具有容噪特性的C4.5算法改进

王 伟 李 磊 张志鸿
(郑州大学信息工程学院 郑州 450001)

摘 要 针对有噪声的高维数据引起决策树预测准确率下降的问题,利用容噪主成分分析(Noise-free Principal Component Anlysis,NFPCA)算法思想对C4.5算法改进而形成NFPCA-in-C4.5算法。该算法一方面将高维数据噪声控制问题转化为拟合数据特征与控制平滑度相结合的最优化问题,从而获得主成分空间;另一方面在决策树自顶向下构建新节点的过程中,再将主成分空间恢复到原始数据空间来避免降维过程中属性特征信息永久消失。实验结果表明NFPCA-in-C4.5算法兼具降维和容噪功能,避免了降维中由特征信息损失和噪声残留造成的预测模型准确率大幅降低的问题。

关键词 高维数据噪声,容噪,主成分分析,C4.5算法
中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.12.057

Improvement of C4.5 Algorithm with Free Noise Capacity

WANG Wei LI Lei ZHANG Zhi-hong
(School of Information Engineering, Zhengzhou University, Zhengzhou 450001, China)

Abstract Against the decline of decision tree prediction accuracy rate for high-dimensional data with noise, this paper used the theory of noise-free principal component analysis(NFPCA) algorithm to improve C4.5 algorithm, forming the NFPCA-in-C4.5 algorithms. On one hand, the algorithm transforms the noise suppression problem into an optimization problem of a combination of fitting the data and controling the smoothness, getting the space of principal components. On the other hand, it lets the space of principal components back to the space of original data during the process of building a new node in the decision tree from the top to down, to avoid the loss of characteristic information permanently in the dimension reduction process. Experimental results show that NFPCA-in-C4.5 algorithm has effects of dimensionality reduction and noise reduction, and avoids significant reduction of prediction accuracy rate, which is caused by the loss of information and noise.

Keywords High-dimensional data with noise, Noise tolerance, Principal component analysis, C4.5 algorithm

1 引言

在实际应用中经常会碰到高维数据,如金融交易数据、环境监测数据、图像数据、Web数据等。由于这种数据普遍存在,使得对高维数据挖掘的研究有着非常重要的意义[1]。一般的机器学习算法和模型都假设数据是精确的且含噪声少,而现实世界并非如此,会产生各种噪声,使其数据有所偏差[2],在高维数据环境中更是如此,高维数据中的冗余特征和噪声特征往往由于维数过多、过度拟合等原因,造成分类预测模型效果不佳[3]。针对此问题,许多国内外研究者在此领域进行了大量的探索。

Mantas等[4]提出一种Credal-C4.5算法,该算法利用模糊概率思想获得特征与类变量之间的概率,同时使用一种新的属性分裂准则来评价此种模糊概率分布;陈家俊等[5]提出一种基于多尺度粗糙集模型的决策树改进算法,该算法引入尺度变量和尺度函数,采用不同尺度下的近似分类精度选择测试属性,同时结合抑制因子对决策树进行修剪。上述算法都采用新的属性选择准则来处理噪声,降低噪声残留,但同时却忽略了高维数据的高维特性,使得生成的决策树预测模型过度复杂,而这往往造成预测准确率不佳[6]。

孟凡荣等[7]提出一种PCA-DT算法,该算法运用主成分分析方法[8]来对信息增益和相关性度量这两个因子降维,得到一个综合指标,以此来选择优先划分的条件属性,但是该算法仅将高维数据降噪处理放在决策树算法之前,作为纯粹的预处理操作,未能与分类算法进行深度融合;而且高维数据的降维处理也造成一定的信息损失,这些都使得针对含噪声高维数据的分类预测模型的准确率不高。

本文提出一种NFPCA-in-C4.5算法,该算法利用文献[9]提出的NFPCA算法思想对C4.5算法进行改进,避免了降维过程中因特征信息损失而造成的预测准确率大幅下降的问题,同时提升了决策树算法的容噪能力,相比于传统C4.5算法生成的决策树模型,其预测准确率更具稳定性。

