计算机研究与发展 ISSN 1000-1239/CN 11-1777/TP
Journal of Computer Research and Development 45(7): 1189-1194, 2008

基于局部信息熵的加权子空间离群点检测算法

倪巍伟[1] 陈 耿[2] 陆介平[3] 吴英杰[1] 孙志挥[1]
[1](东南大学计算机科学与工程学院 南京 210096)
[2](南京审计学院审计信息工程实验室 南京 210029)
[3](江苏省镇江市科技局 江苏镇江 212002)
(niww2007@yahoo.com.cn)

Local Entropy Based Weighted Subspace Outlier Mining Algorithm

Ni Weiwei[1], Chen Geng[2], Lu Jieping[3], Wu Yingjie[1], and Sun Zhihui[1]
[1](*College of Computer Science and Engineering, Southeast University, Nanjing* 210096)
[2](*Laboratory of Audit Information Engineering, Nanjing Audit University, Nanjing* 210029)
[3](*Zhenjiang Science and Technology Bureau of Jiangsu Province, Zhenjiang, Jiangsu* 212002)

Abstract Outlier mining has become a hot issue in the field of data mining, which is to find exceptional objects that deviate from the most rest of the data set. However, along with the increase of dimension, some unusual characteristic appearance becomes possible, such as spatial distribution of the data, and the distance of full attribute space is no longer meaningful, which is called "curse of dimensionality". Phenomena of "curse of dimensionality" deteriorate lots of existing outlier detection algorithms' validity. Concerning this problem, a local entropy based weighted subspace outlier mining algorithm SPOD is proposed, which generates outlier subspace and weighted attribute vector of each data object by analyzing entropy of each attribute on the neighborhood of this data object. For a given data object, those outlier attributes which constitute this object's outlier subspace, are assigned with bigger weight. Furthermore definitions such as subspace weighted distance are introduced to make a density-based outlier processing upon the data set and get each data point's subspace outlier influence factor. The bigger this factor is, the bigger the possibility of the corresponding data point becoming an outlier is. Theoretical analysis and experimental results testify that SPOD is suitable for datasets with high dimension, and is efficient and effective.

Key words high dimensional data; outlier detection; information entropy; subspace mining; weighted vector

摘 要 离群点检测作为数据挖掘的一个重要研究方向,可以从大量数据中发现少量与多数数据有明显区别的数据对象."维度灾殃"现象的存在使得很多已有的离群点检测算法对高维数据不再有效.针对这一问题,提出基于局部信息熵的加权子空间离群点检测算法SPOD.通过对数据对象在各维进行邻域信息熵分析,生成数据对象相应的离群子空间和属性权向量,对离群子空间中的属性赋以较高的权值,进一步提出子空间加权距离等概念.采用基于密度离群点检测的思想,分析计算数据对象的子空间离群影响因子,判断是否为离群点.算法能够有效地适应于高维数据离群点检测,理论分析和实验结果表明算法是有效可行的.

关键词 高维数据;离群点检测;信息熵;子空间挖掘;权向量
中图法分类号 TP311

收稿日期:2007-08-15;修回日期:2008-01-09
基金项目:江苏省自然科学基金项目(BK2006095);教育部高等学校博士学科点专项科研基金项目(20040286009)

近年来,基于数据挖掘概念的离群点检测技术由于其独特的知识发现功能而得到了较深入的研究,提出了一些有效的检测算法并获得了应用.诸如Johnson等人提出的基于深度的算法DEEPLOC[1],Knorr等人提出的基于距离的算法FindAllOutsD[2],Breunig等人提出的带离群度的离群点检测算法LOF[3],Papadimitirou等人提出的LOCI算法等[4-7].

这些算法对于普通低维度数据具有良好的性能.当数据维度较高时,由于高维数据"维度灾殃"现象的存在,常规意义下全维度空间的距离不再有意义,这些算法在准确性和效率上可能急剧恶化.子空间挖掘是解决高维空间数据挖掘的一种有效方法,在关联规则发现、聚类等领域,子空间数据挖掘已经有了较多的应用.

