·先进计算与数据处理· 文章编号: 1000-3428(2017)09-0043-08 文献标志码: A 中图分类号: TP181

鲁棒自表达的低秩属性选择算法

胡荣耀[1],刘星毅[2],程德波[1],何 威[1],罗 奘[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.广西钦州学院,广西 钦州 535000)

摘 要: 针对无监督属性选择算法无类别信息和未考虑属性的低秩问题,提出一种基于自表达方法的低秩属性选择算法。在损失函数中使用低秩和自表达方法描述属性间的相关结构,利用K均值聚类算法得到所有样本的伪类标签进行属性选择,采用稀疏学习方法中的$l_{2,p}$-范数参数p控制属性选择结果的稀疏性,并通过子空间学习方法使属性选择结果达到全局最优。实验结果表明,与无监督属性选择算法相比,该算法在6个公开数据集上均具有较高的分类准确率及稳定性。

关键词: 属性选择; 子空间学习; K均值聚类; 低秩约束; 稀疏学习

中文引用格式: 胡荣耀,刘星毅,程德波,等. 鲁棒自表达的低秩属性选择算法[J]. 计算机工程,2017,43(9):43-50.

英文引用格式: HU Rongyao,LIU Xingyi,CHENG Debo,et al. Robust Low-rank Self-representation Feature Selection Algorithm[J]. Computer Engineering,2017,43(9):43-50.

Robust Low-rank Self-representation Feature Selection Algorithm

HU Rongyao[1],LIU Xingyi[2],CHENG Debo[1],HE Wei[1],LUO Yan[1]

(1. Guangxi Key Lab of Multi-source Information Mining and Security,Guangxi Normal University,Guilin,Guangxi 541004,China; 2. Qinzhou University,Qinzhou,Guangxi 535000,China)

【Abstract】 Since unsupervised feature selection algorithms do not have label information and also ignore the low-rank characteristics of the data,this paper proposes a new low-rank feature selection algorithm based on self-representation method. In the loss function,low rank and self-representation methods are used to describe the correlation structure between features,and the K-means clustering method is used to obtain the pseudo labels of samples to realize feature selection. Then,$l_{2,p}$-norm parameter p in sparse learning method is adopted to control the sparsity of feature selection results. Through subspace learning method,the result of feature selection is globally optimal. The experimental results on six public datasets demonstrate that the proposed feature selection algorithm has higher classification accuracy and better stability compared with the unsupervised feature selection algorithm.

【Key words】 feature selection; subspace learning; K-means clustering; low-rank constraint; sparse learning

DOI: 10.3969/j.issn.1000-3428.2017.09.009

0 概述

高维特征表示物体(或样本)能准确且多样化地表现物体的特性。因此,高维度数据经常在机器学习及相关领域中存在,如文本数据、人脸数据和生物信息数据等[1]。但是,高维数据的使用不仅增大了数据存储空间及增加了运算时间的复杂度,而且在对数据处理过程中容易造成维灾难等现象。因此,对高维数据进行属性约简以降低数据的维度,从而寻找到具有代表性的低维属性子集成为机器学习领域的研究热点[2-3]。

属性约简括能降低处理时间,得到更具有泛化能力和更紧实的学习模型等[4-5]。常见的属性约简方法包含子空间学习和属性选择2种方法[6-7]。子空间学习是把高维属性从高维度空间投影到低维空间,并保持数据间的关联结构,常见的方法有线性判别分析(Linear Discriminant Analysis,LDA)[8]、邻域保持嵌入(Neighborhood Preserving Embedding,NPE)[9]等。属性选择是依据某种评估准则,选取一个最优属性子集来代表整个原始属性集,常见的方法有Lasso回归[10]和偏最小二乘回归[11]等。早期的属性选择通过评价函数计算每个属性与类别之间的相互关系得分,然后依据得分进行属性选择。现有方法[12-14]是利用属性间相关性得到一个响应矩阵,把属性选择问题转

基金项目:国家自然科学基金(61263035,61573270);中国博士后科学基金(2015M570837);广西自然科学基金(2015GXNSFCB139011);广西研究生教育创新计划项目(YCSZ2016046)。

