(LOF)通过计算对象邻域范围内的局部密度来进行孤立点检测,其中 LOF 值大的对象则更有可能被标记为孤立点。为了区分孤立点数据和正常数据,LOF 方法通过设置阈值来进行控制。因此,LOF 方法中至少需要设置两个参数,一个是阈值,另外一个是邻域里邻居个数 K。其中 K 值和阈值的设置直接关系到算法的检测结果,如果阈值设置得越小,意味着孤立点的个数将大大增加,这样使得误检率(又称假阳性率,false-detection rate)增加;如果阈值设置得过大,将导致丢失许多真正的孤立点。为了消除这些参数设置带来的影响,近年来大部分的研究者采用 Top-N LOF 检测方法。该方法对 LOF 进行了改进,将阈值设置改进为选取 LOF 值排在前 N 位的数据对象作为孤立点检测出来,但是该方法同样会因为 N 的设置而影响算法的准确率。

而与 LOF 算法同时期提出的 Top-N K 最近邻孤立点检测算法(kNN),是经典的基于距离的孤立点检测算法,该算法把离群度表示为一个数据对象到它 K 近邻邻居的距离之和,离群度越大,越可能是孤立点。而 Top-N kNN 算法把具有最大离群度的前 N 个对象看做孤立点。Top-N kNN 算法和 Top-N LOF 算法存在同样的由参数设置导致的问题。

另一方面,这些参数设置一般是由用户完成的,因此这些参数的设置还包含了用户对孤立点的潜意识观念。然而让用户直接提供一些正常数据和孤立点数据的样本显得更加实际,因为这些用户是相关领域的专家,他们更擅长采用人为的方式找出孤立点并提供成相关样本。

目前基于距离的方法大部分采用欧氏距离来计算距离,这样就没考虑到数据对象属性之间的相互影响。在很多实际的数据集中,某些属性是有关联性的,例如在某种鱼类生长情况的数据集中,鱼身长度和重量以及生长时间这三个属性存在着

收稿日期: 2011-11-21; **修回日期**: 2012-03-05　　**基金项目**: 国家科技支撑计划资助项目(2012BAH08B00)

作者简介: 刘胜宗(1986-),男,湖南邵阳人,博士研究生,主要研究方向为数据挖掘、智能信息处理等(lsz179@163.com);樊晓平(1961-),男,教授,博导,博士,主要研究方向为无线传感器网络、网络化系统控制、智能信息处理、智能交通系统等;廖志芳(1968-),女,副教授,硕导,博士,主要研究方向为数据挖掘、推荐系统等.

关联关系,但这种关联关系不是简单的线性关系。本文将选取能同时考虑对象属性间相互影响的 Mahalanobis 距离进行计算。本文提出结合 Mahalanobis 距离和基于样本的方法对 Top-N kNN 算法进行改进。

1 基于距离的孤立点检测问题描述

一个数据对象表示为若干个属性值的集合。基于距离的孤立点表示为数据集中与大多数点之间的距离均大于某个阈值的点,具体描述为在数据集 X 中,当且仅当 X 中至少有 r 部分的对象与 O 距离大于设定的阈值 θ 时,O 是一个孤立点。对孤立点进行检测之前,需要给出以下定义。

定义 1 k-近邻邻居集。给定一个含有若干数据对象的集合 D,一个数据对象 p,一个从 p 到 D 中对象的距离集合 dist,一个正整数 k,对于 D 中某一对象 q:如果恰好存在 $k-1$ 个对象 r 使得 $\mathrm{dist}(p,q)\geqslant\mathrm{dist}(p,r)$,则称 q 是 p 关于 D 的第 k 近邻邻居,又称为第 k 近邻,记做 $q=nn_k(p,D)$,而 $\mathrm{dist}(p,q)$ 称为 p 的 k-距离(记做 k-distance),距离不超过 k-距离的范围称为 k 邻域,p 的 k-近邻邻居集合则用 Q 表示为

$$Q=\{q\mid q\in D,\mathrm{dist}(p,q)<k\text{-distance}\}\cup\{nn_k(p)\}$$

如果 p 在 D 中,则 $p=nn_1(p,D)$,否则 $p=nn_1(p,D)$ 是离 p 最近的邻居。一般地,在后面的讨论中,都将 p 看做是在 D 中。