针对高维数据空间离群点发现问题,提出一种基于信息熵的加权子空间密度离群点检测算法SPOD(subspace outlier detection),算法根据数据点在各维度上的局部信息熵确定相应的子空间,并对数据点间的距离进行重新定义,实现了基于密度的高维数据离群点检测.

1 问题描述和相关工作

1.1 优选子空间选取

Christian等人在文献[8]中提出了一种局部优选子空间(local subspace preference)思想,假设d维数据集D的属性集为$A=\{A_1,\cdots,A_d\}$,D中数据点p在属性A_i上的投影记为$\Pi_{A_i}(p)$,$N_\varepsilon(p)$为p的ε邻域(ε为距离半径),相关定义如下:

定义**1**. 属性方差(variance along an attribute).对$p\in D$,$A_i\in A$,$N_\varepsilon(p)$关于A_i的方差定义为

$$VAR_{A_i}(N_\varepsilon(p))=\frac{\sum_{q\in N_\varepsilon(p)}(dist(\Pi_{A_i}(p),\Pi_{A_i}(q)))^2}{|N_\varepsilon(p)|}.$$

定义**2**. 优选子空间维(subspace preference dimensionality).对$p\in D$,$\delta>0$,若$VAR_{A_i}(N_\varepsilon(p))\leqslant\delta$称$A_i$为$N_\varepsilon(p)$的优选子空间维.

根据定义可以求出各个数据点的优选子空间属性集.对高维数据集,这种方法存在一些不足:邻域半径ε和阈值δ的设置非常困难,而ε和δ又直接关系到数据点子空间的确定,进而影响挖掘算法的效率.在文献[8]提出的局部优选子空间思想的基础上,提出适用于高维数据离群点检测的基于信息熵的加权子空间选取方法.

1.2 基于信息熵的子空间选取

熵是信息论中用来描述信息和随机变量不确定性的重要工具,设X为随机变量,其取值集合为$S(X)$,$P(x)$表示X可能取值的概率,则X的熵定义为

$$E(X)=-\sum_{x\in S(X)}P(x)\log_2(P(x)).$$

变量的不确定性越大,熵也就越大,把它搞清楚所需要的信息量也就越大;熵值越小,不确定性越小.在此基础上,引入"局部属性熵"定义.

定义**3**. 数据点p的k-距离(k-distance).存在$o\in D$,使得至少有k个数据点$o'\in D$满足$dist(p,o')\leqslant dist(p,o)$,并且最多有$k-1$个数据点$o'\in D$满足$dist(p,o')<dist(p,o)$,则$p$的$k$-距离$k$-$dist(p)=dist(p,o)$.

定义**4**. p的k邻域(k neighborhood).$N_k(p)=\{x\in D|\ dist(p,x)\leqslant k\text{-}dist(p)\}$.

定义**5**. 局部属性熵(local entropy of attribute).对$p\in D$,$A_i\in A$,p关于A_i的局部属性熵定义为

$$LEA_{A_i}(p)=-\sum_{q\in N_k(p)}\frac{dist(\Pi_{A_i}(p),\Pi_{A_i}(q))-d_{\min}}{d_{\max}-d_{\min}}\cdot\log_2\left(\frac{dist(\Pi_{A_i}(p),\Pi_{A_i}(q))-d_{\min}}{d_{\max}-d_{\min}}\right),$$

$$d_{\max}=\max\{dist(\Pi_{A_i}(p),\Pi_{A_i}(q))\ |q\in N_k(p)\},$$

$$d_{\min}=\min\{dist(\Pi_{A_i}(p),\Pi_{A_i}(q))\ |q\in N_k(p)\}.$$

局部信息熵$LEA_{A_i}(p)$描述了数据点p及其邻近数据点在属性A_i上投影值的分布情况,其数值越大,说明以p为中心的数据点在属性A_i上表现出的不稳定性(非规范性)越大;数值越小,说明以p为中心的数据点在属性A_i上分布越趋规范.生成数据点p在各维上的局部属性熵后,进一步对p的离群子空间进行定义.

