文章编号:1001 - 9081(2006)10 - 2401 - 04

维规约技术综述

许明旺[1],施润身[2]
(1. 同济大学 电子与信息工程学院,上海 200433;2. 同济大学 研究生院,上海 200092)
(iqdoctor3@ hotmail. com)

摘 要:从属性选择和维变换两个方面对维规约技术进行了概括。首先对属性选择的基本思想和常用算法进行简要介绍;然后对维变换技术中的几种应用最广泛的方法进行了详细研究,主要包括主成分分析及其相关算法、独立成分分析、因子分析、投影寻踪等方法,简要给出了这些方法间的联系和区别,最后指出了维规约技术的现状和发展方向。

关键词:维规约;主成分分析;独立成分分析;因子分析;投影寻踪
中图分类号: TP311.13 **文献标识码**:A

Survey on dimension reduction techniques

XU Ming-wang[1], SHI Run-shen[2]
(1. *College of Electronics and Information Engineering, Tongji University, Shanghai* 200433, *China;*
2. *Graduate School, Tongji University, Shanghai* 200092, *China*)

Abstract: Dimension reduction techniques were discussed from the two aspects: feature selection and dimension transformation. Firstly, the basic theory and famous algorithms of feature selection were roughly introduced. Then, several most popular dimension transformation techniques were analyzed in detail including Principal Components Analysis and its related methods, Independent Components Analysis, factor analysis, projection pursuit, etc. Meanwhile, connection and distinction between them were provided. Finally, the present situation and future development of dimension reduction techniques were pointed out.

Key words: dimension reduction; Principal Components Analysis(PCA); Independent Components Analysis(ICA); factor analysis; projection pursuit

0 引言

数据挖掘又称为数据库中的知识发现(Knowledge Discovery in Database),是指从大量的数据中提取隐含在其中的、人们事先不知道的、但又潜在有用的信息和知识的过程,它是目前国际上数据库和信息决策领域最前沿的研究方向之一。数据挖掘过程一般包括数据采集、数据预处理,数据开采及知识评价和呈现等[1]。

现实世界的数据一般是脏的、不完整的和不一致的。数据预处理(Data Preprocessing)技术可以改进数据的质量,从而有助于提高其后的挖掘过程的精度和性能。作为数据挖掘的关键,根据统计,在一个完整的数据挖掘过程中,数据预处理要花费60%左右的时间,而后的挖掘工作仅占总工作量的10%左右。数据预处理包括数据清理、数据集成、数据变换和数据规约。数据清理可以用于填充空缺的值,平滑数据,找出孤立点以及纠正数据的不一致性等;数据集成将来自不同数据源的数据整合成一致的数据存储;数据变换将数据变换成适于挖掘的形式;数据规约技术,如数据立方体聚集、维规约、数据压缩、数值规约和离散化都可以用来得到数据的鼓乐表示,使信息内容的损失最小[2]。

维规约技术,从数学角度讲就是,对于给定 p 维的数据向量 $\boldsymbol{x} = \{x_1, x_2, x_3, \cdots, x_p\}$,在某种条件下,寻找一个能反映原始数据信息的较低维的表示,即 $\boldsymbol{s} = \{s_1, s_2, s_3, \cdots, s_k\}$,使得 $k \leqslant p$(理想情况 $k \ll p$),$\boldsymbol{s}$ 的向量有时又被称为潜隐向量[3]。维规约技术,其形式分为两种:属性选择(Feature Selection)和维变换(Dimension Transformation,又称属性变换 Feature Transformation)[4]。两者中,前者通过选择属性子集代表原属性集来达到维规约目的,后者则通过线性或非线性方法将高维属性空间变换到低维属性空间从而达到维规约目的。

