基于超图的稀疏属性选择算法*

雷 聪[1]，钟 智[2†]，胡晓依[1]，方 月[1]，余 浩[1]，郑 威[1]
(1. 广西师范大学 广西多源信息挖掘与安全重点实验室，广西 桂林 541004；2. 广西师范学院 计算机与信息工程学院，南宁 530023)

摘 要： 针对噪声或者离群点通常会增加矩阵的秩的问题，提出一个在低秩限制下的基于超图的稀疏属性选择算法。算法利用其他属性稀疏地表达每一个属性来获得属性自表达系数矩阵，再利用超图正则化因子获取数据的局部结构将子空间学习嵌入到属性选择的框架中；同时，利用 $l_{2,p}$-范数惩罚自表达系数矩阵和损失函数，挖掘出属性之间的关系和样本间的关系来帮助算法有效地进行属性选择，最终提高模型的预测能力。在 UCI 数据集上的实验结果表明，该算法相比其他对比算法，能更有效地选取重要属性，并取得很好的分类效果。

关键词： 属性选择；属性自表达；子空间学习；超图表示；低秩约束
中图分类号： TP181 **文献标志码：** A **文章编号：** 1001-3695(2018)11-3213-04
doi:10.3969/j.issn.1001-3695.2018.11.003

Hypergraph-based sparse feature selection

Lei Cong[1], Zhong Zhi[2†], Hu Xiaoyi[1], Fang Yue[1], Yu Hao[1], Zheng Wei[1]
(1. *Guangxi Key Laboratory of Multi-source Information Mining & Security, Guangxi Normal University, Guilin Guangxi* 541004, *China*; 2. *College of Computer & Information Engineering, Guangxi Teachers Education University, Nanning* 530023, *China*)

Abstract: It is a fact that, during real data mining applications, noises or outliers can increase the rank of a matrix. This paper proposed a novel feature selection via hypergraph-based sparse structure combined with a low-rank constraint. Specially, it obtained the self-representation coefficient matrix through sparsely represent each feature by other features. Then, obtained the local structure of the data via a hypergraph-based regularizer, so as to integrate the subspace learning into the framework of feature selection. Meanwhile, it obtained the correlation between features via using an $l_{2,p}$-norm regularization to penalize the self-representation matrix. And designed the $l_{2,p}$-norm on the loss function for building the relation among samples. It used the correlation and relation for selecting those features that assisted in improving the performance. Experimental results demonstrate that the proposed method is much better than extant methods at classification performance and stability.

Key words: feature selection; feature self-representation; subspace learning; hypergraph representation; low-rank constraint

随着信息技术的发展，在数据挖掘、机器学习和多媒体索引等新兴领域产生的数据往往都是大数据集，它们包含众多的特征和大规模的记录，并导致了较高的处理时间和空间复杂度。为了有效地利用这些高维数据，且减少相应的处理时间，对数据的预处理就显得尤为重要。因此，如何对高维数据进行有效地属性约简[1]，发现一个具有代表性且规模较小的子集便成为一个重要的研究领域[2]。

一般地，属性约简方法包括属性选择和子空间学习[3,4]。属性选择通常是通过一些特定的模型从数据中提取一些必要的属性，从而达到减少数据维数的目的。属性选择中比较先进的方法有过滤方法[5-7]、封装方法[8,9]和嵌套方法[10,11]。而子空间学习是通过投影矩阵将高维数据投影到低维空间，以此来保持数据之间的关联结构，常见的方法有局部保留投影(locality preserving projection, LPP)[12]、主成分分析(principal component analysis, PCA)[13]、线性判别式分析(linear discriminant analysis, LDA)[14]算法等。

研究表明，能够捕捉到数据间不同的流形结构的子空间学习方法通常比属性选择方法具有更稳定的效果，而属性选择方法更具有解释性[15]，为此，本文首先结合子空间学习和属性选择进行属性约简以获取显著的效果。本文结合属性自表达性和稀疏学习对重要的属性进行提取，其中通过子空间学习的方法来考虑数据的全局结构(利用低秩限制实现)和局部结构(通过超图正则化因子实现)，提出一种更加高效的属性选择算法——基于超图稀疏的属性选择算法(hypergraph-based sparse feature selection，HS_FS)。由于本文同时考虑到了数据的全局结构和局部结构，所以比单一的子空间学习方法具有更好的效果。经实验验证，该算法在分类任务中已取得较好的结果。

1 相关理论背景及简介

给定属性集 $X=[x_1,x_1,\cdots,x_m]\in R^{n\times m}$($n$ 为样本数，m 为属性数)，用原始数据 X 中的全部属性来对属性 x_i($i=1,2,\cdots,m$) $\in R^{n\times 1}$ 进行表示的过程称为属性自表达。根据属性自表达的特性，需要找到一个列向量 $z_i\in R^{m\times 1}$，使得 x_i 能用 Xz_i 重新表示，其中 z_i 为自表达系数。由于某些属性是噪声或者冗余的，所以在重新表示属性的时候要考虑误差，那么新的属性集

收稿日期：2017-06-13；修回日期：2017-07-21 基金项目：国家重点研发计划资助项目(2016YFB1000905)；国家自然科学基金资助项目(61672177,61573270)；国家“973”计划资助项目(2013CB329404)；广西自然科学基金资助项目(2015GXNSFCB139011)；广西多源信息挖掘与安全重点实验室开放基金资助项目(16-A-01-01,16-A-01-02)；广西研究生教育创新计划项目(XYCSZ2017064,XYCSZ2017067,YCSW2017065)

作者简介：雷聪(1991-)，男，湖北大冶人，硕士研究生，主要研究方向为数据挖掘、机器学习；钟智(1963-)，男(通信作者)，广西梧州人，副教授，主要研究方向为机器学习和数据挖掘(CongL_hu@163.com)；胡晓依(1992-)，女，山东临沂人，硕士，主要研究方向为网络与信息安全；方月(1992-)，男，湖北孝感人，硕士研究生，主要研究方向为数据挖掘、机器学习；余浩(1994-)，男，江西上饶人，硕士研究生，主要研究方向为数据挖掘、机器学习；郑威(1989-)，男，吉林延吉人，硕士研究生，主要研究方向为数据挖掘、机器学习.