定义**6**. p的离群属性(outlier attribute).对$p\in D$,$A_i\in A$,若满足如下关系:

$$LEA_{A_i}(p)\geqslant\frac{\sum_{q\in N_k(p)}LEA_{A_i}}{|N_k(p)|},$$

则称属性A_i为p的离群属性.

由定义6可知,若p关于A_i的局部属性熵不小于其k邻域内数据点关于A_i的属性熵均值,说明p较其邻域内数据点关于A_i属性表现出更大的不确定性.从信息熵角度考虑,离群点的存在使得数据集

的不确定性增强[9-10], 而离群点所表现出的不确定性正是通过其在某些属性上的取值分布造成的. 因此, 把满足该条件的属性称为关于 p 的离群属性.

定义 7. p 的离群属性子集(distinct outlier subspace). 对 $p\in D$, p 的离群属性子集 $DIOS(p)=\{A_i\in A\mid A_i$ 为 p 的离群属性$\}$.

p 的所有离群属性构成 p 的离群属性子集, 这个属性子集对应 d 维属性集的投影子空间, 数据点 p 在这个子空间内表现出较强的不确定性.

2 子空间离群点检测算法 SPOD

2.1 LOF 算法

Breunig 等人在 2000 年提出了 LOF 算法, 算法通过对各个数据点的 k 邻域内数据点局部密度的定义, 引申出对数据点离群度的形式化描述, 主要定义如下:

定义 8. p 的核心距离(core distance). $p\in D$, ε 为距离参数值, $MinPts$ 为给定自然数, p 的核心距离

$$core\text{-}distance_{\varepsilon,MinPts}(p)=\begin{cases}\text{不作定义}, & |N_\varepsilon(p)|<MinPts,\\ MinPts\text{-}distance(p), & \text{否则}.\end{cases}$$

定义 9. p 关于 o 的可达距离(reachable distance). p 与 o 为 D 中数据点, $p\in N_\varepsilon(o)$, 则 p 关于 o 的可达距离定义为

$$reachability\text{-}distance_{\varepsilon,MinPts}(p,o)=\begin{cases}\text{不作定义}, & |N_\varepsilon(o)|<MinPts,\\ \max(core\text{-}distance_{\varepsilon,MinPts}(o),d(o,p)), & \text{否则}.\end{cases}$$

定义 10. p 的局部可达密度(local reachable density). $p\in D$, 参数定义如上, p 的局部可达密度定义为

$$lrd_{MinPts}(p)=1\Big/\frac{\sum_{o\in N_{MinPts}(p)}reachability\text{-}distance_{\infty,MinPts}(p,o)}{|N_{MinPts}(p)|}.$$

定义 11. p 的离群因子(outlier factor). $p\in D$, p 的离群因子定义如下:

$$OF_{MinPts}(p)=\frac{\sum_{o\in N_{MinPts}(p)}\frac{lrd_{MinPts}(o)}{lrd_{MinPts}(p)}}{|N_{MinPts}(p)|}.$$

数据点的离群因子值越高, 成为离群点的可能性越大.

2.2 SPOD 算法思想与相关定义

加权子空间离群点检测算法 SPOD 的算法思想与 LOF 相似, 区别在于算法主要在 p 的离群子空间内进行分析, 对离群属性和非离群属性分别赋以不同的权值.