作者简介:胡荣耀(1992—),男,硕士,主研方向为数据挖掘、机器学习;刘星毅,副教授、硕士;程德波、何 威、罗 奘,硕士。

收稿日期:2016-06-28 修回日期:2016-08-11 E-mail:610273948@qq.com

化为一个简单的多输出回归问题。

现有研究表明,有监督属性选择方法由于利用类标签,因此效果要好于无监督属性选择方法。然而,无监督属性选择面对的是无标签学习,因此,在无监督属性选择中利用类别信息(或者伪类别信息)对现有方法具有较大的挑战性。为此本文结合聚类方法、子空间学习和属性选择各自的优点,提出一种高效的属性选择算法,称为鲁棒自表达的低秩属性选择算法(Robust Low-rank Self-representation Feature Selection Algorithm,RS_FS)。本文提出的 RS_FS 算法在无监督中运用伪类别信息进行属性选择,考虑数据低秩特性,并在属性选择框架中嵌入子空间学习等方式进一步改进属性选择效果。

1 鲁棒自表达的低秩属性选择算法优化

1.1 基础模型

假设给定训练集 $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n]\in\mathbb{R}^{n\times d}$,其中,$n$ 和 d 分别表示样本数和属性数,同时,给定类标签矩阵 $\boldsymbol{Y}\in\mathbb{R}^{n\times k}$,其中,$k$ 表示样本类别数。因此,通常对于有监督的分类问题,采用的线性回归模型如下:

$$\min_{\boldsymbol{Z}}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{Z}\|_{\mathrm{F}}^{2} \tag{1}$$

其中,$\boldsymbol{Z}\in\mathbb{R}^{d\times k}$ 表示回归系数矩阵;$\|\cdot\|_{\mathrm{F}}$ 表示 Frobenius 范数,$\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{Z}\|_{\mathrm{F}}^{2}$ 为重构误差。由于式(1)是凸函数,因此易得其最优解为:$\boldsymbol{Z}=(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$。通常任何数据集都伴有不必要且冗余的属性,即数据集矩阵不是满秩就是低秩的。为了更好地突显具有代表性的属性和揭示数据在空间分布中的结构信息和类别信息,本文需要引入低秩来满足非满秩的实际需求。式(1)关于多类问题,对每个类的响应变量分别求解,未考虑到每个类的响应变量之间的相关性。为此,算法引入低秩结构来约束系数矩阵 $\boldsymbol{Z}$,具体如下:

$$\mathrm{rank}(\boldsymbol{Z})=r,r\leqslant\min(d,k) \tag{2}$$

考虑到无监督分类选取一个适当的响应矩阵比较困难。为此,算法在式(2)低秩约束结构的基础上,利用属性自表达的方法把 $\boldsymbol{X}$ 作为响应矩阵,所以,转化为解决如下问题:

$$\begin{aligned}&\min_{\boldsymbol{Z}}\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{Z}\|_{\mathrm{F}}^{2}\\&\text{s.t.}\quad\mathrm{rank}(\boldsymbol{Z})\leqslant\min(d,k)\end{aligned} \tag{3}$$

1.2 低秩属性选择算法的提出

现有研究显示,有监督分类由于包含类标签信息,通常比无监督分类具有更好的效果[15-16]。然而,通常类标签的获得难度较大[17]。为此,对无监督分类进行研究,并且为获得较好的分类效果,本文算法引入伪类标签来最小化类内距离和最大化类间距离。本文目的是采用伪类标签把无监督分类转为有监督分类,试图得到更好的分类效果。因此,本文算法引入 K 均值聚类算法,使得所有样本都能分到一个伪类标签,目的是在最大化类间距离的同时增强分类效果。具体地,给定 c 个类中心,变换后的聚类中心矩阵 $\boldsymbol{G}=[\boldsymbol{g}_1,\boldsymbol{g}_2,\cdots,\boldsymbol{g}_c]\in\mathbb{R}^{d\times c}$,引入 K 均值聚类算法,最小化如下目标函数:

$$\sum_{i=1}^{c}\sum_{\boldsymbol{y}_j\in Y_i}(\boldsymbol{y}_j-\boldsymbol{g}_i)^{\mathrm{T}}(\boldsymbol{y}_j-\boldsymbol{g}_i)=\sum_{i=1}^{n}(\boldsymbol{y}_i-\boldsymbol{u}_i\boldsymbol{G}^{\mathrm{T}})^{\mathrm{T}}(\boldsymbol{y}_i-\boldsymbol{u}_i\boldsymbol{G}^{\mathrm{T}}) \tag{4}$$

