基于流形学习的离群点检测方法

徐雪松[1]，宋东明[1]，张　　[1]，许满武[2]，刘凤玉[1]
（1.南京理工大学计算机科学与技术学院，南京 210094；2.南京大学计算机科学与技术系，南京 210093）

[摘要] 为了提高高维数据集合离群数据挖掘效率，提出了一种基于流形学习的离群点检测算法。局部线性嵌入(locally linear embedding，LLE)算法是流形学习中有效的非线性降维方法，它的优势在于只定义唯一的参数，即邻域数。根据 LLE 算法的思想寻找样本数据的内在嵌入分布，并通过邻域数选取和降维后数据点之间的距离调整，提高了数据集中离群点发现效率，同时利用离群点权值判别式进行权值数据判定，根据权值的大小标识出数据集中的离群点，仿真实验的结果表明了该方法能够有效地发现高维数据集中的离群点。与此同时，该算法具有参数估计简单、参数影响不大等优点，该算法为离群点检测问题的机器学习提供了一条新的途径。

[关键词] 流形学习；离群点检测；高维数据；维数约减；离群数据
[中图分类号] TP391 [文献标识码] A [文章编号] 1009－1742(2009)02－0082－06

1 前言

离群数据(outlier)是在数据集中与众不同的数据，使人怀疑这些数据并非随机偏差，而是产生于完全不同的机制[1]。目前已经出现了一些高效的离群点检测挖掘算法：基于聚类的、统计的、距离的、深度的以及基于密度的方法等 5 种类型[2~6]。但大多数离群点检测算法对高维数据的异常检测效果都不是很理想。高维空间离群点检测与其他数据集的离群点检测差别甚大的原因主要有两个：a. 对高维数据的估计需要的样本个数与维数构成指数增长的关系，即机器学习中的“维数灾难”(curse of dimensionality)[7]；b. 大量的数据分析问题本质上是非线性的，甚至是高度的非线性。一种常用的做法是在保持数据所含感兴趣信息的前提下，尽可能降低数据的维数，即降维。LLE 是 Roweis 和 Saul 于 2000 年提出的一种解决非线性问题及克服维数灾难问题的方法[8]；由文献[9]和[10]中可知基于距离的离群点最早是由 Knorr 和 Ng 提出的，他们把数据看作高维空间中的点，离群点被定义为数据集中与大多数点之间的距离都大于某个阈值的点。基于距离的离群点定义包含并拓展了基于统计的思想，即使数据集不满足任何特定分布模型，它仍能有效地发现离群点。此算法的一个主要缺陷是要计算所有点之间的距离，每计算一个点的距离就要扫描一次数据集，对于大数据集，其 I/O 次数常常使得算法的计算效率非常低。由此，基于 LLE 算法和基于距离方法的融合，提出了一种思路——基于流形学习的离群点检测方法，其基本思想是：该算法利用 LLE 算法对高维非线性数据进行维数约减，对从高维采样数据中恢复得到低维数据集，通过邻域数的选取，结合降维后数据点之间距离调整，并根据文章提出的离群点权值判别式进行权值数据的判别，同时，在判别基础上，设定分段线性处理，再利用局部邻近点加权，最终确定离群点。实验表明此算法能够快速处理带有离群点的非线性高维数据集，结果与对象空间分布顺序无关，并且效率优于已有的同类基于距离的离群点检测算法。

2 LLE 算法简介

LLE 算法是映射数据 $X=\{x_1,x_2,\cdots,x_n\}$，$x_i\in R^D$ 到数据集 $Y=\{y_1,y_2,\cdots,y_n\}$，$y_i\in R^d(D>d)$。该算法主要包括 3 步：

第一步，对高维空间中的每个样本点 $x_i(i=1,$

[收稿日期] 2007－09－18；修回日期 2008－02－12

[作者简介] 刘凤玉(1943－)，女，江苏江阴市人，南京理工大学教授，主要研究方向为人工智能与信息安全；E-mail：liu.fengyu@263.net

2,…,n),计算它和其他 $n-1$ 个样本点之间的距离,根据距离的大小,选择前 K 个与 x_i($i=1,2,\cdots,n$) 最近的点作为其邻近点,采用欧氏距离来度量两个点之间的距离,即 $d_{ij}=|x_i-x_j|$;