1 属性选择

1.1 问题描述

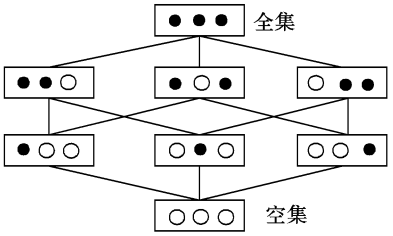

图1 含三个属性的数据集

如图1,对于一个含三个属性的数据集,它共有8个子集,其中黑点代表选中某属性,圆圈代表没有选中该属性。属性子集选择的目标是找出最小属性集,使得数据类的概率分布尽可能地接近使用所有属性的原分布。在压缩的属性集上

收稿日期:2006 - 04 - 05;**修订日期**:2006 - 06 - 08

作者简介:许明旺(1982 -),男,安徽安庆人,硕士研究生,主要研究方向:数据库技术与应用; 施润身(1946 -),男,浙江杭州人,教授,主要研究方向:数据库应用技术、信息系统.

挖掘其他优点,它减少了出现在发现模式上的属性数目,使得模式更易于理解。

1.2 常用搜索方法和算法

如何找出原属性的一个“好的”子集,d 个属性有 2^d 个可能的子集。一般搜索符合要求属性的方法有穷举搜索、启发式搜索等。

穷举搜索指列出并评估所有的属性子集直到选出最好的子集。这种法只有在属性集较小时适用,因为该搜索方法的搜索空间随初始属性集的大小指数增长。

压缩搜索空间的启发式算法通常是当 d 的数目较多的时候使用。一般而言,这些算法是贪心算法,在搜索属性空间时,总是做看上去是最佳的选择。它们的策略是做局部最优选择,期望由此导致全局最优解。在实践中,这种贪心方法是有效的,并可以接近最优解。“最好的”(或“最差的”)属性使用统计意义的测试来选择。这种测试假定属性是相互独立的,也可以使用一些其他属性估计度量,如使用信息增益度量建立分类判定树。

常用的启发式搜索算法有:逐步向前选择(SFS)、逐步向后删除(SBS)、向前选择和向后删除的结合、判定树归纳等,具体参见文献[5]。

1.3 包装模式和过滤模式

在过滤模式中,搜索算法反复产生属性子集,这些子集通过某个评估模式评估。反复进行直至达到某个停止阈值,这样最终的属性子集即为输出。最终的子集通过分类来评估。

在包装模式中,搜索算法反复产生属性子集,这些子集通过分类本身来评估而不是通过一个评估模型来评估。这个过程也是反复进行直到该模式达到足够好的性能。

所以,包装模式时间复杂度较高,但具有更高的精确性。而过滤模式选出的属性子集的性能很大程度上取决于评估模型中评估准则的选取。

2 维变换技术

2.1 主成分分析

2.1.1 原理

主成分分析(Principle Components Analysis,PCA),在不同的领域又被称 KARHUNEN-LOEVE(K-L)方法、HOTELLING 变换以及 EOF(Empirical Orthogonal Function)方法等[4]。它是一种把原来多个指标化为少数几个互不相关的综合指标的多元统计方法,可以达到数据化简、揭示变量间的关系和进行统计解释的目的,为进一步分析数据的性质和统计特征提供重要信息。

假定待压缩的数据有 p 维,有 n 组观测数据,表示为 $x=(x_1,x_2,\cdots,x_p)$,要搜索其主成分 $Y=(y_1,y_2,\cdots,y_m)$,使得 $m\leqslant p$,这样将原来的数据投影到一个较小的空间,达到维规约的目的。其模型为:

$$y_1 = l_{11}x_1 + l_{12}x_2 + \cdots + l_{1p}x_p$$
$$y_2 = l_{21}x_1 + l_{22}x_2 + \cdots + l_{2p}x_p$$
$$\cdots$$
$$y_m = l_{m1}x_1 + l_{m2}x_2 + \cdots + l_{mp}x_p$$

该方程满足 $l_{i1}^2 + l_{i2}^2, + \cdots + l_{ip}^2 = 1(i=1,2,\cdots,m)$,且 $y_1, y_2,\cdots,y_m$ 互不相关。其中 y_1 是 $x_1,x_2,\cdots,x_p$ 线性组合中方差最大的,故而含有最大信息量,称为第一成分,依次类推。