其中,$\boldsymbol{u}_i=[u_{i1},u_{i2},\cdots,u_{ic}]$ 是通过聚类指示矩阵 $\boldsymbol{U}=[\boldsymbol{u}_1^{\mathrm{T}},\boldsymbol{u}_2^{\mathrm{T}},\cdots,\boldsymbol{u}_n^{\mathrm{T}}]^{\mathrm{T}}\in\mathbb{R}^{n\times c}$ 转化后的 $\boldsymbol{x}_i$ 集合。由于 $\boldsymbol{y}_i=\boldsymbol{x}_i\boldsymbol{Z}$,因此通过 $\boldsymbol{Z}$ 的线性转化后可以得到类内最小间距。将式(4)代入到式(3)中可以得到:

$$\begin{aligned}&\min_{\boldsymbol{Z}}\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{Z}\|_{\mathrm{F}}^{2}+\alpha\sum_{i=1}^{n}(\boldsymbol{x}_i\boldsymbol{Z}-\boldsymbol{u}_i\boldsymbol{G}^{\mathrm{T}})^{\mathrm{T}}(\boldsymbol{x}_i\boldsymbol{Z}-\boldsymbol{u}_i\boldsymbol{G}^{\mathrm{T}})\\&\text{s.t.}\,\mathrm{rank}(\boldsymbol{Z})\leqslant\min(d,k)\end{aligned} \tag{5}$$

其中,回归系数矩阵 $\boldsymbol{Z}\in\mathbb{R}^{d\times k}$,具有低秩约束结构。式(5)相当于一个有监督分类模型。

为了使模型能更好地应用于属性选择,同时,考虑到 l_F-范数对于噪声数据和离群数据比较敏感,而一个离群数据样本只对应原始数据矩阵 $\boldsymbol{X}$ 的一行,所以,在模型中嵌入 $l_{2,p}$-范数作为正则化项,并通过 $p(0<p<2)$ 的控制获得具有稀疏结构的系数矩阵,去除噪声属性和冗余属性的影响。新模型如下:

$$\begin{aligned}&\min_{\boldsymbol{Z}}\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{Z}\|_{\mathrm{F}}^{2}+\beta\|\boldsymbol{Z}\|_{2,p}\\&\qquad+\alpha\sum_{i=1}^{n}(\boldsymbol{x}_i\boldsymbol{Z}-\boldsymbol{u}_i\boldsymbol{G}^{\mathrm{T}})^{\mathrm{T}}(\boldsymbol{x}_i\boldsymbol{Z}-\boldsymbol{u}_i\boldsymbol{G}^{\mathrm{T}})\\&\mathrm{rank}(\boldsymbol{Z})=r\leqslant\min(d,k)\end{aligned} \tag{6}$$

其中,β 为正参数,用来调节正则化项对模型的惩罚。通过调节参数,可以获得稀疏结构化表示的系数矩阵 $\boldsymbol{Z}$,令不相关或者弱相关属性在系数矩阵 $\boldsymbol{Z}$ 中均收缩为 0,而在系数矩阵 $\boldsymbol{Z}$ 中越重要的属性具有越大的系数权重。因此,算法可以选择出更具有代表性的属性子集,通过新构成的属性集进行分类以获得更好的分类效果。

一般地,秩最小化问题是一个非凸状且 NP 难的问题,但秩约束的目标函数是可解的,即能够得到全局解[18]。为更好地优化低秩结构并且考虑类别间的关联结构,算法令 $\boldsymbol{Z}=\boldsymbol{A}\boldsymbol{B}$ 拥有一个低秩 r,最终得到的算法目标函数为:

$$\begin{aligned}&\min_{\boldsymbol{A},\boldsymbol{B},\boldsymbol{U},\boldsymbol{G}}\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{A}\boldsymbol{B}\|_{\mathrm{F}}^{2}+\alpha\|\boldsymbol{X}\boldsymbol{A}\boldsymbol{B}-\boldsymbol{U}\boldsymbol{G}^{\mathrm{T}}\|_{\mathrm{F}}^{2}\\&\qquad+\beta\|\boldsymbol{A}\boldsymbol{B}\|_{2,p}\\&\text{s.t.}\,\mathrm{rank}(\boldsymbol{A}\boldsymbol{B})\leqslant\min(m,k)\end{aligned} \tag{7}$$