可以表示为

$$x_i = Xz_i + b \quad i = 1, 2, \cdots, n \tag{1}$$

其中: b 是重构误差。由于属性自表达系数是依赖于全部属性而不仅仅依赖某个单独的属性,所以,能够有效地降低离群点对模型的影响,增强模型的鲁棒性。

2 算法描述和优化

2.1 算法描述

假设给定训练集 $X \in \mathrm{R}^{n\times m}$,其中,$n$ 和 m 分别表示样本数和属性数。由于样本的类标签难以获取或者获取成本较高,无监督属性选择方法已经成为研究的热点[16]。近些年来,由于属性自表达特性的存在,使得无监督属性选择大行其道。本文依据属性自表达特性来获取属性之间的联系:

$$X = XZ + eb \tag{2}$$

其中: $Z \in \mathrm{R}^{m\times m}$ 为自表达系数矩阵; $e \in \mathrm{R}^{n\times 1}$ 是元素全为 1 的列向量; $b \in \mathrm{R}^{1\times m}$ 代表偏差项。

为了合理地利用样本间的关系,本文用 $l_{2,p}$-范数来估计余量,即 $\min\limits_{Z,b} \| X - XZ - eb \|_{2,p}$。同时,为了去除冗余和离群数据,添加一个正则化项 $R_1(Z)$,由于 $l_{2,p}$-范数能够有效地发现稀疏结构,使 Z 中一些不重要的行的系数变小或者直接为 0,这样有利于排除无关的属性,使算法更加有效,故本文采用 $l_{2,p}$-范数作为正则化项惩罚目标项,即 $R_1(Z) = \| Z \|_{2,p} = \left[\sum\limits_{i=1}^{m}\left(\sum\limits_{j=1}^{m} Z_{ij}^2\right)^{\frac{p}{2}}\right]^{\frac{1}{p}}$,其中 $0<p<2$。为保证样本数据在空间投影变换后数据的近邻关系保持稳定,本文借鉴了 LPP 算法,并且将其中的普通图改成超图,在之前提出的模型上嵌入一个超图拉普拉斯正则化项 $R_2(Z) = \lambda_2 \mathrm{tr}(Z^{\mathrm{T}} X^{\mathrm{T}} L_H XZ)$,这样不仅考虑了数据的局部信息,又考虑到了样本间更深层次的关系。进而可以获得如下函数:

$$\min_{Z,b} \| X - XZ - eb \|_{2,p} + \lambda_1 \| Z \|_{2,p} + \lambda_2 \mathrm{tr}(Z^{\mathrm{T}} X^{\mathrm{T}} L_H XZ) \tag{3}$$

其中: $L_H \in \mathrm{R}^{n\times n}$ 表示由数据构成的归一化超图拉普拉斯矩阵; λ_1 和 λ_2 为控制参数。

为了减少存储空间和进一步地提高分类效率,本文必须获取更加低秩的矩阵,为此,本文对 Z 进行低秩约束[17]。假设 $\mathrm{rank}(Z) = r, r \leqslant \min(m, n)$,一个低秩限制[17]的 Z 可以表示为: $Z = AB$,其中 $A \in \mathrm{R}^{m\times r}$,而 $B \in \mathrm{R}^{r\times m}$。综上所述,最终得到以下目标函数:

$$\min_{A,B,b} \| X - XAB - eb \|_{2,p} + \lambda_1 \| AB \|_{2,p} + \lambda_2 \mathrm{tr}(B^{\mathrm{T}} A^{\mathrm{T}} X^{\mathrm{T}} L_H XAB)$$
$$s.t. \quad rank(AB) \leqslant \min(m, n) \tag{4}$$

其中: AB 表示由 $A \in \mathrm{R}^{m\times r}$ 和 $B \in \mathrm{R}^{r\times m}$ 组成的自表达系数矩阵。低秩结构意味着, $\mathrm{rank}(AB) = r \leqslant \min(m, n)$。使用低秩限制能搜寻属性之间的隐含关系,并且已有文献证明,实际上低秩限制就是采用 LDA 方法进行子空间学习[18],而 LDA 能够利用数据的全局信息。通过稀疏学习使矩阵 AB 中大多数的行收缩为零,而 AB 中非零系数所对应的属性则被作为最具判别性的属性子集,以此帮助更好地进行属性选择。

本文提出的 HS_FS 算法具有以下优点:

a) 由于算法采用了属性自表达来建模,这种方式对噪声和离群点不太敏感,在建模的时候加入了一个偏差项使模型更加准确,接着,在含有偏差项的属性自表达模型中加入 $l_{2,p}$ 稀疏正则化因子,使该属性自表达模型能够自动地选择重要的属性,故相比子空间学习方法具有较好的解释性。

b) 通过在模型中嵌入一个超图拉普拉斯正则化项来考虑数据之间的局部信息,同时利用低秩表示来探寻数据的全局信息,所以该算法能够在很大程度上保持数据之间局部信息和全局信息的完整,从而使得该算法具有更强的分类能力。

c) 对本文的目标函数提出了一种有别于交替方向乘子法的优化求解方式,即对目标函数按顺序迭代执行低秩属性选择和子空间学习的优化求解算法,不断交替地迭代优化使得目标函数值在每次迭代过程中逐步收敛,最终取得全局最优解。