第二步,对每个 x_i($i=1,2,\cdots,n$),找到它的 K 个近邻点之后,计算该点和它的每个近邻点之间的权值 $w_j^{(i)}$,即最小化:

$$\min(w)=\sum_{i=1}^{n}\left|x_i-\sum_{j=1}^{K}w_j^{(i)}x_j\right|^2 \tag{1}$$

其中,$\sum_{j=1}^{k}w_j^{(i)}=1$,如果 x_j($j=1,2,\cdots,n$) 不是 x_i($i=1,2,\cdots,n$) 的近邻,则 $w_j^{(i)}=0$;

第三步,根据高维空间中的样本点 x_i($i=1,2,\cdots,n$) 和它的近邻 x_j($j=1,2,\cdots,K$) 之间的权值 $w_j^{(i)}$ 来计算低维嵌入空间中的值 y_i 和 y_j。由于在低维空间中尽量保持高维空间中的局部线性结构,而权值 $w_j^{(i)}$ 代表着局部信息,所以固定权值 $w_j^{(i)}$,使下面的损失函数最小化:

$$\min\in(\boldsymbol{Y})=\sum_{i=1}^{n}\left|y_i-\sum_{j=1}^{K}w_j^{(i)}y_i\right|^2=tr(Y^{\mathrm{T}}\boldsymbol{M}\boldsymbol{Y}) \tag{2}$$

其中,$\boldsymbol{M}=(I-W)^{\mathrm{T}}(I-W)$。

要求 $\sum_{i=1}^{n}y_i=0$ 且 $\frac{1}{n}\sum_{i=1}^{n}y_iy_i^{\mathrm{T}}=1$,以使 $\min\in(\boldsymbol{Y})$ 对平移、旋转和伸缩变化都具有不变性,使 $\min\in(\boldsymbol{Y})$ 最小化的解为矩阵 $\boldsymbol{M}$ 的最小几个特征值所对应的特征向量构成的矩阵 $\boldsymbol{Y}$。取 $\boldsymbol{M}$ 最小的 $m+1$ 个特征值对应的特征向量,去掉其中最小的特征值对应的特征向量,剩余的 m 个特征向量组成的矩阵就是低维空间中所得特征向量。

3 基于流形学习的离群点检测方法

3.1 邻域数的选取和改进距离定义

最近邻域最佳选取的原因在于大量的最近邻域可以促成流形小规模结构的消除及整个流形的平滑。相反,太少邻域可能误将连续的流形划分成脱节的子流形。可参照文献[11]原则按下式进行选取:

$$k=\mathrm{argmin}(1-\rho^2\boldsymbol{D}_x\boldsymbol{D}_y) \tag{3}$$

$\boldsymbol{D}_x$ 是输入空间 X 的欧氏距离矩阵,$\boldsymbol{D}_y$ 是得到嵌入的低维流形 Y 的欧氏距离矩阵,ρ 是他们之间的相关系数。

同时,对于降维后数据集分布中含有离群点的样本集,近邻点个数 k 的选取对实验结果影响较大。在样本点分布稀疏的区域,k 个近邻点所组成的局部邻域显然要比在样本点分布比较密集的区域大,可以通过以下对数据点之间距离的调整来降低它受样本点分布的影响。

对于 LLE 算法降维后分布中含有离群点的样本集,由于 LLE 算法能够实现高维输入数据点映射到低维坐标系,同时保留邻接点之间的关系,即固有的几何结构得到保留。因此,对降维后所得数据集,可以调整数据点之间的距离,以有利于离群点的发现。