其中,$\boldsymbol{X}\in\mathbb{R}^{n\times d}$;$\boldsymbol{A}\in\mathbb{R}^{d\times r}$;$\boldsymbol{B}\in\mathbb{R}^{r\times d}$;重构的回归系数矩阵 $\boldsymbol{Z}$、矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 同样具有低秩结构;$\boldsymbol{U}$ 表示聚类指示矩阵;$\boldsymbol{G}$ 表示聚类中心矩阵;$\|\cdot\|_{\mathrm{F}}$ 表示矩阵的 Frobenius 范数。

算法 1 RS_FS 算法

输入 训练数据集 $\boldsymbol{X}\in\mathbb{R}^{n\times d}$,正则化参数 α,β,SVM 分类器中的参数 $(c,g)\in[-10:2:10]$,分类器类型选择线性模式

输出 分类准确率

1) 通过自表达方法得到自表达矩阵 X。

2) 依据所选择的式(7),求解全局最优解得到自表征矩阵 $Z^* = A^* B^*$。

3) 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。

4) 对新的属性集构成的样本进行 SVM 分类。

目标函数执行了一个 LDA 算法的过程,同时算法引入伪类标签进行有监督属性选择,所以,本文算法有效利用了属性间的线性判别能力。

1.3　低秩属性选择算法的优化

自表达的低秩属性选择算法的优化过程具体如下:

首先,将式(7)化简如下:

$$\min_{A,B,U,G} \| X - XAB \|_F^2 + \alpha \| XAB - UG^T \|_F^2 + \beta \mathrm{tr}(B^T A^T DAB) \tag{8}$$

其中,$D \in \mathbb{R}^{d \times d}$ 是一个对角矩阵,D 中的每个元素定义如下:

$$d_{ii} = \frac{1}{(2/p)(\| z^i \|_2)^{2-p}}$$
$$\text{s.t. } i = 1,2,\cdots,d;\ 0 < p < 2 \tag{9}$$

其中,z^i 表示矩阵 $Z^* = A^* B^*$ 的第 i 行;A^* 和 B^* 分别为 A 和 B 的最优解;Z^* 为重构系数矩阵 Z 的最优解。

然后,分 3 步来优化目标函数,具体为:

1) 固定矩阵 A,B,G,优化矩阵 U。当矩阵 A,B 固定后,则式(8)的第 1 项和第 3 项就变为常数,而第 2 项就可以被作为 K 均值的目标函数,即可以对原始矩阵 X 中的每个元素都分配一个聚类标签,同时,聚类中心矩阵 $G = [g_1, g_2, \cdots, g_c]$ 固定以后,可以得出优化的 U。

$$U_{ij} = \begin{cases} 1, j = \arg\min_k \| x_i AB - g_k^T \|_F^2 \\ 0, \text{其他} \end{cases} \tag{10}$$

式(10)表明通过 XAB 转换原始数据后,直接进行 K 均值算法,可使得转换后的每个数据都分配到一个标签。

2) 固定矩阵 U,A,B,优化矩阵 G。当固定聚类指示矩阵 U 以及矩阵 A 和 B 后,再对式(8)中的 G 求偏导数,并令所得偏导数等于 0。

$$\begin{aligned} &\alpha \frac{\partial \mathrm{tr}(XAB - UG^T)^T (XAB - UG^T)}{\partial G} = 0 \\ &\Rightarrow 2\alpha GU^T UG^T - 2\alpha GU^T XAB = 0 \\ &\Rightarrow G^T = [U^T U)^{-1}]^T U^T XAB \\ &\Rightarrow G = B^T A^T X^T U (U^T U)^{-1} \end{aligned} \tag{11}$$

3) 固定矩阵 U 和 G,优化矩阵 A 和 B。固定聚类指示矩阵 U 和聚类中心矩阵 G,将式(11)带入到式(8)中,并令式(8)等于 L,得到:

$$\begin{aligned} L &= \| X - XAB \|_F^2 + \beta Tr(B^T A^T DAB) \\ &\quad + \alpha \| XAB - U [U^T U)^{-1}]^T U^T XAB \|_F^2 \\ &= \alpha \mathrm{tr}[XAB - U [U^T U)^{-1}]^T U^T XAB)^T \\ &\quad \times (XAB - [U^T U)^{-1}]^T U^T XAB)] \\ &\quad + \mathrm{tr}[(X - XAB)^T (X - XAB)] + \beta \mathrm{tr}(B^T A^T DAB) \\ &= \alpha \mathrm{tr}(B^T A^T X^T XAB - 2B^T A^T X^T U [U^T U)^{-1}]^T U^T XAB \\ &\quad + B^T A^T X^T U (U^T U)^{-1} U^T U [U^T U)^{-1}]^T U^T XAB) \\ &\quad + \mathrm{tr}(B^T A^T X^T XAB - 2B^T A^T X^T X) + \beta \mathrm{tr}(B^T A^T DAB) \\ &= \mathrm{tr}(B^T A^T X^T XAB + \alpha B^T A^T X^T XAB - \alpha B^T A^T X^T \\ &\quad \times U [U^T U)^{-1}]^T U^T XAB + \beta B^T A^T DAB - 2B^T A^T X^T X) \\ &= \mathrm{tr}(B^T A^T (X^T X + \alpha X^T X - \alpha X^T U [U^T U)^{-1}]^T \\ &\quad \times U^T X + \beta D) AB - 2B^T A^T X^T X) \end{aligned} \tag{12}$$

通过上述化简,目标函数式(7)可转化为:

$$\begin{aligned} &\min_{A,B} \mathrm{tr}(B^T A^T (X^T X + \alpha X^T X - \alpha X^T U [(U^T U)^{-1}]^T \\ &\quad \times U^T X + \beta D) AB - 2B^T A^T X^T X) \\ &\text{s.t. } \mathrm{rank}(AB) = r \leqslant \min(m, k) \end{aligned} \tag{13}$$

其中,令 $M = X^T X - \alpha X^T U [(U^T U)^{-1}]^T U^T X + \alpha X^T X + \beta D$。

定理 1　固定 U 和 G 优化 A 和 B 的过程就是对伪类标签执行 LDA 的过程。

在固定聚类指示矩阵 U 和聚类中心矩阵 G 后,为更新 B,对式(13)中的 B 求偏导数,并令所得偏导数为 0。

$$\frac{\partial L}{\partial B} = 0 \Rightarrow B = (A^T MA)^{-1} A^T X^T X \tag{14}$$

由式(14)可以得出,$B^T = X^T XA [A^T MA)^{-1}]^T$,然后,把得到的矩阵 B 和 B^T 的表达式带入到式(13)中得到:

$$\begin{aligned} L &= \mathrm{tr}(B^T A^T MAB - 2B^T A^T X^T X) \\ &= \mathrm{tr}(X^T XA [A^T MA)^{-1}]^T \times A^T MA \times (A^T MA)^{-1} \\ &\quad \times A^T ABX^T X - 2X^T XA [A^T MA)^{-1}]^T A^T X^T X) \\ &= \mathrm{tr}(X^T XA [A^T MA)^{-1}]^T A^T X^T X \\ &\quad - 2X^T XA [A^T MA)^{-1}]^T \times A^T X^T X) \\ &= -\mathrm{tr}(X^T XA [A^T MA)^{-1}]^T \times A^T X^T X) \\ &= -\mathrm{tr}(X^T XA (A^T MA)^{-1} A^T X^T X) \\ &= -\mathrm{tr}\{(A^T MA)^{-1} A^T X^T XX^T XA\} \\ &= -\mathrm{tr}(A^T MA)^{-1} A^T X^T XX^T XA \end{aligned}$$

目标函数 L 只包含变量 A,因此,优化问题转化为:

$$\begin{aligned} &\max_A \mathrm{tr}\{(A^T MA)^{-1} A^T X^T XX^T XA\} \\ &= \max_A \mathrm{tr}\{(A^T (X^T X + \alpha X^T X - \alpha X^T U [U^T U)^{-1}]^T \\ &\quad \times U^T X + \beta D) A)^{-1} A^T X^T XX^T XA\} \end{aligned} \tag{15}$$

其中:

$$\begin{aligned} S_t &= \alpha X^T X - \alpha X^T U [U^T U)^{-1}]^T U^T X + X^T X + \beta D \\ S_b &= X^T XX^T X \end{aligned} \tag{16}$$