本文第2节介绍了PCA算法处理含噪声高维数据的不足和噪声影响主成分的机理,进而引入NFPCA算法;第3节

到稿日期:2014-12-22 返修日期:2015-03-13 本文受河南省烟草专卖局科学研究与技术开发项目(HYKJM201335)资助。
王 伟(1989-),男,硕士生,主要研究方向为数据挖掘,E-mail:ww6195207@163.com;李 磊(1974-),男,博士,讲师,主要研究方向为信息安全、数据挖掘;张志鸿(1965-),男,博士,教授,CCF高级会员,主要研究方向为服务计算、信息安全、数据挖掘。

介绍了对传统 C4.5 算法改进后的 NFPCA-in-C4.5 算法，并分别阐述了该算法的两个阶段，即决策树构建阶段和决策树预测阶段；第 4 节通过设计 3 组实验对比分析传统 C4.5 算法与 NFPCA-in-C4.5 算法在不同噪声水平中的预测准确率，从而验证所提算法的有效性；最后总结全文。

2 NFPCA 算法

2.1 PCA 算法的局限性

主成分分析法(Principal Component Analysis，PCA)算法是由 Hotelling[10]提出的一种用低维数据表示高维复杂数据最主要特征的多元统计分析方法。其特征提取的主要思想[11]是：首先在原始数据空间中找到能最大程度表示原始数据方差的一组正交向量，然后将原始数据空间从 n 维投影到这组正交向量构成的 m 维特征空间，从而在不改变原始数据结构的情况下，得到两两独立且正交的由原始属性线性组合而构成的主成分。

然而，当处理的高维数据夹杂噪声时，传统主成分分析(PCA)的缺点非常明显：由于从数据集中提取出的主成分是所有特征的线性组合，因此，如果数据中噪声水平较高，将会导致主成分的计算受到噪声污染。其噪声影响机制如下。

假设数据矩阵为 X，维数为 n，噪声符合高斯分布但噪声水平未知。通过奇异值分解 SVD[12]得到特征值 d_i 和特征向量 u_i，而其无噪声时的特征矩阵为 $V=[v_i](i=1,\cdots,n)$。

当数据矩阵 X 含噪声水平为 0 时，则满足：

$$Xv_i=d_iu_i \tag{1}$$

经奇异值分解得到的特征向量 u_i 是数据矩阵 X 中属性的线性组合，若 X 被噪声影响而发生变化，那么 u_i 也必定会发生变化。这意味着线性方程(1)的解可能有较大误差，会导致在数据集特征向量 u_i 中存在噪声。此时的 u_i 并不是无噪声影响下 X 的特征向量，而是夹杂着噪声的影响，我们将这种影响称为特征向量 u_i 受到的噪声污染。

因此，通过传统的 PCA 算法得到的含有噪声的特征向量组成的主成分矩阵 V 可能是失效的，并不能很好地体现无噪声数据集的主要特征；而当主成分作为输入变量代替原有的数据集属性特征时，由于主成分中噪音的存在，也在很大程度上降低了分类算法的预测能力[9]。

2.2 NFPCA 算法的基本思想

NFPCA 算法是在传统 PCA 算法的基础上进行改进的，使其在降维过程中进行容噪，最终找到非常接近无噪声数据的特征向量矩阵 V 和特征值 λ。

NFPCA 算法的基本思想是：利用噪声数据有明显振荡行为的特性，通过控制它的振荡行为来达到控制向量中噪声的目的[13]，首先将振荡行为控制和主要特征保留的问题转化为最优化问题，然后运用共轭梯度法(CGLS)求解获得特征向量矩阵 V 和特征值 λ。

NFPCA 算法用如下准则描述数据的振荡程度：假设一个向量 $\chi=(\chi_1,\cdots,\chi_n)^{\mathrm{T}}\in R^n$，定义

$$\Omega(\chi)=\sum_{i=1}^{n-1}(\chi_{i+1}-\chi_i)^2 \tag{2}$$

式(2)可以用来表示向量 $n\times m$ 的振荡程度。在某种意义上，如果 $\Omega(\chi)$ 值越大，那么向量 χ 的振荡程度越大；$\Omega(\chi)$ 值越小，意味着向量的振荡程度越小或更加平滑。