本文算法的伪代码如下:

算法 1 HS_FS 算法伪代码

输入: 训练样本 $X \in \mathrm{R}^{n\times m}$,参数 λ_1、λ_2、p。
输出: 分类准确率。
1 通过训练样本得出类指示矩阵;
2 根据文献[19]得到超图拉普拉斯矩阵 L_H;
3 依据所选择的模型调用算法 2 求解全局最优解,得到自表达系数矩阵 Z;
4 利用最优解 Z 对原始属性集 X 进行属性选择后得到的属性集作为样本新的属性集;
5 对新的属性集构成的样本采用 SVM 分类。

2.2 算法优化

由于式(4)中使用了 $l_{2,p}$-范数,所以目标函数无法在一个封闭的形式下求解。本文提出一种新的交替优化方法来解决该问题。具体地分为两步:

a) 固定 A、B,优化 b。

当固定 A、B 后,优化问题式(4)将变为

$$\min_b \| X - XAB - eb \|_{2,p} \tag{5}$$

式(5)可以看成是关于 b 的函数,所以对式(5)求导,并使导数为 0,可以得到

$$2e^{\mathrm{T}} eb + 2e^{\mathrm{T}} XAB - 2e^{\mathrm{T}} X = 0 \tag{6}$$

通过简单的数学转换便可以得到

$$b = \frac{1}{n} e^{\mathrm{T}} X - \frac{1}{n} e^{\mathrm{T}} XAB \tag{7}$$

b) 固定 b,优化 A 和 B。

当固定 b 之后,将式(7)代入(4)可以得到

$$\min_{A,B} \| X - XAB - e\left(\frac{1}{n} e^{\mathrm{T}} X - \frac{1}{n} e^{\mathrm{T}} XAB\right) \|_{2,p} + \lambda_1 \mathrm{tr} \| AB \|_{2,p} + \lambda_2 \mathrm{tr}(B^{\mathrm{T}} A^{\mathrm{T}} X^{\mathrm{T}} L_H XAB) \tag{8}$$

让 $H = I_n - \frac{1}{n} ee^{\mathrm{T}} \in \mathrm{R}^{n\times n}$,而 $I_n \in \mathrm{R}^{n\times n}$ 是一个单位矩阵,那么式(8)可以改写成

$$\min_{A,B} \| HX - HXAB \|_{2,p} + \lambda_1 \mathrm{tr} \| AB \|_{2,p} + \lambda_2 \mathrm{tr}(B^{\mathrm{T}} A^{\mathrm{T}} X^{\mathrm{T}} L_H XAB) \tag{9}$$

也就等于

$$\min_{A,B} \mathrm{tr}((HX - HXAB)^{\mathrm{T}} P(HX - HXAB)) + \lambda_1 \mathrm{tr}(B^{\mathrm{T}} A^{\mathrm{T}} QAB) + \lambda_2 \mathrm{tr}(B^{\mathrm{T}} A^{\mathrm{T}} X^{\mathrm{T}} L_H XAB) \tag{10}$$

其中: $P \in \mathrm{R}^{n\times n}$ 和 $Q \in \mathrm{R}^{m\times m}$ 都是对角矩阵,且 $P_{ii} = 1/(2/p) \| (HX - HXAB)^i \|_2^{2-p}$ $(i = 1, \cdots, n, 0<p<2)$, $Q_{jj} = 1/(2/p) \| (AB)^j \|_2^{2-p}$ $(j = 1, \cdots, m, 0<p<2)$ ($(AB)^j$ 表示矩阵 AB 的第 j 行,$(HX - HXAB)^i$ 同理)。

将式(10)看成是 B 的函数,并且令其对应的偏导数为 0,便能够得到以下式子:

$$B = (A^{\mathrm{T}}(X^{\mathrm{T}} H^{\mathrm{T}} PHX + \lambda_1 Q + \lambda_2 X^{\mathrm{T}} L_H X) A)^{-1} A^{\mathrm{T}} X^{\mathrm{T}} H^{\mathrm{T}} PHX \tag{11}$$

将式(11)代入式(10)可以得到

$$\max_A tr((A^{\mathrm{T}}(X^{\mathrm{T}} H^{\mathrm{T}} PHX + \lambda_1 Q + \lambda_2 X^{\mathrm{T}} L_H X) A)^{-1} A^{\mathrm{T}} X^{\mathrm{T}} H^{\mathrm{T}} PHXX^{\mathrm{T}} H^{\mathrm{T}} P^{\mathrm{T}} HXA) \tag{12}$$

可以发现

$$S_t = X^{\mathrm{T}} H^{\mathrm{T}} PHX + \lambda_1 Q + \lambda_2 X^{\mathrm{T}} LHX,$$
$$S_b = X^{\mathrm{T}} H^{\mathrm{T}} PHXX^{\mathrm{T}} H^{\mathrm{T}} P^{\mathrm{T}} HX \tag{13}$$