其中,S_t 和 S_b 分别表示 LDA 中的类内矩阵和类间离散度矩阵。因此,通过求解式(15)可以得到最终的矩阵 A:

$$\boldsymbol{A}=\underset{\boldsymbol{A}}{\arg\max}\ \mathrm{tr}\{(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{S}_t\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{S}_b\boldsymbol{A}\} \tag{17}$$

由于式(17)是一个 LDA 求解问题,因此式(17)对应的全局最优解是:求 $\boldsymbol{S}_t^{-1}\boldsymbol{S}_b$ 中非零属性值所对应的属性向量。如果 $\boldsymbol{S}_t$ 是奇异的,那么计算 $\boldsymbol{S}_t^{+}\boldsymbol{S}_b$ 中非零属性值对应的属性向量,而 $\boldsymbol{S}_t^{+}$ 表示 $\boldsymbol{S}_t$ 的伪逆矩阵。由于系数矩阵的最优解 $\boldsymbol{Z}^{*}=\boldsymbol{A}^{*}\boldsymbol{B}^{*}$ 的列空间和 $\boldsymbol{A}$ 的列空间相同,因此本文方法与特殊的正则化 LDA 具有同样的列空间。

算法 2 LDA 求解问题的优化

输入 训练数据集 $\boldsymbol{X}\in\mathbb{R}^{n\times d}$,正则化参数 α,β,低秩参数 r,聚类中心数目 c,属性中控制关联结构度的参数 p

输出 矩阵 $\boldsymbol{A}\in\mathbb{R}^{d\times r}$,$\boldsymbol{B}\in\mathbb{R}^{r\times d}$和聚类指示矩阵 $\boldsymbol{U}\in\mathbb{R}^{n\times c}$

1) 初始化 $t=0$;

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

3) 通过在 $\boldsymbol{XAB}$ 上运用 K 均值聚类算法初始化矩阵 $\boldsymbol{U}^{(t)}$;

4) 初始化 $\boldsymbol{D}^{(t)}=\boldsymbol{I}\in\mathbb{R}^{d\times d}$;

5) 重复步骤 6) ~步骤 11):

6) 通过式(10)计算 $\boldsymbol{U}^{(t+1)}$;

7) 通过式(11)计算 $\boldsymbol{G}^{(t+1)}$;

8) 通过式(14)计算 $\boldsymbol{B}^{(t+1)}$;

9) 通过式(17)计算 $\boldsymbol{A}^{(t+1)}$;

10) 更新对角矩阵 $\boldsymbol{D}^{(t+1)}\in\mathbb{R}^{m\times m}$,第 i 个对角元素由式(9)计算得出,其中,$z^{i}=[\boldsymbol{A}^{(t+1)}\boldsymbol{B}^{(t+1)}]^{i}$;

11) 更新 $t=t+1$;

12) 直到目标函数式(7)收敛;

13) 结束。

本文提出算法具有如下优点:

1) 利用线性回归模型有效结合自表达方法,确保能选取最大限度地反映原始数据主要属性的自表达矩阵。同时,利用 K 均值聚类算法得到伪类标签来最大化类间距和最小化稀疏学习中的残差数据。

2) 在新模型中合理利用低秩属性选择,即运用 $l_{2,p}$-范数来考虑属性及其对应响应变量间的相关性。通过调节 p 和惩罚参数来得到合理的稀疏结构,同时,利用低秩约束来考虑不同类别之间的相关性并揭示数据在空间分布中的结构信息和类别信息,以消除输入数据中的冗余属性和不相关属性。

3) 在新的回归模型中,融入子空间学习中的线性判别分析算法来调整子空间属性选择的结果,以此使得本文算法比单一的子空间学习或者属性选择方法具有更强的分类能力。

4) 对本文的目标函数提出一种区别于交替方向乘子法的优化求解方法,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方法的优化求解算法,不断交替地迭代优化使得结果达到全局最优。该优化算法能保证目标值在每次迭代过程中逐步收敛趋近于全局最优解,最终取得全局最优解。

2 优化算法收敛性证明

由于式(8)是非平滑的,且有 4 个变量需要优化,而正则化项也是非平滑的,因此使得目标函数求解更加困难。但本文提出了一个简单且有效的算法来求解该问题,下文给出算法 2 收敛性证明的具体步骤。