当存在噪声时，$Xv_i\neq d_iu_i$，为了尽量拟合数据中的主要特征，此时的目标函数为：

$$\min_{v_i}\|Xv_i-y_i\|_2 \tag{3}$$

其中，$y_i=d_iu_i$。

因此，为了在保持平滑的同时，尽量保证拟合数据的主要特征趋势，需要将 $\Omega(v_i)$ 和 $\|Xv_i-y_i\|_2$ 同时考虑。因此采用将两者的和值最小化的方法，既保证了拟合数据中的主要特征，又避免了过度振荡带来的影响。所以，可以将其转化为如下最优化问题：

$$\min_{v_i}\{\|Xv_i-y_i\|_2^2+\lambda_i^2\|Lv_i\|_2^2\} \tag{4}$$

其中，$L=\begin{pmatrix}-1 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & \ddots \\ & & -1 & 1\end{pmatrix}\in R^{(n-1)\times n}$，式(4)可等价转化为：

$$\min_{v_i}\left\|\begin{pmatrix}X\\ \lambda_iL\end{pmatrix}v_i-\begin{pmatrix}y_i\\ 0\end{pmatrix}\right\|_2^2 \tag{5}$$

通过 NFPCA 算法可以获得正则化的主成分向量，其伪代码如算法 1 所示。

算法 1 NFPCA 算法

输入：数据矩阵 X，主成分个数 k；

输出：$V=[v_1,\cdots,v_k]$

步骤 1 运用 SVD 计算奇异向量 u_i 和奇异值 $V=v_i$，其中 $i=1,\cdots,k$

步骤 2 for i=1:k do

$y_i=d_iu_i$

运用共轭梯度法(如算法 2 所示)计算正则化参数 λ_i，解决如下最小二乘问题：$v_i=\underset{v}{\operatorname{argmin}}\left\|\begin{pmatrix}X\\ \lambda_iL\end{pmatrix}v_i-\begin{pmatrix}y_i\\ 0\end{pmatrix}\right\|_2^2$

end for

使用共轭梯度法(CGLS)解决算法 1 中最小化问题[14,15]，以便获得过滤掉噪声的解 v_i。CGLS 算法的伪代码如算法 2 所示。

算法 2 解决最小二乘问题的共轭梯度法(CGLS)

步骤 1 $v_i^{(0)}$ 作为初始近似解，

$A=\begin{pmatrix}X\\ \lambda_iL\end{pmatrix}$，$b=\begin{pmatrix}y_i\\ 0\end{pmatrix}$，$\chi^{(0)}=v_i^{(0)}$，

$r^{(0)}=b-A\chi^{(0)}$，$p^{(0)}=s^{(0)}=A^{\mathrm{T}}r^{(0)}$，$\gamma=\|s^{(0)}\|_2^2$

步骤 2 for k=0 until $\gamma_k>$tolerance do

$q^{(k)}=Ap^{(k)}$，$\alpha_k=\gamma_k/\|q^{(k)}\|_2^2$

$\chi^{(k+1)}=\chi^{(k)}+\alpha_kp^{(k)}$，$\gamma^{(k)}=\gamma^{(k)}-\alpha_kq^{(k)}$

$s^{(k+1)}=A^{\mathrm{T}}\gamma^{(k+1)}$，$\gamma_{k+1}=\|s^{(k+1)}\|_2^2$，$\beta_k=\gamma_{k+1}/\gamma_k$

$p^{(k+1)}=s^{(k+1)}+\beta_kp^{(k)}$

end for

步骤 3 计算 $v_i=\chi^{(k+1)}$

相比于传统的 PCA 算法，利用上述思想提出的 NFPCA 算法具有容噪特性，不仅可以降低数据维数，更削弱了噪声对数据的污染程度，为决策树节点中数据集由高维含噪声空间映射到低维低噪空间提供了途径。

3 NFPCA-in-C4.5 算法设计

本文利用 NFPCA 算法思想提出一种改进的决策树算法，即 NFPCA-in-C4.5 算法，解决了 C4.5 决策树算法在高维数据环境下容噪效果不佳的问题。该算法在自顶而下构建决策树的过程中，随着节点的不断建立，首先在原始数据空间