定义 12. 属性权向量(attribute weighted vector). 对 $p\in D$, p 的属性权向量 $\boldsymbol{AWV}(p)=\langle\omega_1,\cdots,\omega_d\rangle$, 其中 ω_i 为属性 A_i 关于 p 的权重:

$$\omega_i=\begin{cases}\lambda, & \lambda>1,\ A_i\in DIOS(p),\\ 1, & A_i\notin DIOS(p).\end{cases}$$

定义 13. 子空间加权距离(subspace weighted distance). 对 $p,q,r\in D$, q 和 r 关于 p 的子空间加权距离定义为 $SWDist_p(q,r)$:

$$SWDist_p(q,r)=\sqrt{\sum_i\omega_i(\Pi_{A_i}(q)-\Pi_{A_i}(r))^2},$$

其中 ω_i 为属性 A_i 关于 p 的权重.

定义 14. p 的加权 k 距离(weighted k distance). 存在 $o\in D$, 使得至少有 k 个数据点 $o'\in D$ 满足 $SWDist_p(p,o')\leqslant SWDist_p(p,o)$, 并且最多有 $k-1$ 个数据点 $o'\in D$ 满足 $SWDist_p(p,o')<SWDist_p(p,o)$, 则 p 的加权 k 距离 $k\text{-}wdist(p)=SWDist_p(p,o)$.

定义 15. p 的加权 k 邻域(weighted k neighborhood). $WN_k(p)=\{x\in D\mid SWDist_p\leqslant k\text{-}wdist(p)\}$.

定义 16. p 的局部密度(local density). $den(p)=1/k\text{-}wdist(p)$.

定义 17. p 的子空间离群影响因子(subsPace outlier influence factor). p 的子空间离群影响因子 SPOIF 定义如下:

$$SPOIF(p)=\frac{den_{avg}(WN_k(p))}{den(p)},$$

其中 $den_{avg}(WN_k(p))=\frac{\sum_{q\in WN_k(p)}den(q)}{|WN_k(p)|}$.

算法 SPOD 在生成数据点 p 的离群属性子空间 $DIOS(p)$ 后, 采用点 p 的属性权向量 $AWV(p)$ 计算加权距离, 分析点 p 的加权 k 邻域内数据点的分布情况, 计算出点 p 的子空间离群影响因子. 分析可知, $SPOIF$ 值越大, 离群可能性越大. 算法采用对各个数据点的离群属性子空间的属性赋以较大的权值 λ, 其他属性权值固定为 1 的方法, 对数据集进行基于密度的离群点检测, 可以有效解决高维数据集的离群点检测问题. 同时 SPOD 算法也可以用于低维数据集的离群点检测, 当 λ 值取 1 时, SPOD 算法即退化为 LOF 算法.

SPOD 算法流程分为两部分: 1)生成数据点的离群子空间和属性权向量; 2)进行基于加权密度的离群点检测. 在不借助数据集索引结构的情况下, 生成数据点的离群子空间和属性权向量的时间复杂度为 $O(|D|^2)$, 进而对数据点进行离群点检测时间复杂度也为 $O(|D|^2)$. 分析可知, SPOD 算法具有与 LOF 算法同一级别的时间复杂度.

2.3 算法描述

算法 **1**. SPOD.

输入: 维数据集 D, k, λ; 离群因子阈值 t;

输出: 数据集 D 的离群点集合 *Outlier*.

步骤:

```
Initialization;
For each p in D{
  N_k(p)=GenNNK(p, D, k);}
  /*生成 k 邻域点集合 */
  L= N_k(p);
  For(i=1; i≤d; i++){
    LEA_{A_i}(p)=GenLEA(L, A[i]);}
    /*生成局部属性熵 */
For each p in D{
  L= N_k(p);
  For(i=1; i≤d; i++){
    temp= genAVGE(L, A[i])/*计算 L
    中数据点在 A[i] 上的平均熵值 */
    If(LEA_{A_i}(p)≥temp){AWV[p][i]=λ}
    /*生成属性权向量, 初始值均为 1 */
}}
For each p in D{
Genkwdist(p); /*生成加权 k 距离 */
WN_k[p]=genWN(p, D, AWV);
/*计算加权 k 邻域 */
SPOIF[p]=genSPOIF(WN[p], AWV, D);
/*计算 p 的加权子空间离群影响因子 */
If(SPOIF[p]> t){ /*判断离群性 */
  p is marked an outlier;
  insert(p, AVW[p])into Outlier;
  /*将 p 及对应离群子空间加入离群点集
  合 */
}Else{p is marked as non-outlier;}}
```