其中: S_t 和 S_b 分别表示 LDA 中的类内离散矩阵和类间离散度矩阵。因此,通过求解式(12)可以得到 A 如下:

$$A=\arg\max_{A}\{\mathrm{tr}((A^{\mathrm{T}}S_tA)^{-1}A^{\mathrm{T}}S_bA)\} \tag{14}$$

式(14)的求解问题明显是 LDA 求解问题,因此,上式对应的全局最优解是: 求 $S_t^{-1}S_b$ 中非零属性值所对应的属性向量。

综上所述,整理出算法 2 来求解,得到自表达系数矩阵 Z:

算法 2　优化求解式(4)的伪代码

输入: 训练样本 $X\in\mathrm{R}^{n\times m}$,控制参数 λ_1,λ_2,低秩参数 r,参数 p。

输出: 矩阵 $A\in\mathrm{R}^{m\times r}$ 和 $B\in\mathrm{R}^{r\times m}$;

1 初始化 $t=1$;

2 通过随机初始化矩阵 $A^{(t)}$ 和 $B^{(t)}$,得到初始化矩阵 $Z^{(t)}=A^{(t)}B^{(t)}$;

3 初始化 $P^{(t)}=I\in\mathrm{R}^{n\times n}$;

4 初始化 $Q^{(t)}=I\in\mathrm{R}^{m\times m}$;

5　重复:

5.1 通过式(14)计算 $A^{(t+1)}$;

5.2 通过式(11)计算 $B^{(t+1)}$;

5.3 通过式(7)计算 $b^{(t+1)}$;

5.4 更新对角矩阵 $P^{(t+1)}\in\mathrm{R}^{n\times n}$,第 i 个对角元素根据 $P_{ii}=1/(2/p)\|(HX-HXAB)^{i}\|_2^{2-p}(i=1,\cdots,n)$ 来计算;

5.5 更新对角矩阵 $Q^{(t+1)}\in\mathrm{R}^{m\times m}$,第 j 个对角元素根据 $Q_{jj}=\dfrac{1}{(\frac{2}{p})\|(AB)^{j}\|_2^{2-p}}(j=1,\cdots,m)$ 来计算;

5.6 更新 $t=t+1$;

6　直到式(4)收敛;

3　优化算法收敛性证明

本章将证明目标函数式(4)在每次的迭代过程中是单调递减的。注意到,目标函数式(4)实际上等价于

$$\begin{aligned}&\min_{A,B}\mathrm{tr}((HX-HXAB)^{\mathrm{T}}P(HX-HXAB))+\\&\lambda_1\mathrm{tr}(B^{\mathrm{T}}A^{\mathrm{T}}QAB)+\lambda_2\mathrm{tr}(B^{\mathrm{T}}A^{\mathrm{T}}X^{\mathrm{T}}L_HXAB)\end{aligned} \tag{15}$$

由此可得

$$\begin{aligned}&\mathrm{tr}\left[(HX-HXA^{(t+1)}B^{(t+1)})^{\mathrm{T}}P^{(t)}(HX-HXA^{(t+1)}B^{(t+1)})\right]+\\&\lambda_1\mathrm{tr}(B^{(t+1)\mathrm{T}}A^{(t+1)\mathrm{T}}Q^{(t)}A^{(t+1)}B^{(t+1)})+\\&\lambda_2\mathrm{tr}(B^{(t+1)\mathrm{T}}A^{(t+1)\mathrm{T}}X^{\mathrm{T}}LHXA^{(t+1)}B^{(t+1)})\leqslant\\&\mathrm{tr}\left[(HX-HXA^{(t)}B^{(t)})^{\mathrm{T}}P^{(t)}(HX-HXA^{(t)}B^{(t)})\right]+\\&\lambda_1\mathrm{tr}(B^{(t)\mathrm{T}}A^{(t)\mathrm{T}}Q^{(t)}A^{(t)}B^{(t)})+\\&\lambda_2\mathrm{tr}(B^{(t)\mathrm{T}}A^{(t)\mathrm{T}}X^{\mathrm{T}}LHXA^{(t)}B^{(t)})\end{aligned} \tag{16}$$

让 $D=HX-HXAB$,可以得到

$$\begin{aligned}&\sum_{i=1}^{n}\frac{\|D^{i(t+1)}\|_2^{2(2-p)}}{(\frac{2}{p})\|D^{i(t)}\|_2^{2-p}}+\lambda_1\sum_{j=1}^{m}\frac{\|Z^{j(t+1)}\|_2^{2(2-p)}}{(\frac{2}{p})\|Z^{j(t)}\|_2^{2-p}}+\\&\lambda_2\mathrm{tr}(Z^{(t+1)\mathrm{T}}X^{\mathrm{T}}LHXZ^{(t+1)})\leqslant\\&\sum_{i=1}^{n}\frac{\|D^{i(t)}\|_2^{2(2-p)}}{(\frac{2}{p})\|D^{i(t)}\|_2^{2-p}}+\lambda_1\sum_{j=1}^{m}\frac{\|Z^{j(t)}\|_2^{2(2-p)}}{(\frac{2}{p})\|Z^{j(t)}\|_2^{2-p}}+\\&\lambda_2\mathrm{tr}(Z^{(t)\mathrm{T}}X^{\mathrm{T}}LHXZ^{(t)})\end{aligned} \tag{17}$$

对于每个 i,能够得到

$$\begin{aligned}&\|D^{i(t+1)}\|_2^{2-p}-\frac{\|D^{i(t+1)}\|_2^{2(2-p)}}{(\frac{2}{p})\|D^{i(t)}\|_2^{2-p}}\leqslant\\&\|D^{i(t)}\|_2^{2-p}-\frac{\|D^{i(t)}\|_2^{2(2-p)}}{(\frac{2}{p})\|D^{i(t)}\|_2^{2-p}}\end{aligned} \tag{18}$$