2.1.2 主成分的计算

通常,对数据集 x 主成分的求解可转化为求 x 的协方差矩阵的特征根和标准正交向量的问题,以下是其主要过程:

1)对原始数据样本集 $(x_{ij})_{n\times p}$ 进行标准化处理,即 $x'_j=(x_j-\mu_j)/\sigma_j, j\in 1,\cdots,p$。式中 μ_j,σ_j 分别为特征变量 x_j 的均值和标准差。

2)建立标准化数据的协方差矩阵 $\boldsymbol{V}$,求解 $\boldsymbol{V}$ 的 $k(k\leqslant p)$ 个不为 0 的特征值和与特征值对应的标准正交特征向量。

3)根据贡献率 $\lambda_i/\sum_{j=1}^{k}\lambda_j$,确定满足累计贡献率 $\sum_{i=1}^{m}\lambda_i/\sum_{j=1}^{k}\lambda_j$($m\leqslant k$)的 m 个主成分。

4)建立主成分方程,计算主成分值,形成新的训练样本集和测试样本集,其中主成分方程为:$y_i=\sum_{j=1}^{p}a_{ij}x'_j, i=1,2,\cdots,m$,其中 $a_i=(a_{i1},a_{i2},\cdots,a_{ip})$ 是 $\boldsymbol{V}$ 的第 i 个特征值对应的标准正交向量[6]。

2.1.3 主成分分析的发展

1)均值法。在主成分分析中原始数据标准化是为了避免各指标变量的量纲和数量级对协方差矩阵的影响,但同时它也消除了各指标在变异程度上的差异信息。均值化方法是一种较好的改进方法。均值化就是用各指标的均值除它们相应的原始数据。均值化处理不改变指标间的相关系数,相关矩阵的全部信息都在相应的协方差矩阵中得到反映。可见均值化处理后的协方差矩阵不仅消除了指标量纲与数量级的影响,还能包含原始数据的全部信息,因此在进行主成分分析前,可以用均值化方法进行无量纲化处理[6]。

2)非线性主成分分析。主成分分析法是一种线性降维法,表现为各主成分是原始变量的线性组合。因此,当原始数据不具备线性的基本特点时,若简单地进行线性处理,必然会导致结果的偏差,因此有必要对传统主成分分析中的“线性化”进行改进。一般可直接对它们进行函数处理:描绘原始数据列 x_{ij} 的散点图,若散点图呈现出某种曲线特征,如呈现出对数曲线特征时,则可令 $y_{ij}=\ln x_{ij}$,再经过中心化变换利用主成分分析法,可提高降维效果。对具有“非线性”特征的原始数据进行函数处理后,作为主成分分析的指标就是原始指标的函数,求出的主成分就出现非线性的形式。对原始数据进行变换不仅会明显提高降维效果,用更少的主成分更多反映原始指标信息,而且会使评价模型更具科学性[7]。

3)其他相关算法方法。新近的 PCA 相关算法还有:概率主成分分析方法(PPCA)、核主成分分析方法(Kernel PCA)等。概率 PCA 是传统 PCA 的延伸,它由 Michael E Tipping, Christopher M. Bishop 等人首先提出,其目的是为 PCA 定义一个恰当的概率模型。在传统 PCA 中,子空间外的信息只是简单的丢弃,然而在 PPCA 中,这些信息将作为高斯噪声进行估计。对于概率 PCA 模型,可以通过最大可能函数或 EM 算法来估计模型参数而得到最佳概率模型。PPCA 定义了一个恰当的概率模型,这个模型能很容易地延伸为混合模型,同时该模型的参数能用 EM 算法训练获得[8,9]。核主成分分析方法不是直接计算特征向量,而是将其转化为求核矩阵的特征向量和特征值,这避免了在特征空间求特征向量,而数据在特征向量上的投影转换为求核函数的线性组合,有效地简化了计算[10]。