3 实验结果

这部分对 SPOD 算法的性能进行实验分析. 实验平台配置如下: Intel 1.8GHz/512MB, Windows 2000 (Server 版), 代码用 Visual C++ (6.0)实现, 性能比较需要, 将文献[7]中子空间选取方法也应用于第 2.2 节提出的离群点检测, 称为 LSPOD(local subspace preference outlier detection)算法, 并将其性能与 SPOD 算法进行实验比较.

实验所使用的数据共有 2 种, 第 1 种是网络入侵检测数据集 KDD-CUP1000, 该数据集中的数据对象分为五大类, 包括正常的连接、各种入侵和攻击等. 为了进行实验, 对 KDD-CUP1000 数据适当的修改, 使得攻击(即离群点)占数据集的 5%. 选择了其中的 34 个连续值属性维, 对非数值属性维进行数值化处理. 第 2 种是人工生成数据, 可以通过输入参数来控制产生数据集的结构与大小, 参数包括数据集的大小、维数、各维上的取值范围等. 文中用'*B*'、'*C*'、'*D*'分别表示数据集记录数、所含聚类个数、数据空间维数.

1) 算法的精确度和执行效率

为了测试算法的精确度和执行效率, 对算法 SPOD, LSPOD 和 LOF 的性能进行对比, 采用 KDD-CUP1000 数据集和仿真数据集 B1000C6D20 作为测试集, 均匀加入 3%和聚类具有较大偏差的离群点. 实验中取 $k=6$, $t=1.3$, $\lambda=1.2$, LSPOD 算法的邻域半径 ε 和阈值 δ 采用对数据集的采样数据进行预分析的方法设定. 精度采用以下量度对算法进行评价:

$$Precision = \frac{\text{Number of correct outliers}}{|Outlier|}.$$

实验结果如图 1 和图 2 所示, SPOD 虽然效率没有 LOF 高, 但运行时间仍为同一量级, 算法与 LSPOD 效率相近; 由于算法 SPOD 采用基于熵的离群子空间进行加权离群点检测, 较好地适应高维数据的特点, 在算法精度上远优于 LSPOD 算法和算法 LOF.

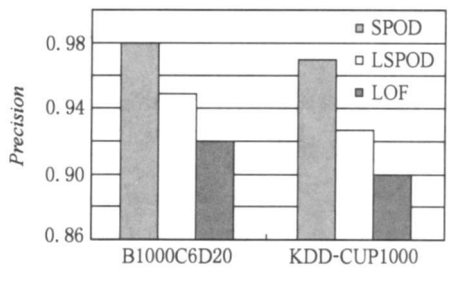

Fig. 1 Precisions of SPOD, LOF.

图 1 不同算法的精度对比

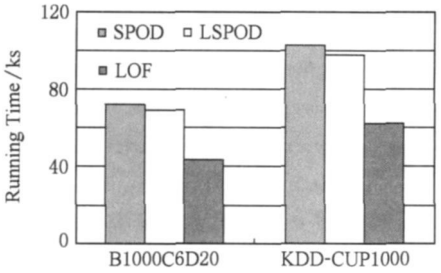

Fig. 2 Running times of SPOD, LOF.