我们知道在样本点分布稀疏的区域,近邻点所组成的局部邻域应该要比在样本点分布比较稠密的区域大,所以需要对原有的欧氏距离公式改进,改进距离如下:

$$d_{ij}(y_i,y_j)=\frac{|y_i-y_j|}{\sqrt{M(i)M(j)}} \tag{4}$$

其中,$M(i)$,$M(j)$ 分别表示 y_i($i=1,2,\cdots,n$),y_j($j=1,2,\cdots,n$) 和其他点之间的平均值,采用改进的距离寻找离群点。

$d_{ij}(y_i,y_j)$ 的分子是普通的欧氏距离,分母是数值,所以容易证明给出的新距离满足距离定义的要求,即

1) $d_{ij}(y_i,y_j)\geqslant 0$,当且仅当 $y_i=y_j$ 时成立,满足距离的非负性;

2) 满足距离对称性要求:

$$d_{ij}(y_i,y_j)=d_{ij}(y_j,y_i);$$

3)满足三角不等式要求,即

$$d_{ij}(y_i,y_j)+d_{it}(y_i,y_t)\geqslant d_{it}(y_j,y_t)。$$

新的距离使处于样本点分布较密集区域的样本点之间的距离增大,而使处于样本点分布较稀疏的区域的样本点之间的距离缩小,这样会使降维后的样本数据集整体分布趋于均匀化,有利于近邻点计算,从而使数据集中的离群点更加突出。同时距离公式可设定所需的距离尺度用于下面的判别定理。

3.2 离群点的权值判别定理

经过 LLE 算法降维,包括离群点的低维数据集是通过权值 W 计算而得,离群点权值的变化情况可由以下定理判别。

令 y_0 代表 y_i,y_i 所代表相应的真实值,$U(y_0)$ 代表 y_0 的邻域,设 $y_1,y_2,\cdots,y_k\in U(y_0)$,则

$$y_0=\sum_{i=1}^{k}W^iy_i,\ \sum_{i=1}^{k}W^i=1$$

令 $y_i' = y_i + d_i(i=0,1,2,\cdots,k)$ 代表相应的离群点,以及相应的

$$y_0' = \sum_{i=1}^{k} W^{i'} y_i', \quad \sum_{i=1}^{k} W^{i'} = 1$$

若再令

$$\boldsymbol{Y}^0 = (y_1, y_2, \cdots, y_k),$$
$$\boldsymbol{Y}^{0'} = (y_1', y_2', \cdots, y_k')$$

以及

$$\boldsymbol{W} = (W^1, W^2, \cdots, W^k)^{\mathrm{T}},$$
$$\boldsymbol{W}' = (W^{1'}, W^{2'}, \cdots, W^{k'})^{\mathrm{T}}$$

于是有

$$y_0 = \boldsymbol{Y}^0 \boldsymbol{W}, \quad y_0' = \boldsymbol{Y}^{0'} \boldsymbol{W}'$$

其中, $\boldsymbol{Y}^0$ 代表 y_0 点的邻域矩阵。

定理　在上述记号下,各离群点之间,不同真实值之间,以及 $\boldsymbol{W}'$ 与离群点之间是相互独立的,各离群点是同均值(0),同方差的,并且记 $\delta\boldsymbol{W} = \boldsymbol{W}' - \boldsymbol{W}$,有如下判别式:

$$E\| \boldsymbol{W}' \|^2 \geqslant E\| \delta\boldsymbol{W} \|^2 \frac{\lambda_{\min} l}{k(k+1)\sigma^2} \tag{5}$$

其中, $\| \cdot \|$ 取为欧几里德范数, $l = \mathrm{ran}k(\boldsymbol{Y}^0)$, $\lambda_{\min}$ 为 $\boldsymbol{Y}^{0\mathrm{T}} \boldsymbol{Y}^0$ 的最小非零特征值, d_i 代表 d 的第 i 个分量:

$$\sigma^2 = \sum_{i=1}^{k} \sigma_i^2, \ \sigma_i^2 = Var(d_i)(i=1,2,\cdots,k)。$$