2.2 因子分析

因子分析(Factor Analysis)认为,在所收集到数据的众多变量中,必定存在某些是高度相关的,把这些高度相关的变量

组成各组。这样同一组内变量具有高度相关,而与其他的各组变量却只有较小的相关或是不相关。这些组内高度相关的变量可以设想是一个共同的东西在影响着它们而导致高度相关。这个共同的东西称之为公共因子,剩余的部分称为特殊因子。

同 PCA 一样,因子分析也是一种线性方法。因子分析可以说是主成分分析推广与发展。如果说主成分分析是将原指标加以综合、归纳,那么因子分析可以说是将原指标给予分解、演绎。分子分析又称为析因分析或因素分析。

考虑 p 个成分的随机观测向量 $\boldsymbol{x}$。有均值为 u。因子模型要求线性相依,其中有 m 个公共因子 $F_1\ F_2\cdots F_m$ 和 p 个特殊因子 $\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_p$ 组成。因子分析的模型如下[11]:

$$x_1 - u_1 = l_{11}F_1 + l_{12}F_2 + \cdots + l_{1m}F_m + \varepsilon_1$$

$$\cdots$$

$$x_p - u_p = l_{p1}F_1 + l_{p2}F_2 + \cdots + l_{pm}F_m + \varepsilon_p$$

l_{ij} 称为因子载荷。因子载荷的统计意义就是第 i 个变量与第 j 个公共因子的相关系数,即表示变量 x_i 依赖于 F_j 的份量(比重),心理学家将它称为载荷。

因子分析的基本任务是:1) 建立因子载荷阵[11]。2) 给出各公共因子 $F_1\ F_2\cdots F_m$ 的合理解释及名称。3) 若有必要时(即一时难以找到合理解释的公共因子),进一步作因子旋转[11]。为了对公因子 F 能够更好的解释,可通过因子旋转的方法得到一个好解释的公因子。所谓对公因子更好解释,就是使每个变量仅在一个公因子上有较大的载荷,而在其余的公因子上的载荷比较小。

2.3　投影寻踪

2.3.1　投影法的基本思想

投影法(Projection Pursuit,PP)也是一种线性的方法,但不同于主成分分析或因子分析,可以处理高于二阶的信息,因此对于非高斯数据集非常有用[4]。

PP 利用计算机技术,把高维数据通过某种组合,投影到低维(1~3 维)子空间上,并通过极小化(或极大化)某个投影指标,寻找出能反映原高维数据结构或特征的投影,在低维空间上对数据结构进行分析,以达到研究和分析高维数据的目的。

PP 方法的一般方案是:1) 选定一个初始模型;2) 把数据投影到低维空间上,找出数据与现有模型相差最大的投影,这表明在此投影中含有现有模型中没有反映出来的结构;3) 将上述投影中包含的结构合并到现有模型上,得到改进了的新模型。然后再从此新模型出发重复以上步骤,直到数据与模型在任何投影空间都没有明显的差别为止。

2.3.2　PP 回归模型的算法

将 PP 法与传统的统计分析法相结合可以产生很多新的分析方法。其中的投影寻踪回归(PPR)模型如下:

设 $y = f(\vec{x})$ 和 $\vec{x} = (x_1,x_2,\cdots,x_k)$ 分别是一维和 k 维随机变量。为了能真实反映高维非线性数据的特征,根据 PP 法的基本算法,经过多次迭代后,就可以采用一系列岭函数 $Gm(z)$ 的和去逼近回归函数:

$$f(\vec{x}) \sim \sum_{m=1}^{M}\beta_m G_m(z) = \sum_{m=1}^{M}\beta_m G_m(\vec{\alpha}\cdot\vec{x}) = \sum_{m=1}^{M}\beta_m G_m\left(\sum_{j=1}^{k}\alpha_{jm}\cdot x_j\right)$$