而对于每个 j 可以得到

$$\begin{aligned}&\|Z^{j(t+1)}\|_2^{2-p}-\frac{\|Z^{j(t+1)}\|_2^{2(2-p)}}{(\frac{2}{p})\|Z^{j(t)}\|_2^{2-p}}\leqslant\\&\|Z^{j(t)}\|_2^{2-p}-\frac{\|Z^{j(t)}\|_2^{2(2-p)}}{(\frac{2}{p})\|Z^{j(t)}\|_2^{2-p}}\end{aligned} \tag{19}$$

然后,结合式(18)和(19)便可以得到

$$\begin{aligned}&\sum_{i=1}^{n}(\|D^{i(t+1)}\|_2)^{2-p}-\frac{\|D^{i(t+1)}\|_2^{2(2-p)}}{(\frac{2}{p})\|D^{i(t)}\|_2^{2-p}}+\\&\lambda_1\sum_{j=1}^{d}(\|Z^{j(t+1)}\|_2)^{2-p}-\frac{\|Z^{j(t+1)}\|_2^{2(2-p)}}{(\frac{2}{p})\|Z^{j(t)}\|_2^{2-p}}\leqslant\\&\sum_{i=1}^{n}(\|D^{i(t)}\|_2)^{2-p}-\frac{\|D^{i(t)}\|_2^{2(2-p)}}{(\frac{2}{p})\|D^{i(t)}\|_2^{2-p}}+\\&\lambda_1\sum_{j=1}^{d}(\|Z^{j(t)}\|_2)^{2-p}-\frac{\|Z^{j(t)}\|_2^{2(2-p)}}{(\frac{2}{p})\|Z^{j(t)}\|_2^{2-p}}\end{aligned} \tag{20}$$

综上所述,再结合式(7)和 $H=I_n-\dfrac{1}{n}ee^{\mathrm{T}}$ 可知

$$\begin{aligned}&\|X-XZ^{(t+1)}-eb\|_{2,p}+\lambda_1\|Z^{(t+1)}\|_{2,p}+\\&\lambda_2\mathrm{tr}(Z^{(t+1)\mathrm{T}}X^{\mathrm{T}}L_HXZ^{(t+1)})\leqslant\\&\|X-XZ^{(t)}-eb\|_{2,p}+\lambda_1\|Z^{(t)}\|_{2,p}+\\&\lambda_2\mathrm{tr}(Z^{(t)\mathrm{T}}X^{\mathrm{T}}LHXZ^{(t)})\end{aligned} \tag{21}$$

上述不等式表明算法 1 在每次的迭代中,目标函数都是单调递减的,所以本文提出的算法能够最终达到全局收敛的结果。

4　实验结果和分析

4.1　实验数据集和对比算法

本文在六个数据集上测试所提出的属性选择算法的性能,数据集 Ionosphere、Sonar、ecoli、Yeast、Movements 和 Arrhythmia 均来源于 UCI[20],数据集详情如表 1 所示。

表 1　数据集信息统计

数据集	样本数	属性数	类数
Ionosphere	351	34	2
Sonar	208	60	2
ecoli	336	343	8
Yeast	1 484	1 470	10
Arrhythmia	452	279	13
Movements	360	90	15

所有实验均在 Windows 7 系统下运行,使用 MATLAB2014a 软件进行测试。本文实验选择五种对比算法来与本文提出的算法进行比较,对于所有算法,均采用 5 折交叉验证方法来选择实验中所涉及到的所有参数,且 SVM 分类器中参数 $(c,g)\in[-10:2:10]$,同时,为了减少运行时间,SVM 分类器类型统一选择线性模式。TRACK 方法[21],通过结合迹比率形式和 K-means 聚类方法来搜索具有判别的属性; RSR 方法[22],通过自表征的方法选择一个最具代表性的响应矩阵,然后嵌入到稀疏学习模型中进行属性选择,同时,系数矩阵中的系数大小即表示对应属性重要性的强弱; CSFS[23],是一种通过 $l_{2,1}$-范数正则化来进行半监督属性选择的方法; FSR_ALM[24],利用 ALM(augmented lagrangian multiplier)来避免调参,同时使用 $l_{2,0}$-范数来提取 k 个最佳属性; LDA[14](linear discriminant analysis),是一种考虑数据全局结构的方法。

分析以上算法,TRACK 没有考虑数据的全局结构; RSR 没有考虑数据之间的相关性; CSFS 虽然通过 $l_{2,1}$-范数去除了冗余的属性,但是没有考虑数据的局部结构和属性之间的关系; FSR_ALM 仅仅关注了类标签和属性之间的关系; LDA 只是单纯地考虑数据之间的整体结构。

4.2　实验结果和分析

本文中参数设置如下: r 是由 m 和 n 约束得到,即: $r\leqslant\min(m,n)$,$\lambda_1,\lambda_2\in\{10^{-3},10^{-1},1,10^{1},10^{3}\}$,控制关联结构度的参数 p 取值为 $(0<p<2)$。本文采用分类准确率作为评价指标,分类准确率越高表示算法效果越好。实验利用 10 折交叉验证的

方法把原始数据划分为训练集和测试集(其中,9 份为训练集,1 份为测试集),然后运用 SVM 进行分类得到分类准确率。为了公平起见,所有算法均保证在同一实验环境下进行,最后提取 10 次运行的实验结果的均值加减均方差来评估各算法的性能,各算法在六个数据集上实验结果对比如图 1 ~6 所示,具体数据结果如表 2 所示。

