基于局部线性加权的离群点检测方法

徐雪松 宋东明 张 谞 张 宏 刘凤玉

(南京理工大学计算机科学与技术学院 南京 210094)

摘 要 为了提高高维数据集合离群数据挖掘效率,在分析了传统的离群数据挖掘算法优点和缺点的基础上,提出了一种基于局部线性加权的离群点检测算法。该算法利用 LLE 算法的思想寻找样本数据的内在嵌入分布,并通过距离公式和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 局部线性嵌入,高维数据,非线性降维,离群数据

Research of Detection of Outliers Based on Locally Linear Weighted Value

XU Xue-song SONG Dong-ming ZHANG Xu ZHANG Hong LIU Feng-yu

(Department of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract The data dimension 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 locally linear weighted value is proposed after analyzing the advantages and disadvantages of the classical outlier mining algorithm in the paper. With the idea of Local Linear Embedding, the algorithm tries to find distribution of internal embedding of the samples, and determines value-weighted by combining distance formula with 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.

Keywords Locally linear embedding, High dimensional data, Nonlinear dimensionality reduction, Outliers

1 引言

离群数据(outlier)就是明显偏离其它数据、不满足数据的一般模式或行为、与存在的其它数据不一致的数据[1]。离群点检测的目的在于找出隐含在海量数据中相对稀疏而孤立的异常数据模式,这是离群点检测与关联规则等传统的面向数据主体的数据挖掘的主要区别。早期,对数据集进行预处理时,通常把离群点当作噪声,或修正离群点值,以减少其对正常数据的影响。尽管离群点检测是为了发现数据集中极少数的一些数据,然而离群数据挖掘常常比其他类型的挖掘来得更有价值,因为 1 万个正常的记录很可能只覆盖一条规则,而 10 个离群很可能就意味着 10 条不同的规则。实际生活中,离群检测有着很广泛的应用,如网络入侵检测、信用卡恶意透支、贷款证明的审核等。目前已经出现了一些高效的离群点检测挖掘算法,它们可分为基于聚类的、统计的、距离的、深度的以及基于密度的方法等五种类型[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:①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,这在机器学习中称作著名的"维数灾难"(Curse of Dimensionality)[7]。②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,对此我们不能利用已有的快速成熟的线性模型进行研究。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(Locally Linear Embedding)是 Roweis 和 Saul 于 2000 年提出的一种非线性降维方法[8],它是解决非线性问题及克服维数灾难问题的关键。该算法能够实现高维输入数据点映射到一个低维坐标系,同时保留了邻接点之间的关系,这样固有的几何结构就能够得到保留。此算法不仅能够有效地发现数据的非线性结构,同时具有平移、旋转等不变特性。由文献[9,10]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,将 LLE 算法和基于距离的方法融合,本文提出了一种思路——基于局部线性加权的离群点检测方法,其基本思想是:为了适合挖掘高维复杂数据集中的离群点,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结合本文提出的距离公式,并根据本文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。实验表明了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。

本文第 2 节简单介绍 LLE 算法;第 3 节详细介绍提出的离群点检测方法;第 4 节是实验及其结果分析;最后是结束语。

徐雪松 博士生,主要研究方向为离群数据发现技术、信息安全;宋东明 博士生,主要研究方向为系统仿真;张 谞 博士生,主要研究方向为分布式虚拟现实;张 宏 教授,博士生导师,主要研究方向为信息安全;刘凤玉 教授,博士生导师,主要研究方向为人工智能与信息安全。

2 LLE算法

LLE是一种依赖于局部线性的算法，它认为在局部意义下，数据结构是线性的，或者说局部意义下的点在一个超平面上。主要利用局部的线性来逼近全局的非线性，保持局部的几何结构不变，通过相互重叠的局部邻域来提供整体的信息，从而保持整体的几何性质。

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 $Y=\{y_1,y_2,\cdots,y_n\}$，$y_i\in R^d(D>d)$。该算法主要包括三步：

第一步，对高维空间中的每个样本点 $x_i(i=1,2,\cdots,n)$，计算它和其它 $n-1$ 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，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：

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

其中，$\sum_{j=1}^{n}w_j^{(i)}=1$，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，因此固定权值 $w_j^{(i)}$，使下面的损失函数最小化：

$$\min\in(Y)=\sum_{i=1}^{n}|y_i-\sum_{j=1}^{K}w_j^{(i)}y_j|^2=tr(Y^TMY) \tag{2}$$