下，对节点所含数据集进行 NFPCA 算法处理，在降维处理的同时降低节点中数据集中的噪声，从而获得相对无噪声的主成分特征向量和样本估计数据集；然后在主成分空间下运用基于信息熵的属性选择方法，对主成分属性进行选择，形成属性分裂条件，进而在主成分空间中将估计样本集划分成多个子集，放入对应的新节点中；最后在新节点中将主成分空间下的数据集恢复到原始数据空间下对应的子集，再反复按此方式迭代，直至满足叶节点形成条件。

与传统的 C4.5 算法类似，NFPCA-in-C4.5 算法主要由两阶段构成，即决策树构建阶段和决策树预测阶段。

3.1 决策树构建阶段

设原始数据集 X 有 n 条记录，每条样本记录有 m 个属性，x_{ij} 为第 i 条记录第 j 个属性值，整个原始数据集 X 构成的 $n\times m$ 矩阵为：

$$X_{n\times m}=[x_{ij}]_{n\times m}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

原始数据集 X 中 $A_1,\cdots,A_{m-1},A_m$ 为各属性列，A_m 为类别属性，其余属性是非类别属性且为数值型。定义数据矩阵 $T=[A_1,\cdots,A_{m-1}]$，矩阵 T 经过 NFPCA 算法处理后在主成分空间下得到对应的估计矩阵为 T^*，则原始样本数据集 X 在主成分空间下对应的数据集 $X^*=[T,A_m]$，数据集 X^* 的各主成分属性矩阵为 $V=(V_1,\cdots,V_e)$，同时也是 X^* 的非类别属性，A_m 为 X^* 类别属性。

假设主成分空间下的数据集 X^* 中类别属性 A_m 的值为 $c_p(p=1,\cdots,k)$，那么对于 A_m 的 k 个不同取值 c_p，将数据集 X^* 划分为 $C_1,\cdots,C_k$；而非类别属性 $V_i(i=1,\cdots,e)$ 的值为 $v_q(q=1,\cdots,t)$，那么对于 V_i 的 t 个不同取值，将数据集 X^* 划分为 $S_1,\cdots,S_t$；同时，当非类别属性 $V_i=v_q$ 且类别属性 $A_m=c_p$ 时，对应地将数据集 X^* 划分为 C_{pq}，其中 $p=1,\cdots,k$ 且 $q=1,\cdots,t$。

以根节点为当前节点，NFPCA-in-C4.5 算法构建决策树模型的流程如图 1 所示，具体步骤如下。

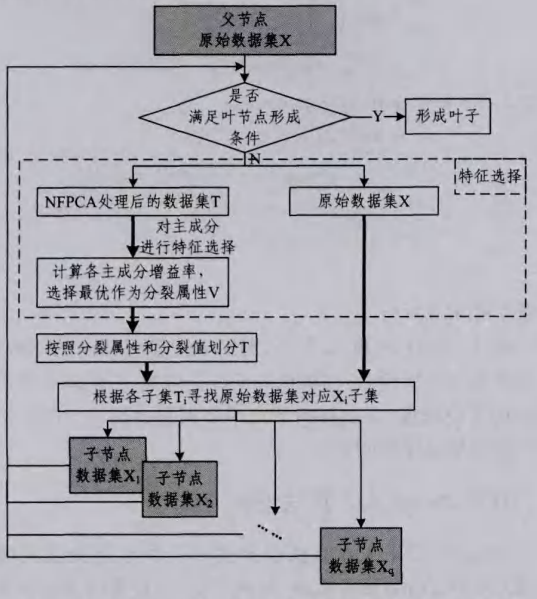

图 1 NFPCA-in-C4.5 算法在决策树构建阶段的流程

步骤 1 假设当前节点中的原始样本数据集为 X，经 NFPCA 算法进行降维降噪处理得到特征向量矩阵 V 与特征值 λ。其中各特征值 λ_i 满足从大到小的排列顺序 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_m\geqslant 0$，特征向量 v_i 满足 $\|v_i\|=1$。