定义 2 k-近邻距离和。给定一个包含 N 个数据对象的集合 D,一个 D 中对象的相互距离集合 dist,一个 D 中的对象 p,一个整数 k,$1\leqslant k\leqslant N$,那么 p 关于 D 的 k-近邻距离和定义为 $W_k(p,D)$,其计算式为

$$W_k(p,D)=\sum_{i=1}^{k}\mathrm{dist}(p,nn_i(p,D)) \tag{1}$$

传统的 k-近邻孤立点检测方法采用 $W_k(p,D)$ 表示某个对象与其邻居的相似度,例如某个对象 p 的 $W_k(p,D)$ 越小,表示 p 与其邻居的相似性越大。在传统的 kNN 算法和 Top-N kNN 算法中,均使用 k-近邻距离和来表示 p 与其邻居之间的相似性或者差异度,并用来区分孤立点与正常对象。由于这种方式只考虑 p 与其邻居之间的这种关系,而没有考虑到 p 的整个邻域范围的密集性,因此准确率不高;同时容易受到孤立点聚集的影响,使得一些正常的数据对象被错误地当做孤立点,造成误检率上升,而这些被误检的对象往往处于正常数据集的边缘位置。为了避免这种情况,同时提高孤立点检测的准确率,本文对 k-近邻距离和进行改进,采用 k-近邻内部距离和替代 k-近邻距离和。

定义 3 k-近邻内部距离和。给定一个包含 N 个数据对象的集合 D,一个 D 中对象的相互距离集合 dist,一个 D 中的对象 p,一个整数 k,$1\leqslant k\leqslant N$,那么 p 关于 D 的 k-近邻内部距离和定义为 $W_k^*(p,D)$,其计算式为

$$W_k^*(p,D)=\sum_{q\in Q,r\in Q,q\neq r}\mathrm{dist}(q,r) \tag{2}$$

其中:Q 是定义 1 中的 k-近邻邻居集合。

某个对象 p 的 $W_k^*(p,D)$ 值越小,表示 p 与其邻居的偏离也越小,p 的邻域范围较密集;相反,$W_k^*(p,D)$ 值越大,表示 p 与其邻居的偏离也就越大,p 的邻域范围较稀疏,p 很有可能成为孤立点。

由于 $W_k^*(p,D)$ 受 k 参数影响较大,因此采用 k-近邻内部距离和与 k-近邻邻居间的边数 $\frac{k(k-1)}{2}$ 的比值来表示某对象与 k-近邻邻居间的分散度 $D_k(p)$,其计算式为

$$D_k(p)=\frac{2W_k^*(p,D)}{k(k-1)} \tag{3}$$

那么 $D_k(p)$ 值越大,表示对象 p 越可能成为孤立点。而孤立点检测问题就是将数据集中该值大于某阈值的异常数据对象全部检测出来。

2 解决方法

对孤立点进行检测,首先要进行参数的设定,在上面问题的定义中,参数 k 和 θ 的取值对算法的性能产生直接的影响,因此,如何选取这些参数是至关重要的问题。算法在针对关联属性的数据集中,因为欧式距离对对象的各个属性采用同等的影响来进行计算,所示并不能体现关联属性间的这种相互约束关系;同时为了减小值域较大的属性对检测结果产生偏差,本文采用 Mahalanobis 距离和均值标准差来处理数据。

2.1 基于样本的参数选取

本文提出基于样本的参数选取方法,通过对孤立点和正常数据的计算分析,自适应地选取合适的 k 和 θ 值。

在 kNN 和 LOF 孤立点检测算法中,都存在参数 k 的选取问题,而 k 值的选取直接影响到孤立点检测算法的效果。在一些研究中,对 k 的选取采用穷尽试探法,将 k 从 1 开始一步步地试探,直到选取出最有效的 k 值。这种方法的优点是简单,但缺点很明显:过于盲目,容易出现局部最优的情况;时间损耗较大,过多地计算无用的 k 值。因此本文提出一种新的基于样本的参数选取方法。根据用户提供的样本(包括已标记的正常数据和孤立点数据),按照如下策略进行:

在 p 对象的 k 邻域里,内部的邻居点相互之间的分散度 $D_k(p)$ 能够区分孤立点与普通数据对象之间的差异,因此使用 $D_k(p)$-k 图来观察 k 对 $D_k(p)$ 的影响。图 1 表示某一孤立点和一正常数据对象的 $D_k(p)$-k 图。

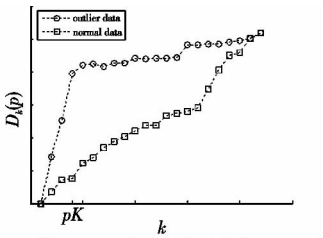

图1 某孤立点对象和正常对象的$D_k(p)$-k图

图 1 中,孤立点数据与正常数据有明显的区分,孤立点数据在 k 值较小时,其邻居相互距离较远,因此 $D_k(p)$ 上升得很快;k 增大到一定大小时,受正常数据的影响,随着 k 的增大,$D_k(p)$ 趋于平缓,而正常数据上升一直较平缓;当 k 到达某大小时,受孤立点影响,有一段上升速度明显加快。因此 k 应该取两条线之间的纵向间隔最大时的 k 值,文中所选的 pK 位置使得两线之间的距离差值最大,因此选取 $k=pK$,这样就将孤立点和正常数据的差异最大化,保证了后续的孤立点检测算法的准确性。从图 1 中也可以看出,当 $k=1$ 或 $k=N$(N 表示数据集的大小)时,所有数据对象的 $D_k(p)$ 值相等,$k=1$ 时 $D_k(p)$ 为 0,而 $k=N$ 时为所有对象相互间的距离和除以总边数 $N(N-1)/2$。

在 Top-N kNN 孤立点检测过程中,N 的选取往往是根据经验或者是事先的估计进行的,表示将权值进行排序,排在前 N 位的当做孤立点输出。可以设想,当 N 取值小于实际孤立点个数时,会导致漏检率增大,使得一些孤立点未能正常地被检测出来;当 N 取值大于实际孤立点个数时,会把正常数据也当做孤立点输出,增大错检率。因此设置一个阈值 θ,当权值大于此阈值时,才把对象当做孤立点输出。在 k 值选取过程中,同时确定该阈值 θ,θ 值由式(4)给出:

$$\theta = \frac{\overline{D_k(p_O)} + \overline{D_k(p_N)}}{2} \tag{4}$$

其中:p_O 表示样本孤立点对象;p_N 表示正常数据对象。在样本集中,θ 为孤立点对象的平均 $D_k(p)$ 值和正常对象的平均 $D_k(p)$ 值的平均值。

2.2　基于 Mahalanobis 距离的数据处理

在孤立点检测之前,首先需要对数据进行预处理。由于对象的不同属性间的值域不同,有些值域的范围非常庞大,而有些却很小,这样会造成值域大的对结果的影响偏大;同时特定的数据有特定的单位,不同的单位度量会对距离的计算结果产生影响。所以,为了保证后面的实验结果,首先对数据进行标准化处理,标准化处理方法采用标准差处理法。标准化方式如下:

设 $X = \{X_i \mid X_i = (X_{i1}, X_{i2}, \cdots, X_{ij}, \cdots, X_{im}), i = 1, 2, \cdots, n; j = 1, 2, \cdots, m\}$ 是待进行孤立点数据挖掘的数据集合,X_i 表示第 i 个对象,共有 n 个对象;X_{ij} 表示第 i 个对象的第 j 个属性值,共有 m 个属性。

令 $\overline{X_j}$、σ_j 分别表示第 j 个属性的均值和标准差,即

$$\overline{X_j} = \frac{1}{n}\sum_{i=1}^{n} X_{ij} \tag{5}$$

$$\sigma_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \overline{X_j})^2} \tag{6}$$