证明:由 $y_i' = y_i + d_i(i=0,1,\cdots,k)$ 可见

$$y_i' = \boldsymbol{Y}^{0'} \boldsymbol{W}' + \sum_{i=1}^{k} \boldsymbol{W}_i' d_i \Rightarrow y_0$$

$$= \boldsymbol{Y}^0 \boldsymbol{W}' + \sum_{i=1}^{k} \boldsymbol{W}_i' d_i - d_0$$

进一步有 $\boldsymbol{Y}^0(\boldsymbol{W}' - \boldsymbol{W}) = \sum_{i=1}^{k} \boldsymbol{W}_i'(d_0 - d_i)$, $\sum_{i=1}^{k} W_i' = \sum_{i=1}^{k} \boldsymbol{W}_i = 1$

即 $\boldsymbol{Y}^0 \delta\boldsymbol{W} = \sum_{i=1}^{k} \boldsymbol{W}_i'(d_0 - d_i)$ (6)

由于离群点是独立的,则有

$$E(\delta\boldsymbol{W}^{\mathrm{T}} \boldsymbol{Y}^{0\mathrm{T}} \boldsymbol{Y}^0 \delta\boldsymbol{W})$$
$$= E\Big[\sum_{i=1}^{k} \boldsymbol{W}_i'(d_0 - d_i)^{\mathrm{T}} \sum_{i=1}^{k} \boldsymbol{W}_i'(d_0 - d_i) \Big]$$
$$= \sum_{i=1}^{k} E\boldsymbol{W}_i^2 \sigma^2 + \sum_{i=1}^{k} E(\boldsymbol{W}_i' \boldsymbol{W}_j') \sigma^2$$
$$\leqslant (k+1)\sigma^2 E\| \boldsymbol{W}' \|^2 \tag{7}$$

记 $\boldsymbol{Y}^{0\mathrm{T}} \boldsymbol{Y}^0$ 的正交分解为 $\boldsymbol{Y}^{0\mathrm{T}} \boldsymbol{Y}^0 = \boldsymbol{A}^{\mathrm{T}} \wedge \boldsymbol{A}$,其中 $\boldsymbol{A}$ 为正交矩阵, $\wedge = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & \lambda_k \end{bmatrix}$。由于, $\mathrm{ran}k(\boldsymbol{Y}^{0\mathrm{T}} \boldsymbol{Y}^0) = \mathrm{ran}k(\boldsymbol{Y}^0)$,所以

$$E(\delta\boldsymbol{W}^{\mathrm{T}} \boldsymbol{Y}^{0\mathrm{T}} \boldsymbol{Y}^0 \delta\boldsymbol{W}) = E(\delta\boldsymbol{W}^{\mathrm{T}} \boldsymbol{A}^{\mathrm{T}} \wedge \boldsymbol{A}\delta\boldsymbol{W})$$
$$\geqslant \lambda_{\min} \sum_{i=1, \lambda_i=0}^{k} E(\delta\boldsymbol{W}_i^2) \tag{8}$$

通过对 A 进行行初等变换(改变两行的位置),然后计算相应的式(2),将所得结果相加可得

$$E(\delta\boldsymbol{W}^{\mathrm{T}} \boldsymbol{Y}^{0\mathrm{T}} Y^0 \delta\boldsymbol{W}) = E(\delta\boldsymbol{W}^{\mathrm{T}} \boldsymbol{A}^{\mathrm{T}} \wedge \boldsymbol{A}\delta\boldsymbol{W})$$
$$\geqslant \lambda_{\min} E\| \delta\boldsymbol{W} \|^2 \frac{1}{k} \tag{9}$$

其中, $\lambda_{\min} = \min_{1 \leqslant i \leqslant k} \{\lambda_i > 0\}$

结合式(1)可知

$$E\| \delta\boldsymbol{W} \|^2 \leqslant \frac{k(k+1)}{\lambda_{\min} l} \sum_{i=1}^{k} E\boldsymbol{W}_i^2 \delta^2 \tag{10}$$

即, $E\| \boldsymbol{W}' \|^2 \geqslant E\| \delta\boldsymbol{W} \|^2 \frac{\lambda_{\min} l}{k(k+1)\sigma^2}$。