步骤 2 计算 $R_{m\times m}$ 矩阵的特征值对应的贡献率 ω_i 和累计贡献率 η_e：

$$\omega_i=\lambda_i/\sum_{k=1}^{m}\lambda_k,i=1,2,\cdots,m \tag{6}$$

$$\eta_e=\sum_{i=1}^{e}\lambda_i/\sum_{k=1}^{m}\lambda_k \tag{7}$$

将累计贡献率 η_e 第一次达到 90% 以上时的特征值 $\lambda_i(i=1,2,\cdots,e)$ 作为主成分。

步骤 3 计算原始数据集 $X_{n\times m}$ 在该主成分维度空间下的估计矩阵 $X^*_{n\times e}$：

$$X^*_{n\times e}=X_{n\times m}V_{m\times e} \tag{8}$$

步骤 4 根据主成分空间下数据集 X^* 中各类别概率 $P(C_p)$，计算类别信息熵：

$$H(C)=-\sum_{p=1}^{k}P(C_p)\log_2 P(C_p) \tag{9}$$

其中在类别 $A_m=c_p$ 的概率：

$$P(C_p)=|C_p|/|X^*| \tag{10}$$

计算主成分 $V_i=v_i$ 时的分裂信息熵，从中选择分裂时信息熵最大的值作为分裂点。

$$SplitInfo(X^*,V_i)=-\sum_{q=1}^{t}P(S_q)\log_2 P(S_q) \tag{11}$$

其中 $P(S_q)=|S_q|/|X^*|$。

步骤 5 从数据集 X^* 中选取一个主成分属性 V_i，计算在主成分 $V_i=v_p$ 条件下，各类别 c_j 的条件信息熵为：

$$H(C/V_i)=-\sum_{q=1}^{t}P(S_q)\left[-\sum_{p=1}^{k}P(C_{pq})\log_2 P(C_{pq})\right] \tag{12}$$

其中主成分 $V_i=v_i$ 条件下，类别 $A_m=c_j$ 的条件概率为：

$$P(C_{pq})=|C_{pq}|/|X^*| \tag{13}$$

步骤 6 计算信息增益，进而计算信息增益率：

$$InfoGain(X^*,V_i)=H(X^*)-H(X^*,V_i) \tag{14}$$

$$GainRatio(X^*,V_i)=\frac{InfoGain(X^*,V_i)}{SplitInfo(X^*,V_i)} \tag{15}$$

步骤 7 将各主成分中信息增益率最高的主成分 V_i 作为最佳属性特征，同时将此主成分 V_i 达到最佳时的分裂点 v_i 作为分支划分条件，将经 NFPCA 算法处理得到的数据集 X^* 进行划分，得到对应各子节点的子集 X_i^*；根据子集 X_i^* 在原始数据集中所对应的序号子集 M_i，找到与 X_i^* 对应的原始数据子集 X_i，将其对应放入新生成的各子节点中，然后对子节点中原始数据空间下的数据集 X_i 再次执行步骤 1，直至符合叶子节点条件时终止。