则标准化数据的公式(X_{ij}^* 为处理后的数据)为

$$X_{ij}^* = \frac{X_{ij} - \overline{X_j}}{\sigma_j} \tag{7}$$

经过标准化后的对象属性都是无量纲的值,既保证了孤立点检测不受值域大的属性影响,也不受不同单位度量的影响。

由于对象的属性间存在关联关系,传统的欧氏距离体现不了这种关联关系,因此,本文采用更能体现各种特性之间联系的 Mahalanobis 距离来度量对象之间的差异。

对于数据集 $X = \{X_i \mid X_i = (X_{i1}, X_{i2}, \cdots, X_{ij}, \cdots, X_{im}), i = 1, 2, \cdots, n; j = 1, 2, \cdots, m\}$,$i$ 表示对象的下标,j 表示对象属性的下标,则可以假定该数据集中的任意对象都是服从同一分布,那么任意两对象 X_i、X_j 之间的 Mahalanobis 距离可以表示为

$$D(X_i, X_j) = \sqrt{(X_i - X_j)^{\mathrm{T}} \Sigma^{-1} (X_i - X_j)} \tag{8}$$

其中:Σ^{-1} 表示数据集 X 的协方差矩阵的逆矩阵,而数据集 X 的协方差矩阵 Σ 为

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1m} \\ \Sigma_{21} & \Sigma_{22} & \cdots & \Sigma_{2m} \\ \vdots & \vdots & & \vdots \\ \Sigma_{m1} & \Sigma_{m2} & \cdots & \Sigma_{mm} \end{bmatrix} \tag{9}$$

其中:

$$\Sigma_{kl} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ik} - \overline{X_k})(X_{il} - \overline{X_l}) \tag{10}$$

$1 \leqslant k, l \leqslant m$,$\overline{X_k}$ 表示 X 在第 k 个属性上的均值。

2.3　改进的孤立点检测算法

以下为该方法相应的算法描述。孤立点检测算法的步骤如下:

输入:样本数据集矩阵 Y(孤立点和正常数据已标出)和待测数据集矩阵 X。

输出:孤立点集合 O。

a)数据标准化处理。根据式(5)和(6)分别计算样本数据集 Y 和待测数据集 X 的各个属性的均值和方差;然后根据式(7)进行标准化处理,得到标准的样本数据集 Y^* 和标准的待测数据集 X^*。

b)根据式(8)~(10)计算出标准的样本数据集 Y^* 的相应的 Mahalanobis 距离矩阵 $D(Y_i^*, Y_j^*)$。

c)从 Y^* 任取一个孤立点对象 m 和一个正常对象 n,根据式(2)(3),计算它们的 k 从 1 到 $|Y^*|$ 相应的 $D_k(p)$ 值,分别记为 $D_k(m)$、$D_k(n)$。令 $D_k(m) - D_k(n)$,将其结果的最大值对应的 k 值作为本方法所需的 k 值。

d)将 Y^* 的距离矩阵 $D(Y_i^*, Y_j^*)$ 根据式(2)(3)计算出各个对象的分散度 $D_k(p)$。

e)根据步骤 c)中求取的每个对象的分散度,计算出标准样本数据集中所有孤立点的分散度平均值和所有正常点的分散度平均值,并根据式(4)求出 θ 的值。

f)将步骤 a)中得到的标准待测数据集 X^* 根据式(8)~(10)计算 X^* 的 Mahalanobis 距离矩阵 $D(X_i^*, X_j^*)$,并根据式(2)(3)计算出 X^* 中各个对象的分散度 $D_k(p)$。

g)将 X^* 中分散度大于 θ 的数据对象输出到孤立点集合 O 中。

h)算法结束。

该算法通过对样本集进行训练,选取最优的参数,并将得到的参数用于待测集的孤立点检测,这样就有效地避免了参数选取的盲目性和局部最优性。

3　实验仿真及分析

在实验过程中,采用 ROC 分析法,选取假阳性率和检测率来衡量算法的效果。假阳性率(false positive rate)表示在检测过程中被误认为孤立点的正常数据对象数目和正常数据对象总数目的比值;检测率(detection rate)表示正确检测出的孤立点对象数目和所有孤立点对象总数目的比值。本文将一些常见的孤立点检测算法在合成数据和实际数据上分别进行了测试,并与本文提出的方法进行了对比。