其中，$M=(I-W)^T(I-W)$。

要求 $\sum_{i=1}^{n}y_i=0$ 且 $\frac{1}{n}\sum_{i=1}^{n}y_iy_i^T=1$，以使 $\min\in(Y)$ 对平移、旋转和伸缩变化都具有不变性，使 $\min\in(Y)$ 最小化的解为矩阵 M 的最小几个特征值所对应的特征向量构成的矩阵 Y。取 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 基于局部线性加权的离群点检测方法

3.1 改进距离定义

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：

$$d_{ij}(y_i,y_j)=\frac{|y_i-y_j|}{\sqrt{M(i)M(j))^2}} \tag{3}$$

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 $y_i=y_j$ 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，有利于离群点的权值计算。同时，距离公式可设定所需的距离阈值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 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

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

若再令

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

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

$y_0=Y^0W$，$y_0'=Y^{0'}W'$

其中，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 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 $\delta W=W'-W$，有如下判别式：

$$E\|W'\|^2\geqslant E\|\delta W\|^2\frac{\lambda_{\min}l}{k(k+1)\sigma^2} \tag{4}$$

其中 $\|\cdot\|$ 取为欧几里德范数，$l=rank(Y^0)$，

$\lambda_{\min}$ 为 $Y^{0T}Y^0$ 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

$y_0'=Y^{0'}W'+\sum_{i=1}^{k}W_i'd_i\Rightarrow y_0=Y^0W'+\sum_{i=1}^{k}W_i'd_i-d_0$ 进一步有

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

即 $Y^0\delta W=\sum_{i=1}^{k}W_i'(d_0-d_i)$ (5)

由于离群点是独立的，则有

$$E(\delta W^TY^{0T}Y^0\delta W)=E[\sum_{i=1}^{k}W_i'(d_0-d_i)^T\sum_{i=1}^{k}W_i'(d_0-d_i)]=\sum_{i=1}^{k}EW_i^2\sigma^2+\sum_{i=1}^{k}E(W_i'W_j')\sigma^2\leqslant(k+1)\sigma^2\sum_{i=1}^{k}EW_i^2=(k+1)\sigma^2E\|W'\|^2 \tag{6}$$

记 $Y^{0T}Y^0$ 的正交分解为 $Y^{0T}Y^0=A^T\wedge A$，其中 A 为正交阵，$\wedge=\begin{bmatrix}\lambda_1 & \cdots & 0\\ \vdots & & \vdots\\ 0 & \cdots & \lambda_k\end{bmatrix}$。

由于 $rank(Y^{0T}Y^0)=rank(Y^0)$，

所以

$$E(\delta W^TY^{0T}Y^0\delta W)=E(\delta W^TA^T\wedge A\delta W)\geqslant\lambda_{\min}\sum_{i=1,\lambda_i=0}^{k}E(\delta W_i^2) \tag{7}$$

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(2)式，将所得结果相加，可得

$$E(\delta W^T Y^{0T} Y^0 \delta W) = E(\delta W^T A^T \wedge A \delta W)$$

$$\geqslant \lambda_{\min} E \| \delta W \|^2 \frac{1}{k} \tag{8}$$

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

结合(1)式，可知

$$E \| \delta W \|^2 \leqslant \frac{k(k+1)}{\lambda_{\min} l} \sum_{i=1}^{k} E W_i^2 \delta^2 \tag{9}$$

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

由上述定理可知，在邻域大小 k 已知情况下，离群点权值 W' 主要由 3 个因素决定：(1)数据点之间距离 d 的大小；(2)邻域的影响，即 $\lambda_{\min}$ 和秩 l 的大小；(3)真实值权值的 $\| W \|$ 的大小。对于(1)可通过距离公式确定，(2)可调整 $\lambda_{\min}$ 和 l 两个值确定，(3)直接由 LLE 算法可知。

3.3 算法描述

基于上面的分析，我们建立基于局部线性加权的离群点检测方法，其工作过程说明如下。

输入：输入样本数据集 $X = \{x_1, x_2, \cdots, x_n\}$，$x_i \in RD$，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 $x_i (i = 1, 2, \cdots, n)$，计算它和其它 $n-1$ 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(1)式；

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 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. 根据距离公式改进降维后样本数据集中各点之间的距离，以使样本数据集中的离群点更加突出；

步骤 6. 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判别式(4)得出；

步骤 7. 由于使用 LLE 算法进行降维，LLE 算法是从保持局部线性假设出发，因为在降维后的数据集中，对从判别式中得到的离群点权值 W'，利用一点的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

采用本文提出的算法对两个样本集进行实验，其中一个样本集是人工构造，另一个样本集来自于 UCI 标准数据库(http://www.ics.uci.edu/mlearn/MLRepository.html)。