3.2 决策树预测阶段

假设原始预测数据集 Z 在决策树模型的顶层根节点中，且其属性类型均为数值型，那么 NFPCA-in-C4.5 算法在决策树预测阶段的流程如图 2 所示。

预测数据集类别的具体步骤如下。

步骤 1 判断本节点是否为叶子节点，若是叶子节点，则本节点数据集 Z 的类别为叶子节点所属类别，然后本节点预测结束；若是非叶子节点，则执行步骤 2。

步骤 2 先使用主成分矩阵 V 计算 $P=ZV$，将其投影到主成分空间，再根据本层节点主成分属性分支条件划分为子集 P_i。

步骤 3 根据主成分空间下各子集 P_i 在原始预测数据集 Z 中对应的序号子集 M_i，找到原始预测数据空间下的对应子集 Z_i。

步骤 4　将预测数据子集 Z_i 作为子节点中的预测数据集；各子节点执行步骤 1，直到全部数据集都获得预测类别。

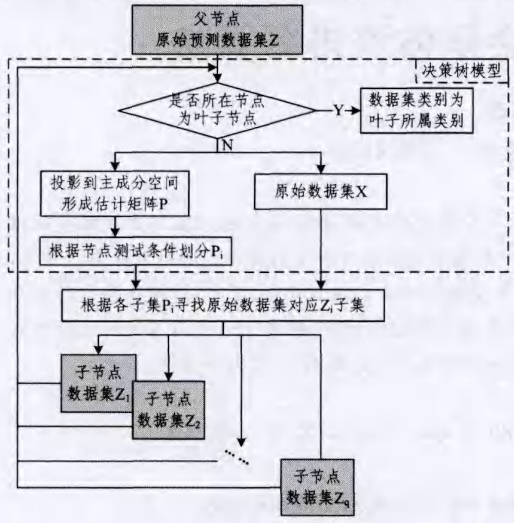

图 2　NFPCA-in-C4.5 算法在决策树预测阶段的流程

4　实验及结果分析

本文使用 UCI 官方提供的 6 组高维数据集（包含 Audiology、Dermatology、Mfeat-pixel、Optdigits、Spambase 和 Splice 数据集）设计了 3 组实验，对比 NFPCA-in-C4.5 算法与 C4.5 算法在不同噪声程度下的预测准确率变化。

4.1　实验环境与数据集描述

实验环境：处理器为 Intel Core i7-2620M @ 2.70GHz，RAM 4GB，操作系统为 Windows 7，算法基于 Matlab 2013a 实现。实验中所用数据集由 UCI 提供，各数据集描述如表 1 所列。

表 1　实验所用数据集描述

数据集	样例数	属性数	类别数
Audiology	226	69	24
Dermatology	366	34	6
Mfeat-pixel	2000	240	10
Optdigits	5620	64	10
Spambase	4601	57	2
Splice	3190	60	3

4.2　实验设计与分析

首先对各数据集中离散型属性采用二进制转换处理，连续属性采用规范化处理。然后设计以下 3 组对比实验进行分析：实验一在无噪声环境下进行，实验二和实验三在各数据集中分别加入 10%高斯噪声水平和 40%高斯噪声水平，以模拟真实环境下数据集各数据属性值受噪声干扰的情况。最后采用十折交叉验证法[16]计算预测准确率，对比 C4.5 算法和 NFPCA-in-C4.5 算法在各组实验中预测准确率的变化。

实验一：对比各数据集在无噪声条件下的准确率，结果如表 2 所列。由表 2 可知，NFPCA-in-C4.5 算法的准确率普遍略低于传统的 C4.5 算法，这是由于 NFPCA-in-C4.5 算法在原数据空间中提取主成分时，损失了部分数据特征，造成分类准确率下降；同时在每次节点划分时又恢复到原数据空间，重新进行主成分提取，大大避免了特征永久丢失的情况发生，因此 NFPCA-in-C4.5 算法与 C4.5 算法的预测准确率只有略微差别。

表 2　在无噪声条件下，C4.5 和 NFPCA-in-C4.5 算法的准确率比较

数据集	C4.5(%)	NFPCA-in-C4.5(%)
Audiology	77.26	76.58
Dermatology	94.1	93.46
Mfeat-pixel	78.66	78.12
Optdigits	90.54	89.93
Spambase	92.45	92.14
Splice	94.11	93.85

实验二：对各数据集加入 10%高斯噪声，对比两种算法的预测准确率变化情况，如表 3 所列。

表 3　在 10%噪声水平下，C4.5 和 NFPCA-in-C4.5 算法的准确率比较

数据集	C4.5(%)	NFPCA-in-C4.5(%)
Audiology	75.03	76.16
Dermatology	92.34	92.74
mfeat-pixel	75.78	77.03
Optdigits	87.49	89.36
Spambase	89.12	90.89
Splice	92.05	93.23

表 3 中传统 C4.5 算法由于受噪声影响，其准确率进一步降低，平均降幅为 2.55%，最大降幅达到 3.3%；而 NFPCA-in-C4.5 算法预测准确率超过 C4.5 算法，这是因为其容噪功能开始发挥作用，避免了噪声带来的错误判断，预测准确率比传统 C4.5 算法最多高出 2%。

实验三：对各数据集加入更强程度的噪声（40%高斯噪声），对比两种算法预测准确率的变化情况，如表 4 所列。

表 4　在 40%噪声水平下，C4.5 和 NFPCA-in-C4.5 算法的准确率比较

数据集	C4.5(%)	NFPCA-in-C4.5(%)
Audiology	68.88	73.68
Dermatology	86.56	91.49
Mfeat-pixel	69.66	74.52
Optdigits	75.93	87.43
Spambase	86.43	88.7
Splice	80.11	91.34