3.1　二维合成数据实验仿真

该数据集包含两个属性 a 和 b,这两个属性并不是关联属性,生成该数据集时,将两个属性的取值范围进行了人为地限制,并按照一定的概率生成一部分不在该范围内的数据点,将其视为孤立点。在运行相关孤立点检测算法时,从该数据集中按一定的比例抽取正常数据和孤立点数据组成样本训练集。

在对合成数据进行孤立点检测时,数据对象总数为 2 000,样本孤立点个数为 10。样本孤立点对象和待测试数据对象的分布如图 2 所示。图中圆圈表示的点为样本孤立点,其余点表示待测试数据对象。

经过运行检测后,发现的孤立点分布如图 3 所示。图中检测出的孤立点使用圆圈标记。在合成数据上进行测试,检测率达到 97% ,而假阳性率则为 0. 16% ,体现出了良好的检测效果。

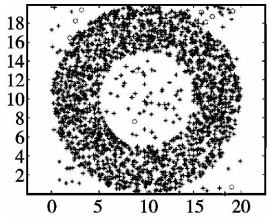

图2 合成数据样本孤立点及待测数据分布

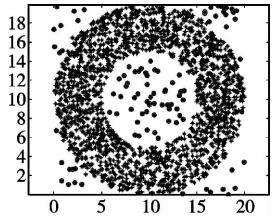

图3 检测出的孤立点分布

3.2 实际数据实验仿真

为了验证算法对于实际数据的有效性,本文从 UCI 公用数据库中选取了鲍鱼数据(abalone data)来进行测试,该数据集含有八维属性,拥有 4 177 个数据实例。选取其中的 Diameter-WholeWeight 两维进行实验,基于这两维的分布如 4 图所示。检测后的孤立点分布如图 5 所示。

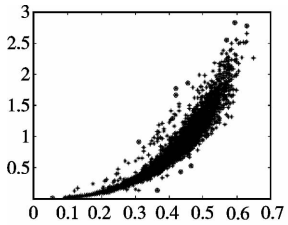

图4 鲍鱼数据直径-体重分布

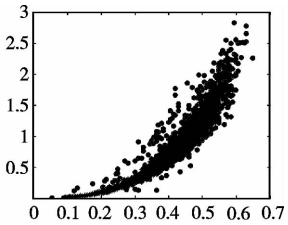

图5 二维鲍鱼数据孤立点检测结果分布

在对鲍鱼数据的 Diameter、WholeWeight 这两维进行检测后,检测率降低到 70% ,假阳性率增加到了 3. 17% ,导致这种情况的原因是这两维组成的子空间不足以代表鲍鱼数据的整体特征。因此,为了体现出相互关联属性之间的影响,对所有维度同时进行孤立点检测。

3.3 实验结果比对及分析

对合成数据和现实数据进行孤立点检测的实验仿真结果表明,本文提出的方法在检测率和假阳性率方面都得到了一定的改善,如表 1 所示。

表 1 四种孤立点检测方法在合成数据和实际数据上的效果比对

方法	合成数据		UCI 鲍鱼数据(八维)		UCI 鲍鱼数据(二维)	
	检测率 DR/%	假阳性率 FPR/%	检测率 DR/%	假阳性率 FPR/%	检测率 DR/%	假阳性率 FPR/%
本文方法	97	0.16	90	1.26	70	3.17
TOP-N kNN (N=400)	90	0.53	71	3.1	65	3.7
TOP-N LOF (N=400)	93	0.37	75	2.65	69	3.28
One class SVM	95	0.26	82	1.9	84	1.87