式中,β_m、α_{jm} 是系数,$G_m(z)$ 是第 m 个岭函数。$z = \vec{\alpha}\cdot\vec{x}$ 为岭函数的自变量,它是 k 维随机变量在 $\vec{\alpha}$ 方向上的投影,$\vec{\alpha}$ 也是 k 维变量。M 是岭函数的个数。在上式中,一方面可以用增加岭函数个数 M 的方法减少模型的误差。另一方面,岭函数 $G_m(z)$ 是用逐段线性函数在各投影方向上不断对数据平滑逼近得到的。因此 PPR 模型更能客观地反映数据本身的内在结构和特征,从而增强了模型的稳定性[12]。

2.4　独立成分分析

2.4.1　独立成分分析原理

独立成分分析(Independent Components Analysis,ICA)是一种寻求线性投影的高阶方法,它不需要相互之间尽可能正交,而是尽可能相互在统计上相互独立。统计上相互独立是一种比不相关更强的条件,后者仅仅涉及到二阶的统计,而前者则关系到所有的高阶统计。如,随机向量 $\boldsymbol{X} = \{x_1,x_2,\cdots,x_p\}$ 是不相关的,则它表示对 $\forall i \neq j, 1 \leqslant i,j \leqslant p$,我们有:

$$D(x_i,x_j) = E\{(x_i - \mu_i)(x_j - \mu_j)\} = E(x_ix_j) - E(x_i)E(x_j) = 0$$

然而,相互独立则要求高阶概率密度函数可分解并写成:

$$f(x_1,x_2,\cdots,x_p) = f_1(x_1)\cdots f_p(x_p)$$

独立肯定意味着不相关,但是反之不然。只有当 $f(x_1,x_2,\cdots,x_p)$ 服从高阶正态分布,两者才能等价。对于高斯分布,主成分就是独立成分[4]。

ICA 的一般线性模型(不考虑噪声)为 $\boldsymbol{X} = \boldsymbol{As}$,其中 $\boldsymbol{x} = (x_1,x_2,\cdots,x_n)^T$ 为观察信号。$\boldsymbol{s} = (s_1,s_2,\cdots,s_m)^T$ 为独立的源信号且各分量服从非高斯分布,$\boldsymbol{A}$ 是 $n \times m$ 混合矩阵。ICA 的目的就是在仅知道 $\boldsymbol{X}$ 的情况下,寻找 $n \times m$ 混合矩阵 $\boldsymbol{A}$ 或 $m \times n$ 解混矩阵(又称为分离矩阵)$\boldsymbol{W}$,使 $\boldsymbol{Y} = \boldsymbol{WX}$,其中 $\boldsymbol{Y} = (y_1,y_2,\cdots,y_m)^T$ 且 $\boldsymbol{Y}$ 的各分量尽可能相互独立,而 $\boldsymbol{Y}$ 逼近 $\boldsymbol{s}$,从而得到源信号 $\boldsymbol{s}$[13]。

2.4.2　ICA 的估计准则

对于上述模型的评估一般包含两个步骤:指定目标函数(又称对比函数)和优化目标函数算法[4]。一般来说,不同的目标函数是由不同的估计准则得到的,然后通过恰当的优化方法来实现独立成分分析,也就是求出混合矩阵 $\boldsymbol{A}$ 和独立成分 $\boldsymbol{s}$,其中这些优化方法大多是基于梯度的方法,具体参见文献[13]、[14]。常用的评估准则有:

1) 非高斯性最大化

非高斯性是独立成分分析的本质特征,而传统的统计学习方法(比如主成分分析方法)假设随机变量服从高斯分布。非高斯性最大化是独立成分分析的一个重要的估计准则。一般来说,衡量非高斯性的测度有峰度(Kurtosis)和负熵(又称净熵 Negentropy)[4]两种。粗略的说非高斯就是相互独立,由中心极限定理,我们知道在一定条件下,相互独立的随机变量的和趋向于正态分布,所以一般来说两个相互独立的随机变量的和比任何一个参与求和的随机变量更加靠近正态分布。事实上,为了求出一个独立成分,我们考虑 $x_i(i = 1,\cdots,n)$ 的线性组合:

$$y = \boldsymbol{b}^T\boldsymbol{x} = \sum_i \boldsymbol{b}_i x_i$$

其中 $\boldsymbol{b}$ 是要估计的向量。如果记 $q^T = \boldsymbol{b}^T\boldsymbol{A}$,则上式可以表示为:

$$y = b^Tx = b^T\boldsymbol{As} = q^T\boldsymbol{s} = \sum_i q_i s_i$$

因此,从这个表达式可以看出,如果 b^T 恰好是混合矩阵逆的某一行时,则这个线性组合就恰好表示了一个独立成分。

事实上,可以由中心极限定理来确定这个独立成分,就像上面提到的,两个独立随机变量的和比其中的任何一个随机变量更加接近于高斯分布,也就是说上述表达式蕴含着随机变量,比任何一个独立成分更加靠近高斯分布,只有当这个随机变量恰好等于其中一个独立成分时,y 离高斯分布最远。因此,我们可以求最优值 b 使得 b^Tx 非高斯性最大化,则 y 就是一个独立成分。作为衡量随机变量 y 非高斯性的一个传统方法,峰度 $kurt(y)$ 在统计学上表示为:

$$kurt(y) = E\{y^4\} - 3(E\{y^2\})^2$$

另外一种度量非高斯性的方法是负熵,定义为:

$$J(y) = H(y_{gauss}) - H(y)$$

其中,y_{gauss} 表示服从高斯分布的随机变量,且与 y 具有相同的方差[13,14]。

2) 极大似然估计方法

极大似然估计方法也是实现独立成分分析的一个很有效的方法,是统计学习的一个最基本的方法,它的应用最为广泛和实用,其基本思想就是,求模型中的参数使得样本的概率分布达到最大。

对于模型 $\boldsymbol{X} = \boldsymbol{As}$ 中 $\boldsymbol{X}$ 的概率分布,可以表示为:

$$P_X(X) = |\ det(W)\ |\ p_s(s) = |\ det(W)\ |\ \prod_i p_i(s_i)$$

其中,$\boldsymbol{W} = \boldsymbol{A}^{-1}$ 表示分离矩阵,p_i 表示各个独立成分的概率密度函数。则上式表示为 $\boldsymbol{W} = (w_1, w_2, \cdots, w_n)^T$ 和 $\boldsymbol{x}$ 的函数:

$$P_X(X) = |\ det(W)\ |\ p_s(s) = |\ det(W)\ |\ \prod_i p_i(w_i^T x)$$

现在假设 T 个观察样本点表示为:$X(1), \cdots, X(T)$,且样本点是相互独立的,则得到样本的似然函数:

$$P(X) = \prod_{t=1}^{T} p(X(t)) = \prod_{t=1}^{T} \{|\ det(W)\ |\ \prod_i p_i(w_i^T X(t))\}$$

最大化样本的似然函数等价于最大化它的对数似然函数[14,15]:

$$L(W) = \log\{p(X)\} = T\log|\ det(W)\ | + \sum_{t=1}^{T}\sum_{i=1}^{n} \log\{p_i(w_i^T X(t))\}$$

3) 基于互信息最小的方法

基于互信息的独立成分分析是很重要的一个内容,这个方法不仅仅局限于独立成分分析中,也在其他的领域中有着广泛的应用。随机向量 $\boldsymbol{y} = (y_1, \cdots, y_n)^T$ 的各个元素之间的互信息表示为:

$$I(y_1, \cdots, y_n) = \sum_{i=1}^{n} H(y_i) - H(y)$$

互信息有一个重要的性质:$I(y_1, \cdots, y_n) \geqslant 0$,当且仅当随机向量 $\boldsymbol{y}$ 的各个分量之间相互独立时取零[13,14]。

3 结语

以上介绍的主要是线性的维规约方法,但现实中非线性情形更具适用性。对于非线性的维规约方法目前也有很多研究和发展,如:主曲线、多维标度法、神经网络、基因遗传算法等[4],具体领域使用的维规约技术的比较、改进的或综合的维规约算法、新的维规约算法、维规约算法的效率等仍将是研究的热点。

本文借鉴国外已有关于维规约的综述[3,15],结合国内对各种相关维规约技术研究的成果,从维规约技术的两个方面(属性选择和维变换)概述了常用的维规约技术思想和原理。

参考文献:

[1] 刘明吉, 王秀峰, 黄亚楼. 数据挖掘中的数据预处理[J]. 计算机科学, 2000, 27(4): 54-57.