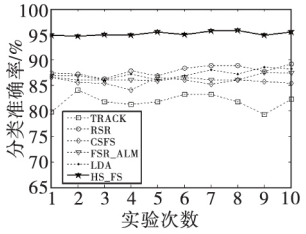

图1 数据集Ionosphere

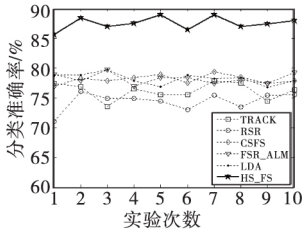

图2 数据集Sonar

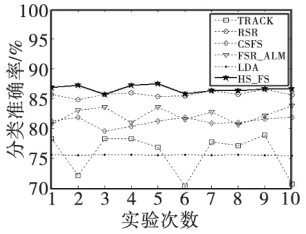

图3 数据集ecoli

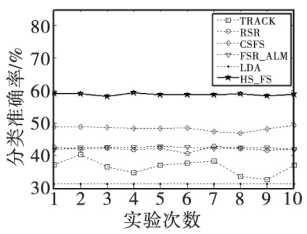

图4 数据集Yeast

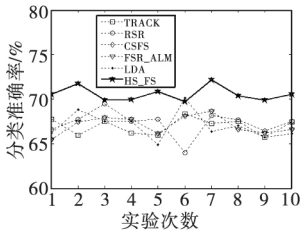

图5 数据集Arrhythmia

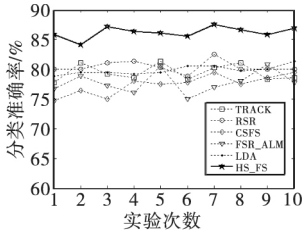

图6 数据集Movements

表 2 准确率(均值±均方差)统计结果

数据集	TRACK	RSR	CSFS	FSR_ALM	LDA	HS_FS
Ionosphere	81.82±1.43	87.88±0.90	85.67±0.71	86.53±0.59	87.09±0.91	95.13±0.39
Sonar	76.17±1.32	74.44±1.42	78.10±0.70	78.25±0.85	78.24±0.86	87.55±1.02
ecoli	75.85±3.24	85.74±0.48	81.13±0.72	82.28±1.19	75.55±0.08	86.61±0.58
Yeast	36.45±2.17	41.96±0.57	48.30±0.67	42.32±0.28	31.20±0.04	58.75±0.37
Arrhythmia	66.95±0.88	67.27±1.36	67.07±0.96	67.47±0.96	67.01±1.47	70.56±0.77
Movements	79.47±1.29	80.42±0.95	77.44±1.56	77.81±1.76	79.86±0.73	86.22±0.91
平均	69.45±1.72	72.95±0.95	72.95±0.89	72.44±0.94	69.83±0.68	80.80±0.67

通过图 1 ~6 可以清楚地看到 HS_FS 算法在六个数据集上的分类准确率,由于 10 折交叉验证的随机性,并不能保证每次的结果都是最好的,但是每个数据集上 10 次实验结果大多高于对比算法,最终的平均分类准确率也是最高的。

对于多类的数据集,如 Yeast 和 Movements 数据集,可以通过控制低秩的数量来得到不同的分类结果,通过图 7 和 8 所示,具有低秩结构的 HS_FS 算法比满秩的效果更佳,分类准确率更高。

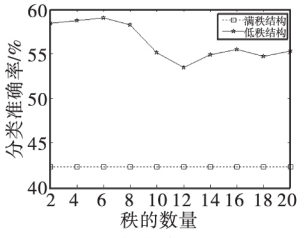

图7 数据集Yeast

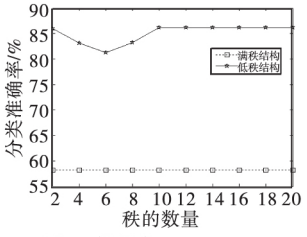

图8 数据集Movements

通过分析表 2 可以看出,HS_FS 与其他四种对比算法比较,均取得了最高的分类准确率,具体地,与 TRACK 算法比较平均提高了 11.35%;比 LDA 算法比较平均提高了 10.97%,证明了本文提出的算法比单一的子空间学习算法性能要好;与 FSR_ALM 算法相比,准确率平均提升了 8.36%;比 RSR 和 CSFS 算法比较平均提高了 7.85%。在 Yeast 数据集上,HS_FS 算法的效果最好,不但比 LDA 高出了 27.55%,而且与 CSFS 算法相比也提高了 10.45%,主要与以下两点有关:a) 考虑了两种数据的内在关系(全局结构和局部结构);b) 迭代的执行低秩属性选择步骤和子空间学习步骤(不断调整属性选择的结果,使其达到最优)。

虽然不同类型的数据集有不同的数据分布,且可能含有不同的干扰因素,但实验结果表明,本文提出的 HS_FS 算法在每个数据集上都取得了最高的平均准确率,并且对于所有对比算法,HS_FS 算法展现了最高的鲁棒性,同时也说明 HS_FS 算法输出了更具判别力的属性集。

5 结束语

本文结合属性层面和样本层面的关系提出了一种新的无监督属性选择算法——HS_FS 算法,即利用 $l_{2,p}$-范数,同时应用在损失函数和正则化两个部分,并结合属性自表达、超图结构、低秩约束不同结构方法来进一步完善提出的模型,因此可以达到较好的效果。该算法融合了属性自表达和子空间学习的优点,在一定程度上弥补了属性自表达在保持数据几何结构方面的不足。经实验证实,本文算法在分类任务的准确率和稳定性上均取得了较大的提升。在今后的工作中,本文将尝试在半监督属性选择方面拓展验证本文提出的算法,并尝试使用更先进的技术改进算法。

参考文献:

[1] Zhu Xiaofeng, Huang Zi, Shen Hengtao, *et al*. Dimensionality reduction by mixed kernel canonical correlation analysis [J]. Pattern Recognition, 2012, 45(8): 3003-3016.