图 2 不同算法的执行时间对比

2) 算法参数的影响

对算法受参数 k 和 λ 的影响情况进行分析,采用仿真数据集 B1000C5D50 测试 k 对算法的影响,取 $t=1.3$, $\lambda=1.2$. 实验结果如图 3 所示,算法 SPOD 受参数 k 的影响较小,而 LOF 算法受 k 的影响较大. 在 k 较小时,随着 k 的增加,LOF 算法的精度增加,k 取 8 时达到一个极大值. 之后随着 k 的继续增加,算法精度下降,这也与高维效应对常规维度离群点检测算法的影响相符. 进一步分析 λ 的取值对 SPOD 算法检测精度的影响,取 $t=1.3$, $k=6$,实验结果如图 4 所示. 由图 4 可知 λ 取 1 时,算法精度较低,这时算法退化为 LOF 算法. 随着 λ 的取值大于 1 后,算法精度激增,之后变化趋缓,这与对离群子空间属性设置较大权向量以适应高维数据集离群点检测的初衷相符.

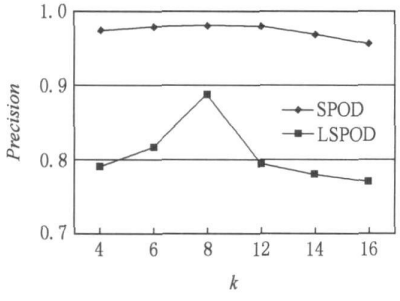

Fig. 3 Scaling of precision with parameter k.

图 3 对参数 k 大小的伸缩性

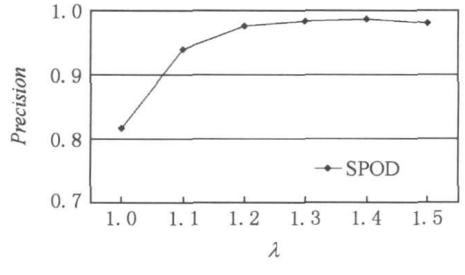

Fig. 4 Scaling of precisions with Parameter λ.

图 4 对参数 λ 的伸缩性

3) 算法对数据集维数的伸缩性

为测试算法对数据集维度的伸缩性,分别生成 10, 20, 30, 40 和 50 维仿真数据集,对 SPOD 和 LOF 算法运行情况进行分析,实验中取 $k=6$, $t=1.3$, $\lambda=1.2$. 由图 5 中可知,随着测试数据集维度的增加,LOF 算法的检测精度与 SPOD 算法检测精度的差距越来越大,SPOD 算法检测精度随着维度的增加变化较小,维持在 0.95 以上.

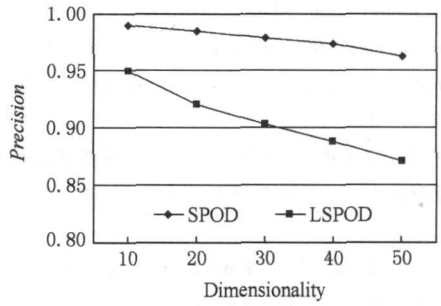

Fig. 5 Scaling of precisions with dataset dimensionality.

图 5 对数据集维数的伸缩性

4 总 结

论文针对高维数据离群点检测问题,提出一种基于熵的加权子空间离群点检测算法,引入离群子空间、属性权向量等概念,对数据点的离群子空间属性赋以较大的权值. 进一步提出子空间加权距离的定义,采用基于密度的离群点检测方法,实现对高维数据集的离群点检测. 理论分析和实验结果表明,算法是有效可行的. 下一步,将结合数据背景知识对子空间离群点的语义解释及算法在复杂数据对象(包括空间对象)上的应用加以研究.

参 考 文 献

[1] Johnson T, Kwok I, Ng R. Fast computation of 2-dimensional depth contours [C] //Gregory Piatetsky-Shapiro ed. Proc of the 4th Int'l Conf on Knowledge Discovery and Data Mining. New York: ACM, 1998: 224-228

[2] Knorr E M, Ng R T. Algorithms for mining distance-based outliers in large datasets [C] //A Gupta, O Shmueli, J Widom, eds. Proc of the 24th Int'l Conf on Very Large Databases. New York: ACM, 1998: 392-403