[2] HAN JW, KAMBER M. 数据挖掘概念与技术[M]. 北京: 机械工业出版社, 2001.

[3] FODOR IK. A Survey of Dimension Reduction Techniques [DB/OL]. Center for Applied Scientific Computing. http://www.llnl.gov/CASC/sapphire/pubs/148494.pdf, 2002.

[4] YE J-P, LI Q, XIONG H, *et al.* IDR/QR: An Incremental Dimension Reduction Algorithm via QR Decomposition [J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(9): 1208-1222.

[5] 李喆, 王仁超, 褚春超. 数据挖掘中的数据预处理与维度优化[J]. 东北林业大学学报, 2003, 31(3): 70-72.

[6] 于之虹, 郭志忠. 改进主成分分析法用于暂态稳定评估的输入特征选择[J]. 电力自动化设备, 2003, 23(8): 17-21.

[7] 万星火, 檀亦丽. 主成分分析原始数据的预处理问题[J]. 中国卫生统计, 2005, 22(5): 327-329.

[8] 刘直芳, 游志胜, 王运琼. 基于概率主成分分析的人脸识别[J]. 红外与激光工程, 2004, 33(1): 71-75.

[9] TIPPING ME, BISHOP CM. Probabilistic Principal Component Analysis [Z], 1999.

[10] 赵广社, 张希仁. 基于主成分分析的支持向量机分类方法研究[J]. 计算机工程与应用, 2004: 37-38.

[11] 余锦华, 杨维权. 多元统计分析与应用[M]. 广州: 中山大学出版社, 2005.

[12] 高建芸, 宋德众, 林秀芳. 影响福建热带气旋年季频数的投影寻踪回归模型[J]. 热带气象学报, 2004, 20(4): 443-448.

[13] 杨竹青, 李勇, 胡德文. 独立成分分析方法综述[J]. 自动化学报, 2002, 28(5): 762-772.

[14] 钟明军. 独立成分分析算法研究及其在功能核磁共振成像中的应用[DB/OL]. http://218.69.114.37/wf/~CDDBN/Y665703/PDF/index.htm, 2006.

[15] BURA E. Dimension Reduction Techniques: A Review [DB/OL]. The George Washington University. http://srccs.snu.ac.kr/Workshop/04Statistics/7.pdf, 2006.

(上接第 2392 页)

[7] ZHOU AY, ZHOU SG, CAO J, *et al.* approaches for scaling DBSCAN algorithm to large spatial database[J]. Journal of computer science and technology, 2000, 15(06): 509-526.

[8] 蔡颖琨, 谢昆青, 马修军. 屏蔽了输入参数敏感性的 DBSCAN 改进算法[J]. 北京大学学报(自然科学版), 2004, 40(3): 480-486.

[9] KRIEGEL H-P, PFEIFLE M. Density-based clustering of uncertain data[A]. Proc. 11th Int. Conf. on Knowledge Discovery and Data Mining [C]. Chicago, IL, 2005. 672-677.

[10] HO TK, KLEINBERG EM. Checkerboard dataset[EB/OL]. http://www.cs.wisc.edu/math-prog/mpml.html, 1996.

[11] NEWMAN DJ, HETTICH S, BLAKE CL, *et al.* UCI Repository of machine learning databases[EB/OL]. http://www.ics.uci.edu/~mlearn/MLRepository.html. Irvine, CA: University of California, Department of Information and Computer Science, 1998.