从表 4 可知，传统 C4.5 算法的准确率进一步下降，且与无噪声条件下相比，此次下降幅度在 7.5%～14.6%内，平均降幅达到 9.9%，这种幅度的下降表明 40%的噪声已对传统 C4.5 算法产生了很大影响；而 NFPCA-in-C4.5 算法的预测准确率比 C4.5 算法提高幅度最高达到 11.5%，平均提升幅度达到 6.6%，且依然保证较高的准确率。

3 组实验的预测准确率下降幅度如图 3 所示，横轴以数字编号（1～6）代表各数据集，从左至右依次是 Audiology、Dermatology、Mfeat-pixel、Optdigits、Spambase 和 Splice 数据集。纵轴为相比于无噪声条件下预测准确率的下降幅度。

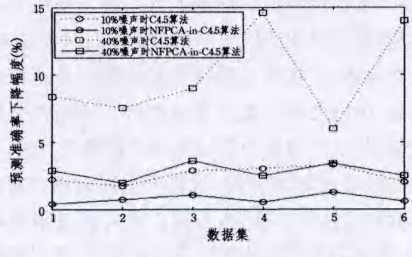

图 3　相比无噪声环境，C4.5 算法和 NFPCA 算法在 10%和 40%噪声环境中预测准确率的下降幅度

结合表 2—表 4 和图 3 中实验结果可以得出以下结论：在无噪声条件下，C4.5 算法与 NFPCA-in-C4.5 算法对 6 组高维数据集的预测准确率基本相同，未出现因降维造成的准

（下转第 287 页）

[20] de Vienne D M, Giraud T, Martin O C. A congruence index for testing topological similarity between trees [J]. Bioinformatics, 2007, 23(23): 3119-3124

[21] Pompei S, Loreto V, Tria F. On the accuracy of language trees [J]. PloS one, 2011, 6(6): e20109

[22] Hayes M, Walenstein A, Lakhotia A. Evaluation of malware phylogeny modelling systems using automated variant generation [J]. Journal in Computer Virology, 2009, 5(4): 335-343

[23] Swofford D L. When are phylogeny estimates from molecular and morphological data incongruent [M]// Phylogenetic Analysis of DNA Sequences. 1991: 295-333

[24] Bryant D. Building trees, hunting for trees, and comparing trees: theory and methods in phylogenetic analysis [D]. Dept. of Math., Univ. of Canterbury, 1997

[25] Day W H. Optimal algorithms for comparing trees with labeled leaves [J]. Journal of Classification, 1985, 2(1): 7-28

[26] Robinson D, Foulds L R. Comparison of phylogenetic trees [J]. Mathematical Biosciences, 1981, 53(1): 131-147

[27] Steel M A, Penny D. Distributions of tree comparison metrics—some new results [J]. Systematic Biology, 1993, 42(2): 126-141

[28] Bogdanowicz D, Giaro K. On a Matching Distance Between Rooted Phylogenetic Trees [J]. International Journal of Applied Mathematics and Computer Science, 2013, 23(3): 669-684

[29] Gabow H N, Tarjan R E. Faster scaling algorithms for network problems [J]. SIAM Journal on Computing, 1989, 18(5): 1013-1036

[30] Bogdanowicz D, Giaro K. Matching split distance for unrooted binary phylogenetic trees [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2012, 9(1): 150-160

[31] Lin Y, Rajan V, Moret B M. A metric for phylogenetic trees based on matching [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2012, 9(4): 1014-1022

(上接第 271 页)

确率大幅下降的情况；在 10%和 40%噪声条件下，C4.5 算法预测准确率下降幅度最大达到了 14.6%，下降幅度较大，而 NFPCA-in-C4.5 算法预测准确率下降幅度始终保持在 3.5%以下，下降程度明显低于 C4.5 算法。因此实验结果说明了在不同程度的噪声环境中，NFPCA-in-C4.5 算法对高维数据的预测准确率具有稳定性，具备了容噪的特性。