由上述定理可知,在邻域大小 k 已知情况下,离群点权值 $\boldsymbol{W}'$ 主要由 3 个因素决定:a.数据点之间距离 d 的大小;b.邻域的影响,即 $\lambda_{\min}$ 和秩 l 的大小;c.真实值权值的 $\| \boldsymbol{W} \|$ 的大小。对于 a 要素可通过距离公式确定,b 要素可通过调整 $\lambda_{\min}$ 和 l 两个值确定,c 要素直接由 LLE 算法可知。

3.3　算法描述

输入:输入样本数据集 $X = \{x_1, x_2, \cdots, x_n\}$, $x_i \in R^D$,邻域参数 k;

输出:低维数据集中的离群点 y_i';

步骤 1　对高维空间中的每个样本点 $x_i(i=1,2,\cdots,n)$,计算它和其他 $n-1$ 个样本点之间的距离,根据距离的大小,选择前 K 个与 $x_i(i=1,2,\cdots,n)$ 最近的点作为其邻近点,常采用欧氏距离来度量两个点之间的距离;

步骤 2　对每个 $x_i(i=1,2,\cdots,n)$,找到它的 K 个近邻点之后,计算该点和它的每个近邻点之间的权值 $w_j^{(i)}$,即最小化式(1);

步骤 3　对最小化所得的每一点的权值 min $\in (w)$ 组成一个权值矩阵,并对 $\boldsymbol{W}$ 进行约束限制;

步骤 4　根据高维空间中的样本点 $x_i(i=1,2,\cdots,n)$ 和它的近邻 $x_j(j=1,2,\cdots,K)$ 之间的权值 $w_j^{(i)}$ 来计算低维嵌入空间中的值 y_i 和 y_j,即 $y_i - \sum_{j=1}^{K} w_j^{(i)} y_j$;

步骤 5　根据式(3)重新选取邻域数 k,并重复步骤 1 至步骤 4;

步骤 6　根据距离公式改进降维后样本数据集中各点之间的距离,以利于发现样本数据集中的离群点;

步骤 7　经过 LLE 算法降维,包括离群点的低维数据集是通过权值 W 计算而得,离群点权值的变化情况可依据判别定理,由式(4)得出;

步骤 8　对从判别式中得到的离群点权值 W',利用离群点的局部重建权值矩阵及近邻点的线性组合来表示出该离群点。

4　算法的实现及其实验结果的评估

采用文章提出的算法对两个样本集进行实验,其中一个样本集是采用算法生成的曲线形柱面,另一个样本集来自于 UCI 标准数据库(http://www.ics.uci.edu/mlearn/MLRepository.htm1)。

整个实验运行在主频为 P42.4 GHz、内存为 512 MB、80 GB 硬盘、操作系统为 Windows XP sp2 的主机上面。基于此环境,对整个算法进行相关性能测试。

第一个实验数据集为曲线形柱面(如图 1 所示),柱面上半部分由 30 个离群点组成,下半部分由 20×20=400 个采样点组成,数据点维数为三维,内在维数为两维。目的是将三维数据降为两维数据,并发现离群点。

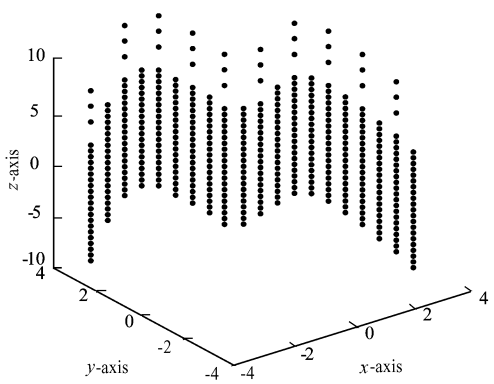

图 1　曲线形柱面数据集

Fig.1　Data set formed by curve shape cylinder

实验中近邻点的个数 k 选择很重要,k 的取值太大,算法不能体现局部特性,而且降维效果显著变差,这是因为测试数据是比较弯曲的流形,容易发生短路现象。反之,算法不能保持样本点在低维空间中的拓扑结构。该处所选取的 k 值尽量使算法中最小化函数达到最小,因此根据式(3)选取 k 为 9。同时由于对降维后的数据点之间距离重新调整,使得降维后的样本数据集整体分布趋于均匀化,有利于近邻点计算,从而使数据集中的离群点更加突出。采用文章提出的算法,将三维的曲线形柱面降维至二维平面,并分离出离群点的效果如图 2 所示。