整个实验运行在主频为 P42.4GHz、内存为 512MB、80GB 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 个点组成，数据点维数为三维，内在维数为两维。

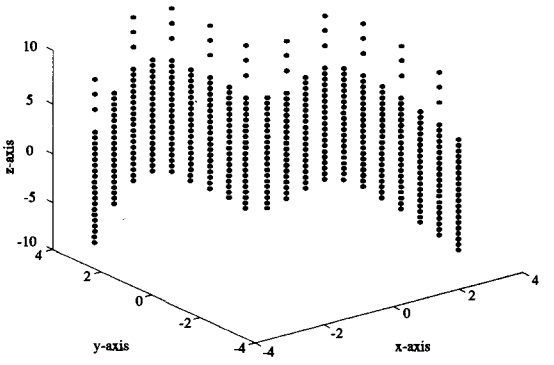

图 1 曲线形柱面数据集

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，使得该算法趋向于主成分分析算法。反之，算法不能保持样本点在低维空间中的拓扑结构。本文所选取的 k 值尽量使算法中最小化函数达到最小，因此选取 k 为 10(通过实验)。采用本文提出的算法，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。

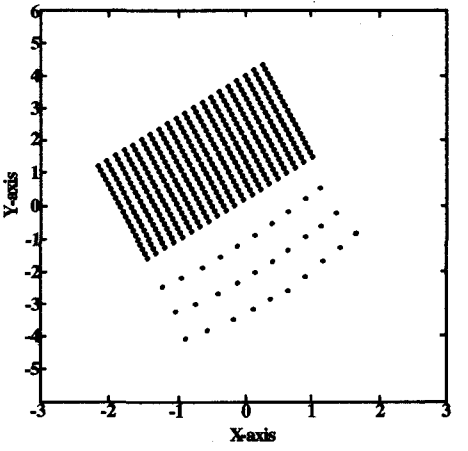

图 2 降维后分离出离群点的数据集

第二个数据集为 IMAGE 数据集，IMAGE 数据集为 UCI 标准数据库之一。该数据集中，训练样本为 1300 个，测试样本为 1010 个，样本的维数为 18。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本文提出的算法对 IMAGE 数据集进行学习和离群点分离的结果，其中本文提出算法中的 k 值分别取 10 和 15。表中结果显示，我们提出的算法的测试分离离群点的错误率明显低于其它基于距离的离群点检测算法的测试错误率。而且，我们的算法具有一个良好的特性，那就是 k 值的估计非常简单，且在一定范围内结果比较稳定。

表 1 不同算法对离群点检测结果

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)
本文提出的算法(k=10)	3.42(±0.66)
本文提出的算法(k=15)	3.67(±0.72)

实验表明我们的方法是正确的，为了更好地应用我们的方法，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，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，实验中我们采用改进的距离公式。可以根据我们挖掘数据的目标，将所感兴趣的离群数据值加权。

实验说明了用本文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1010、维数为18这样的高维大样本，用本文的算法也能快速得到比较满意的结果。

结束语 自离群数据的概念提出以来，在数据挖掘领域一直聚集了大量的研究人员从事离群数据挖掘技术的理论和实际应用的研究。本文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侦测问题，提出一种基于局部线性加权的离群点检测算法。该算法利用LLE算法的思想寻找样本数据的内在嵌入分布，并通过距离公式和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。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，本文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. 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] Franco P, Ian S M. Computational Geometry : an Introduction [M] . New York : Springer - Verlag , 1988

[4] Edwin K M, Raymond N T. Algorithms for Mining Distance - Based Outliers in Large Datasets[R]//Proceedings of the 24th International Conference on Very Large Data Bases. New York:Morgan Kaufmann , 1998:392 -403

[5] Markus B M, Peter K H, Raymond N T , et al . LOF : Identifying Density - based Local Outliers[R] //Chen W, Naughton J F , Bernstein P A, eds. Proceedings of the ACM SIGMOD International Conference on Management of Data. Dallas , Texas: ACM Press , 2000: 93 - 104

[6] Spiros P, Hiroyuki K, Phillip G B. LOCI : Fast Outlier Detection Using the Local Correlation Integral [R]//Proceedings of the 19th International Conference on Data Engineering. 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, eds. Proc. ofthe 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)

[9] Knorr E, Ng R. Finding Intensional Knowledge of Distance—based Outliers[C]//Scotland: Proc of the 25th VLDB Conference. Edinburgh, 1999: 211-222

[10] E M, Ng R T. Algorithms for Mining Distance—based Oufliers in Large Datasets//Proc. of Int. Conf. Very Large Data—bases (VLDB'98). C J, New York, 1998: 392-403