结束语 本文提出的 NFPCA-in-C4.5 算法对含噪声的高维数据兼具降维和容噪功能，避免了降维信息损失和噪声残留造成准确率大幅降低的问题；在不同程度噪声水平下，在高维数据集上进行了 C4.5 算法和 NFPCA-in-C4.5 的算法对比实验，结果表明在噪声水平高的环境下，NFPCA-in-C4.5 算法预测准确率的稳定性得到大幅提高，体现了 NFPCA-in-C4.5 算法对高维数据的容噪特性优势。

完成决策树的构建后，所形成的决策规则由各主成分组成，后续算法将研究决策树规则中各主成分不易理解的问题。

参 考 文 献

[1] 杨风召. 高维数据挖掘中若干关键问题的研究[D]. 上海：复旦大学，2003
Yang Feng-zhao. The Research on A Few Key Issues in High Dimensional Data Mining[D]. Shanghai: Fudan University, 2003

[2] 承文俊，沈建强，谢琪，等. 容噪学习机制及其在 Robocup 中的应用研究[J]. 计算机科学，2004，32(4)：101-103
Cheng Wen-jun, Shen Jian-qiang, Xie Qi, et al. Research on Noise Tolerance Mechanism in Robocup[J]. Computer Science, 2004, 32(4): 101-103

[3] 倪春鹏. 决策树在数据挖掘中若干问题的研究[D]. 天津：天津大学，2004
Ni Chun-peng. Research on Some Problems of Decision Tree in Data Mining[D]. Tianjin: Tianjin University, 2004

[4] Mantas C J, Abellán J. Credal-C4.5: Decision tree based on imprecise probabilities to classify noisy data[J]. Expert Systems with Applications, 2014, 41(10): 4625-4637

[5] 陈家俊，苏守宝，徐华丽. 基于多尺度粗糙集模型的决策树优化算法[J]. 计算机应用，2011，12：3243-3246
Chen Jia-jun, Su Shou-bao, Xu Hua-li. Decision tree optimization algorithm based on multiscale rough set model[J]. Computer Applications, 2011, 12: 3243-3246

[6] Breiman L, Friedman J, Stone C J, et al. Classification and regression trees[M]. CRC press, 1984

[7] 孟凡荣，蒋晓云，田恬，等. 基于主成分分析的决策树构造方法[J]. 小型微型计算机系统，2008(7)：1245-1249
Meng Fan-rong, Jiang Xiao-yun, Tian Tian, et al. Decision Tree Construction Method Based on Principal Component Analysis [J]. Journal of Chinese Computer Systems, 2008(7): 1245-1249

[8] Jolliffe I. Principal component analysis [M]. Wiley Online Library, 2005

[9] Rezghi M, Obulkasim A. Noise-free principal component analysis: An efficient dimension reduction technique for high dimensional molecular data[J]. Expert Systems with Applications, 2014, 41(17): 7797-7804

[10] Hotelling H. Analysis of a complex of statistical variables into principal components[J]. Journal of Educational Psychology, 1933, 24(6): 417

[11] 周斯斯. 谱聚类维数约简算法研究与应用[D]. 西安：西安电子科技大学，2010
Zhou Si-si. Spectral Clustering Based Dimensionality Reduction and Applications[D]. Xi'an: Xi'an University of Electronic Science and Technology, 2010

[12] Golub G. Matrix computations[M]. Johns Hopkins University Press, 1996

[13] Hanke M, Hansen P C. Regularization methods for large-scale problems[J]. Surv. Math. Ind, 1993, 3(4): 253-315

[14] 树方，平文. 数值线性代数[M]. 北京：北京大学出版社，2000
Shu Fang, Ping Wen. Numerical Linear Algebra[M]. Beijing: Beijing University Press, 2000

[15] Björck A. Numerical methods for least squares problems[M]. Siam, 1996

[16] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection[C]// International Joint Conference on Artifical Intelligence. 1995: 1137-1143

[17] 王越，万洪. 一种新的应用变精度粗糙集的决策树构造方法[J]. 重庆理工大学学报(自然科学版)，2013，27(11)：58-64
Wang Yue, Wan Hong. A New Method for Constructing Decision Tree Based on Variable Precision Rough Set[J]. Journal of Chongqing University of Technology (Natural Science), 2013, 28(11): 58-64