[2] Zhu Xiaofeng, Zhang Shichao, Jin Zhi, *et al*. Missing value estimation for mixed-attribute data sets [J]. IEEE Trans on Knowledge and Data Engineering, 2011, 23(1): 110-121.

[3] Zhang Shichao. Shell-neighbor method and its application in missing data imputation [J]. Applied Intelligence, 2011, 35(1): 123-133.

[4] Zhu Xiaofeng, Li Xuelong, Zhang Shichao, *et al*. Robust joint graph sparse coding for unsupervised spectral feature selection [J]. IEEE Trans on Neural Networks & Learning Systems, 2017, 28(6): 1263-1275.

[5] Gheyas I A, Smith L S. Feature subset selection in large dimensionality domains [J]. Pattern Recognition, 2010, 43(1): 5-13.

[6] Tabakhi S, Moradi P, Akhlaghian F. An unsupervised feature selection algorithm based on ant colony optimization [J]. Engineering Applications of Artificial Intelligence, 2014, 32(6): 112-123.

[7] Zhang Shichao, Jin Zhi, Zhu Xiaofeng. Missing data imputation by utilizing information within incomplete instances [J]. Journal of Systems & Software, 2011, 84(3): 452-459.

[8] Unler A, Murat A, Chinnam R B. mr2PSO: a maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification [J]. Information Sciences, 2011, 181(20): 4625-4641.

[9] Zhang Shichao, Li Xuelong, Zong Ming, *et al*. Learning k for kNN classification [J]. ACM Trans on Intelligent Systems & Technology, 2017, 8(3): 43.

[10] Zhu Xiaofeng, Suk H, Wang Li, *et al*. A novel relational regularization feature selection method for joint regression and classification in AD diagnosis [J]. Medical Image Analysis, 2017, 38: 205-214.

[11] Shi Xiaoshuang, Guo Zhenhua, Lai Zhihui, *et al*. A framework of joint graph embedding and sparse regression for dimensionality reduction [J]. IEEE Trans on Image Processing, 2015, 24(4): 1341-1355.

(下转第 3219 页)

度,再采取措施使温度 c 满足要求,即可保证故障概率不变。同理,P_α 分别为 20%、30%、40%、50%,所得到的对于温度的控制线分别如图 1(c) ~(f) 所示。得到的元件最长使用时间分别为 1.211 d、1.936 d、2.773 d、3.763 d,即元件的更换周期。相应的维持可靠性的温度控制线(函数) 分别如式(6) ~(10) 所示。

$$c(T)=\frac{20\arccos\left(1-\frac{1.6}{\mathrm{e}^{-0.1842\times\mathrm{mod}(T/1.211)}}\right)}{\pi} \tag{6}$$

$$c(T)=\frac{20\arccos\left(1-\frac{1.4}{\mathrm{e}^{-0.1842\times\mathrm{mod}(T/1.936)}}\right)}{\pi} \tag{7}$$

$$c(T)=\frac{20\arccos\left(1-\frac{1.2}{\mathrm{e}^{-0.1842\times\mathrm{mod}(T/2.773)}}\right)}{\pi} \tag{8}$$

$$c(T)=\frac{20\arccos\left(1-\frac{1.0}{\mathrm{e}^{-0.1842\times\mathrm{mod}(T/3.763)}}\right)}{\pi} \tag{9}$$

上例分析过程作了一些简化,如只使用了一个元件进行分析,只有两个因素参与分析。在对系统进行分析过程中,可使用文献[10]提供的该系统故障概率分布曲面进行分析。该系统的故障概率分布曲面形式比较复杂,可通过上述方法找到连续的温度控制线,但周期性和函数性没有图中变化明显。另外,上述过程只有一个可控因素和不可控因素,是简单情况,更为复杂的是加入多个可控因素。这些可控因素所表示的空间维度可通过曲面投影\曲面相交或超曲面处理方法归结到三维空间,再进行投影变为二维形式。所以上述方法可以向复杂情况推广。分析所得结果的正确性。根据文献[6]的分析,其得到了系统维持故障概率而限定的工作区域。该工作区域边缘与本文得到的温度控制线类似。但文献[6]中使用方法更为复杂,且没有得到明确的温度与时间的函数关系,这不利于实际操作。本文方法简便,且可得到温控函数,有利于实际操作。另外,本文方法得到的元件更换周期比文献[6]得到的更换周期长,更有利于减少由于元件更换造成的损失。