(上接第153页)

等挖掘方法产生一些归纳规则、一些关系规则和一些模式。这些归纳规则被存于数据库中，根据数据库里先前产生结果用精确表示查询条件而设计的查询语言对数据库进行查询。这个框架适合基于图的数据挖掘。用基于图的数据挖掘方法预先产生子图中的子图和关系，并把它存于归约数据库中，用致力于数据库的查询语言去查询子图和它们之间的关系。这种方法的优点是加快图挖掘的操作，这是因为子图和它们的关系的基本模式已经提前产生了。潜在的缺点是计算量大，提前产生和保存归纳模式的内存大。在一些工作中，这种方法常和下面提到的完全层次搜索方法联合使用。

MolFea系统使用的是归约数据库方法，MolFea系统使用的是level-wise版本空间算法[10]。该方法虽然对原图执行了完全的路径搜索，但是它所挖掘的子结构被限制在图中的连续子结构的范围里。

3.3 归纳逻辑编程(ILP)方法

归纳逻辑程序设计(Inductive Logic Programming, ILP)是计算机科学一个重要的分支学科，是归纳学习(Inductive Learning)与逻辑程序设计(Logic Programming)融合的产物。归纳方法是一种重要的机器学习方法，能够在特殊样例上产生泛化模型；逻辑程序设计使用一阶谓词逻辑表示实体间关系并在此基础上作演绎推理。

3.4 数学图论方法

数学图论方法[11]能够从图中抽取出频繁出现的公共子结构，并且能够执行完全搜索。因为数学图论方法要处理图的同构问题，解决图的同构问题的算法是多种多样的。而精确的图的匹配(同构)的难度相当于NP完全问题(NP完全问题是这样的问题：用确定性的算法在多项式时间内无法解决的问题)的难度，这时就要加上一定的限制条件以减少搜索空间，以便在一定程度上回避NP完全问题的难度，从而达到降低时间复杂度的目的。

4 目前图的数据挖掘技术中存在的问题

基于图的频繁子图挖掘的各种算法及其应用，目前在这方面的研究已经取得了一些成绩，并被应用到一些实际的系统中。

虽然目前存在的算法很多，但执行效率很不高，在处理大型数据库时仍然需要大量的时间和空间；求不精确的频繁子图问题在现实生活中广泛存在，但是目前的算法还停留在最初提出的SUBDUE上，有待进一步提高；在large graph型数据库中进行挖掘的算法也比较少，目前主要都是基于transaction型数据库的，因此对large graph型数据库的频繁子图挖掘算法的研究也是一个重要的研究方面；现实生活中有各种各样的图形，如有向图、无向图、加权图、无连通图等，但目前的算法大部分都是针对连通图的挖掘，对加权图等的挖掘算法很少，因此对加权图等的挖掘算法的研究也是一个重要的研究方向等。

参考文献

[1] Mckay B. Nauty users guide (version 1. 5). Technical Report, TR-CS-90-02. Department of Computer Science, Australian National University, 1990

[2] Inokuchi A, Washio T, Motoda H. Complete mining of frequent patterns from graphs: Mining graph data. Machine Learning, 2003,50:321-354

[3] De Raedt L, Kramer S. The levelwise version space algorithm and its application to molecular fragment finding//IJCAI'01: Seventeenth International Joint Conference on Artificial Intelligence, vol2, 2001:853-859

[4] Agrawal R, Srikant R. Fast algorithms for mining association rules//VLDB'94: Twentyth Very Large Dada Base Conference. 1994:487-499

[5] Yoshida K, Motoda H, Indurkhya N. Graph-based induction as a unified learning framework. J of Applied Intel,1994,4:297-328

[6] Cook J, Holder L. Substructure discovery using minimum description length and background knowledge. J Artificial Intel Research, 1994,1:231-255

[7] Cook D J, Holder B. Graph-based Data Mining. IEEE,2000,15 (2):32-41

[8] Cook J, Holder L B, Djoko S. Scalable discovery of informative structural concepts using domain know] edge. IEEE Expert, 1996,11(15)

[9] Agrawal R, Srikant R. Fast algorithms for mining association rules//VLDB'94: Twentyth Very Large Dada Base Conference. 1994:487-499

[10] De Raedt L, Kramer S. The levelwise version space algorithm and its application to molecular fragment finding//IJCAI' OI: Seventeenth International Joint Conference on Artificial Intelligence. vol2. 2001:853-859

[11] Inokuchi A, Washio T, Motoda H. Complete mining of frequent patterns from graphs : Mining graph data. Machine Learning, 2003,50:321-354