[3] Breunig M M, Kriegel H, Ng R T, *et al*. LOF: Identifying density-based local outliers [C] //W D Chen, J F Naughton, P A Bernstein, eds. Proc of the 2000 ACM SIGMOD Int'l Conf on Management of Data. New York: ACM, 2000: 93-104

[4] Papadimitirou S, Kitagawa H, Gibbons P B, *et al*. LOCI: Fast outlier detection using the local correlation integral[C] //U Dayal, K Ramamritham, T M Vijayaraman, eds. Proc of the 19th Int'l Conf on Data Engineering. Los Alamitos: IEEE Computer Society, 2003: 315-326

[5] Aggarwal C, Yu P. Outlier detection for high dimensional data[C] //SIGMOD 2001. New York: ACM, 2001

[6] Jin Wen, Tung Anthony K H, Han Jiawei, *et al*. Ranking outliers using symmetric neighborhood relationship [C] // Proc of PAKDD 2006. Berlin: Springer, 2006: 577-593

[7] Li Cunhua, Sun Zhihui. GridOF: An efficient outlier detection algorithm for very large datasets [J]. Journal of Computer Research and Development, 2003, 40(11): 1586-1592 (in Chinese)
(李存华, 孙志挥. GridOF: 面向大规模数据集的高效离群点检测算法[J]. 计算机研究与发展, 2003, 40(11): 1586-1592)

[8] Christian Bohm, Karin Kailing, Hans-Peter Kriegel, *et al*. Density connected clustering with local subspace preferences [C] //The 4th Int'l Conf on Data Mining (ICDM). Los Alamitos: IEEE Computer Society, 2004: 27-34

[9] He Zengyou, Xu Xiaofei, Deng Shengchun. A fast greedy algorithm for outlier mining [C] //Proc of PAKDD 2006. Berlin: Springer, 2006: 567-576

[10] He Zengyou, Xu Xiaofei, Deng Shengchun. An optimization model for outlier detection in categorical data [G] //LNCS 3644. Berlin: Springer, 2005: 400-409

Ni Weiwei, born in 1979. Received his Ph. D. in computer science and engineering from the Southeast University. Associate Prefessor. His main research interests include data mining.
倪巍伟, 1979 年生, 博士, 副教授, 主要研究方向为数据库知识发现.

Chen Geng, born in 1965. Received his Ph. D. in computer science and engineering from the Southeast University, professor. His main research interests include pattern recognition and data mining.
陈 耿, 1965 年生, 博士, 教授, 主要研究方向为模式识别、数据库知识发现.

Lu Jieping, born in 1959. Received his Ph. D. in computer science and engineering from the Southeast University, professor. His main research interests include pattern recognition and data mining.
陆介平, 1959 年生, 博士, 教授, 主要研究方向为模式识别、数据库知识发现.

Wu Yingjie, born in 1979. Ph. D. candidate. His main research interests include data mining.
吴英杰, 1979 年生, 博士研究生, 主要研究方向为数据库知识发现.

Sun Zhihui, born in 1941. Professor and Ph. D. supervisor of the Southeast University. Senior member of China Computer Federation. His main research interests include data mining and complicated information system integration.
孙志挥, 1941 年生, 教授, 博士生导师, 主要研究方向为复杂信息系统集成、数据库知识发现.

Research Background

Outlier mining has become a hot research issue. In this paper we present SPOD, a local entropy based weighted subspace outlier mining algorithm, which generates outlier subspace and weighted attribute vector of each data object by analyzing entropy of each attribute on the neighborhood of this data object, and those outlier attributes are assigned with bigger weight. Furthermore, definitions such as subspace weighted distance are introduced to make a density-based outlier processing upon the data set and get each data point's subspace outlier influence factor. The algorithm can deal with the problem of outlier detecting for datasets with high dimensionality efficiently and effectively. Results of experiments show promising availabilities of our approach. Our work is supported by the Natural Science Foundation of Jiangsu Province (BK2006095) and the Doctor Research Foundation of Education Ministry of China (20040286009).