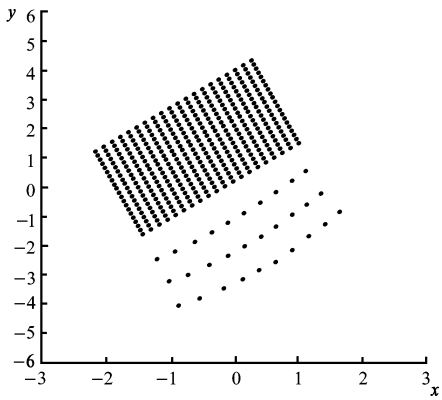

图 2　降维后分离出离群点的数据集

Fig.2　Outliers separated from data set after dimensionality reduction

第二个数据集为 MUSHROOM 数据集,MUSHROOM 数据集为 UCI 标准数据库之一。该数据集中训练样本为 210 个,测试样本为 8 124 个,样本的维数为 22。表 1 给出了分别使用不同的基于距离的离群点检测算法和笔者提出的算法对 MUSHROOM 数据集进行学习和离群点分离的结果,其中提出算法在运行时需要指定不同参数的值,算法中的 k 值根据式(3)选取,其取值范围在 7～12 时效果较好,k 值的选取与测试错误率如图 3 所示。除此之外,还需根据式(4)调整降维后数据集中点之间的距离。笔者提出算法的测试分离离群点的错误率明显低于其他基于距离的离群点检测算法的测试错误率,而且该算法具有良好的特性,即 k 值的估计非常简单,且在一定范围内结果比较稳定。

表 1　不同算法对离群点的检测结果

Table 1　Experimental result of Outliers Detection Based on different algorithms

算法类别	测试错误率/ %
基于索引的算法(基于距离)	22.32(±0.61)
嵌套循环算法(基于距离)	26.73(±0.63)
基于单元算法(基于距离)	18.24(±0.61)
文章提出的算法(k=9)	3.67(±0.66)

实验表明该方法是正确的,为了更好地应用,应注意以下几个方面:

1)如果维数较高,离群点较多,计算时间增加,

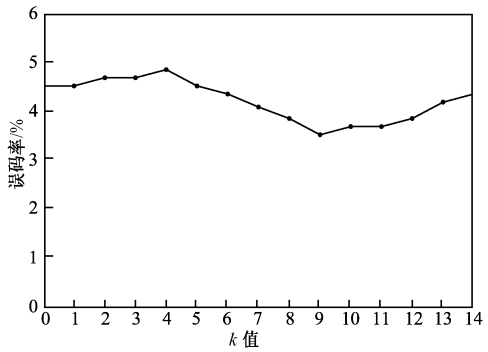

图 3　算法中 k 值对数据集中离群点检测错误率的影响

Fig.3　The effect of k value from algorithm on mistake rate detected in outliers from data set

如离群点个数较少,运算时间明显加快;

2)可以使用增量技术,用户选择某个阈值,计算了一个结果后,不用全部从头开始计算两点距离,可以大大减少运行时间;

3)距离的选取非常重要,实验中采用改进的距离公式,可以根据挖掘数据的目标,将所感兴趣的离群数据值加权。

实验说明了笔者提出的算法既保证了得到的结果相当接近于全局最优解,又保证了能非常快速地得到结果。即使对于样本的个数为8 124、维数为 22 这样的高维大样本,也能快速得到比较满意的结果。