从表 1 中可以看出,对于合成的数据,也就是无关联属性的数据集,这种改善不是很明显,本文方法的检测率比现有的一些较常用算法的检测率只高出几个百分点,甚至与 One class SVM 相差不大,这就意味着在非关联属性的数据集中,本文方法与传统的孤立点检测算法相差不大。这是因为在使用 Mahalanobis 距离计算对象间的差异度时,考虑了数据集分布的协方差,而欧式距离则没有考虑,因此数据集中若存在关联属性,会导致这两种距离计算出的结果出现偏差。从理论上来讲,此时采用 Mahalanobis 距离更加准确,因为它排除了变量间的相关性干扰。当数据集中没有关联属性时,结果就不一样了,这时 Mahalanobis 距离退化成欧氏距离,也就是说,欧式距离是 Mahalanobis 距离的特例。

从表 1 中还可以看出,本文检测算法在八维的 UCI 鲍鱼数据中的性能都要优于其他算法,检测率几乎高出了 20 个百分点,而对于二维的鲍鱼数据集,本文算法的性能受到了限制。这是因为采用 Mahalanobis 距离在计算对象差异度时,数据集的属性维度需要适度,不能太小,太小影响准确性,而太大则可能导致算法的时间开销过于庞大。这也说明了对于具有多维关联属性的数据集而言,本文提出方法更有效。

假阳性率的变化趋势与检测率的趋势相反,当检测率提升时假阳性率则下降,反之亦然。而从表 1 的结果中可以看出,本文方法在提高了检测率的同时,有效地降低了假阳性率,既提高了准确性,又降低了误检率。

4 结束语

本文通过对 k-近邻距离和孤立点检测算法进行改进,采用 k-近邻内部距离和来衡量对象间的相似度,在此基础上利用 Mahalanobis 距离代替欧式距离来进行对象之间的距离度量;同时提出了基于样本集的参数自适应方法,减少了凭经验来设置参数的弊端,提高了孤立点检测准确度,降低了误检率;给出了相应的算法,通过实验仿真验证了本文提出的算法具有一定的优势。

参考文献:

[1] ZHANG Ke, MARCUS H, JIN Hui-dong. A new local distance-based outlier detection approach for scattered real-world data [C]//Advances in Knowledge Discovery and Data Mining. 2009: 813-822.

[2] FABRIZIO A, STEFANO B, CLARA P. Distance-based detection and prediction of outliers[J]. IEEE Tran on Knowledge and Data Engineering, 2006, 18(2): 145-160.

[3] LI Yuan, KITAGAWA H. DB-outlier detection by example in high dimensional datasets[C]//Proc of IEEE International Workshop on Databases for Next Generation Researchers. 2007.

[4] JULIUS E, COLIN P, REINHOLD B, *et al*. A new approach for outlier detection in near real-time[C]//Proc of the 4th UKSim European Symposium on Computer Modeling and Simulation. 2010.

[5] JEROEN H, IDIKO F, ERIC O, *et al*. Outlier detection with one-class classifiers from ML and KDD[C]//Proc of the 8th International Conference on Machine Learning and Applications. 2009.

[6] 田江,顾宏. 孤立点检测的一类支持向量机方法研究[J]. 小型微型计算机系统, 2008, 29(11): 2107-2110.

[7] ISAKSSON C, DUNHAM M H. A comparative study of outlier detection algorithms[C]//Proc of the 6th International Conference on Machine Learning and Data Mining in Pattern Recognition. 2009: 440-453.

[8] EDWIN M, RAYMOND T. Algorithms for mining distance-based outlier in large datasets[C]//Proc of the 24th VLDB Conference. 1998.

[9] HAUTAMAKI V, KARKKAINEN I, FRANTI P. Outlier detection using k-nearest neighbour graph[C]//Proc of the 17th International Conference on Pattern Recognition. 2004.

[10] SAHA B N, RAY N, ZHANG Hong. Snake validation: a PCA-based outlier detection method [J]. IEEE Signal Processing Letters, 2009, 16(6): 549-552.