综上,在 SFT 中将可控因素表示为不可控因素时间的函数,可对故障概率分布进行降维,进而按照故障概率要求维持元件或系统的可靠性稳定。

4　结束语

从维持系统或元件可靠性稳定的角度出发,研究了在 SFT 框架下维持可靠性的方法。

a) 本文提出了可控因素和不可控因素的概念。不可控因素指不能控制或不便控制的因素; 可控因素指可通过技术手段进行控制的因素。在 SFT 中,不可控因素为使用时间,其余均为可控因素。从物理意义上讲,时间是表征事物存在的唯一因素,所以任何可控因素都可以表示为不可控因素时间的函数。

b) 在 SFT 下构造不可控因素表示可控因素的函数需要限定条件,即故障概率的限定。因为通过曲面投影,曲面的相交或超曲面方法最终形成三维空间曲面,其维度分别是不可控因素、可控因素和故障概率。只有在限定故障概率时才可求得可控因素与不可控因素之间的函数关系。

c) 给出了系统或元件可靠性维持方法的步骤,并列举一例进行分析。其结果可以得到直观的温度控制线和精确的温度与时间函数。与已有文献相比,所得元件更换周期更为经济,方法更具有可操作性。

参考文献:

[1] 李家珏,朱钰,邵宝珠,等. 含风能电力系统的充裕性鲁棒决策机制设计[J]. 中国电机工程学报,2016,36(15):4090-4098.

[2] 赵渊,周念成,谢开贵. 计及 FACTS 元件的发输电系统可靠性评估模型[J]. 重庆大学学报: 自然科学版,2006,29(5):19-23.

[3] 王思华,赵琼,马军党,等. 接触网系统可靠性的马尔可夫建模分析[J]. 中国安全科学学报,2013,23(9):39-44.

[4] 徐长航,陈习,朱渊,等. 基于 GO 法的 LNG 接收站气化外输系统可靠性分析[J]. 中国安全科学学报,2013,23(1):61-66.

[5] 于敏,何正友,钱清泉. 基于 SHRN 的地铁综合监控系统可靠性分析[J]. 铁道学报,2012,34(2):70-79.

[6] 崔铁军,马云东. 基于多维空间事故树的维持系统可靠性方法研究[J]. 系统科学与数学,2014,34(6):682-692.

[7] 崔铁军,马云东. 状态迁移下系统适应性改造成本研究[J]. 数学的实践与认识,2015,45(24):136-142.

[8] 崔铁军,马云东. 基于 SFT 和 DFT 的系统维修率确定及优化[J]. 数学的实践与认识,2015,45(22):140-150.

[9] 崔铁军,马云东. 基于空间故障树理论的系统故障定位方法研究[J]. 数学的实践与认识,2015,45(21):135-142.

[10] 崔铁军,马云东. 多维空间故障树构建及应用研究[J]. 中国安全科学学报,2013,23(4):32-37.

[11] 崔铁军,马云东. 连续型空间故障树中因素重要度分布的定义与认知[J]. 中国安全科学学报,2015,25(3):24-28.

[12] 崔铁军,马云东. 离散型空间故障树构建及其性质研究[J]. 系统科学与数学,2016,36(10):1753-1761.

(上接第 3216 页)

[12] Zhu Xiaofeng, Huang Zi, Yang Yang, *et al*. Self-taught dimensionality reduction on the high-dimensional small-sized data [J]. Pattern Recognition, 2013, 46(1): 215-229.

[13] Pyatykh S, Hesser J, Zheng Lei. Image noise level estimation by principal component analysis [J]. IEEE Trans on Image Processing, 2013, 22(2): 687-699.

[14] Zhu Xiaofeng, Suk H, Shen Dinggang. A novel matrix-similarity based loss function for joint regression and classification in AD diagnosis [J]. Neuroimage, 2014, 100: 91-105.

[15] Zhu Xiaofeng, Zhang Lei, Huang Zi. A sparse embedding and least variance encoding approach to hashing [J]. IEEE Trans on Image Processing, 2014, 23(9): 3737-3750.

[16] Wang Tao, Qin Zhenxing, Zhang Shichao, *et al*. Cost-sensitive classification with inadequate labeled data [J]. Information Systems, 2012, 37(5): 508-516.

[17] Zhang Shichao. Decision tree classifiers sensitive to heterogeneous costs [J]. Journal of Systems & Software, 2012, 85(4): 771-779.

[18] Zhu Xiaofeng, Suk H, Lee S W, *et al*. Subspace regularized sparse multitask learning for multiclass neurodegenerative disease identification [J]. IEEE Trans on Biomedical Engineering, 2016, 63(3): 607-618.

[19] Zhu Xiaofeng, Li Xuelong, Zhang Shichao. Block-row sparse multiview multilabel learning for image classification [J]. IEEE Trans on Cybernetics, 2016, 46(2): 450-461.

[20] UCI repository of machine learning datasets [EB/OL]. [2016-05-27]. http://archive.ics.uci.edu/ml/.

[21] Wang De, Nie Feiping, Huang Heng. Unsupervised feature selection via unified trace ratio formulation and K-means clustering (TRACK) [C]//Proc of Joint European Conference on Machine Learning and Knowledge Discovery in Databases. 2014: 306-321.

[22] Zhu Pengfei, Zuo Wangmeng, Zhang Lei, *et al*. Unsupervised feature selection by regularized self-representation [J]. Pattern Recognition, 2015, 48(2): 438-446.

[23] Chang Xiaojun, Nie Feiping, Yang Yi, *et al*. A convex formulation for semi-supervised multi-label feature selection [C]//Proc of the 28th AAAI Conference on Artificial Intelligence. [S. l.]: AAAI Press, 2014: 1171-1177.

[24] Cai Xiao, Nie Feiping, Huang Heng. Exact top-k feature selection via $l_{2,0}$-norm constraint [C]//Proc of the 23rd International Joint Conference on Artificial Intelligence. [S. l.]: AAAI Press, 2013: 1240-1246.