5　结语

在分析了传统的离群数据挖掘算法优点和缺点的基础上,针对非线性高维数据集中的离群点检测问题,提出一种基于流形学习的离群点检测算法。该算法利用 LLE 算法的思想寻找样本数据的内在嵌入分布,并通过邻域数选取和降维后数据点之间的距离调整,提高了数据集中离群点发现效率,根据离群点权值判别式进行权值数据判定,依据权值的大小标识出数据集中的离群点。特别是对于一些线性不可分的数据集,在运用传统离群点检测算法失败的情况下,笔者提出的算法仍然能取得良好的离群点检测效果。

参考文献

[1] 夏火松. 数据仓库与数据挖掘技术[M]. 北京:科学出版社 2004

[2] Barnett V, Lewis T. Outliers in Statistical Data[M]. New York : John Wiley and Sons, Inc, 1994

[3] Preparata F, Shamos M I. Computational Geometry : an Introduction [M]. New York : Springer-Verlag, 1988

[4] Knorr E M, Ng R T. Algorithms for mining distance-based outliers in large datasets[A]. Proceedings of the 24th International Conference on Very Large Data Bases[C]. New York : Morgan Kaufmann, 1998. 392-403

[5] Breunig Markus M B, Kriegel Hans Peter, Ng Raymond T, et al. LOF: identifying density-based local outliers [A]. Chen W, Naughton J F, Bernstein P A, eds. Proceedings of the ACM SIGMOD International Conference on Management of Data [C]. Dallas, Texas : ACM Press, 2000:93-104

[6] Papadimitriou Spiros, Kitagawa Hiroyuki, Gibbons Phillip B. LOCI: fast outlier detection using the local correlation integral [A]. Proceedings of the 19th International Conference on Data Engineering [C].2003. 315-326

[7] Agrawal R, Gehrke J, Gunopulos D, et al. Automatic subspace clustering of high dimensional data for data mining applications[A]. Haas L M, Tiwary A. Proc of the ACM SIGMOD International Conference on Management of Data[C].Seattle:ACM Press,1998:94-105

[8] Roweis S T, Saul L K. Nonlmensionality reduction by locally linear embedding [J]. Science, 2000,290 (22):2323-2325

[9] Knorr E M,Ng R T.Finding intensional knowledge of distance-based outliers[A].Scotland:Pruc of the 25th VLDB Conference Edinburgh [C].1999.211-222

[10] Knorr E M,Ng R T.Algorithms for mining distance-based oufliers in large datasets[A]. New York:Proc of Int Conf Very Large Data-bases (VLDB'98)[C]. 1998.392-403

[11] Olga Kouropteva, Oleg Okun, Matti Pietikäinen. Selection of the optimal parameter value for the locally linear embedding algorithm[A]. FSDK''02, Proc of the 1 st Int Conf on Fuzzy Systems and Knowledge Discovery[C]. 2002. 359-363

The research of detection of outliers based on manifold learning

Xu Xuesong[1], Song Dongming[1], Zhang Xu[1], Xu Manwu[2], Liu Fengyu[1]

(*1. Department of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, China;*

2. Department of Computer Science and Technology, Nanjing University, Nanjing 210093, China)

[**Abstract**] The data dimensionality reduction is the main method that can enhance the outliers mining efficiency based on higher-dimension data set. The research of detection of outliers based on manifold learning is proposed after analyzing the advantages and disadvantages of the classical outlier mining algorithm in the paper. Local Linear Embedding algorithm(LLE) is an effective technique for nonlinear dimensionality reduction in manifold learning. Compared with other dimensionality reduction algorithms, the advantage of the local Linear Embedding algorithm is that it only defines unique parameter, i.e. number of nearest neighbours. With the idea of Local Linear Embedding, the algorithm can select optimal parameter and regulate the distance among data set after data dimensionality reduction, so as to improve efficiency of detection of outliers. The algorithm determines weighted values by discretion formula of weighted outliers. Through these weighted values, the experts can identify the outliers easily. Simulation results illustrate that this algorithm is very efficient. Moreover, our method has the advantage of simple parameter estimation and low parameter sensitivity. Our method gives a new way for the solution of detection of outliers.

[**Key words**] manifold learning; detection of outliers; high dimensional data; dimensionality reduction; outliers