从算法 2 中的步骤 5 ~步骤 13,能够得到在第 t 次迭代中,重构系数矩阵和聚类中心矩阵分别表示为 $\boldsymbol{Z}^{(t)}=\boldsymbol{A}^{(t)}\boldsymbol{B}^{(t)}$ 和 $\boldsymbol{G}^{(t)}$。然后,在第 $t+1$ 次迭代中,依据式(10)规则固定矩阵 $\boldsymbol{A}^{(t)}$,$\boldsymbol{B}^{(t)}$ 和 $\boldsymbol{G}^{(t)}$ 来更新优化 $\boldsymbol{U}^{(t+1)}$ 可以得到如下不等式:

$$\begin{aligned}&\|\boldsymbol{X}-\boldsymbol{XA}^{(t)}\boldsymbol{B}^{(t)}\|_{\mathrm{F}}^{2}+\beta\|\boldsymbol{A}^{(t)}\boldsymbol{B}^{(t)}\|_{2,p}\\&+\alpha\|\boldsymbol{XA}^{(t)}\boldsymbol{B}^{(t)}-\boldsymbol{U}^{(t+1)}(\boldsymbol{G}^{(t)})^{\mathrm{T}}\|_{\mathrm{F}}^{2}\\&\leqslant\|\boldsymbol{X}-\boldsymbol{XA}^{(t)}\boldsymbol{B}^{(t)}\|_{\mathrm{F}}^{2}+\beta\|\boldsymbol{A}^{(t)}\boldsymbol{B}^{(t)}\|_{2,p}\\&+\alpha\|\boldsymbol{XA}^{(t)}\boldsymbol{B}^{(t)}-\boldsymbol{U}^{(t)}(\boldsymbol{G}^{(t)})^{\mathrm{T}}\|_{\mathrm{F}}^{2}\end{aligned}\tag{18}$$

同样地,在第 $t+1$ 次迭代中,依据式(13)规则固定矩阵 $\boldsymbol{U}^{(t+1)}$,更新优化矩阵 $\boldsymbol{A}^{(t+1)}$,$\boldsymbol{B}^{(t+1)}$ 和 $\boldsymbol{G}^{(t+1)}$ 可以得到如下不等式:

$$\begin{aligned}&\|\boldsymbol{X}-\boldsymbol{XA}^{(t+1)}\boldsymbol{B}^{(t+1)}\|_{\mathrm{F}}^{2}+\beta\|\boldsymbol{A}^{(t+1)}\boldsymbol{B}^{(t+1)}\|_{2,p}\\&+\alpha\|\boldsymbol{XA}^{(t+1)}\boldsymbol{B}^{(t+1)}-\boldsymbol{U}^{(t+1)}(\boldsymbol{G}^{(t+1)})^{\mathrm{T}}\|_{\mathrm{F}}^{2}\\&\leqslant\|\boldsymbol{X}-\boldsymbol{XA}^{(t)}\boldsymbol{B}^{(t)}\|_{\mathrm{F}}^{2}+\beta\|\boldsymbol{A}^{(t)}\boldsymbol{B}^{(t)}\|_{2,p}\\&+\alpha\|\boldsymbol{XA}^{(t)}\boldsymbol{B}^{(t)}-\boldsymbol{U}^{(t+1)}(\boldsymbol{G}^{(t)})^{\mathrm{T}}\|_{\mathrm{F}}^{2}\end{aligned}\tag{19}$$

最后,联合不等式(18)和不等式(19)可以得到:

$$\begin{aligned}&\|\boldsymbol{X}-\boldsymbol{XA}^{(t+1)}\boldsymbol{B}^{(t+1)}\|_{\mathrm{F}}^{2}+\beta\|\boldsymbol{A}^{(t+1)}\boldsymbol{B}^{(t+1)}\|_{2,p}\\&+\alpha\|\boldsymbol{XA}^{(t+1)}\boldsymbol{B}^{(t+1)}-\boldsymbol{U}^{(t+1)}(\boldsymbol{G}^{(t+1)})^{\mathrm{T}}\|_{\mathrm{F}}^{2}\\&\leqslant\|\boldsymbol{X}-\boldsymbol{XA}^{(t)}\boldsymbol{B}^{(t)}\|_{\mathrm{F}}^{2}+\beta\|\boldsymbol{A}^{(t)}\boldsymbol{B}^{(t)}\|_{2,p}\\&+\alpha\|\boldsymbol{XA}^{(t)}\boldsymbol{B}^{(t)}-\boldsymbol{U}^{(t)}(\boldsymbol{G}^{(t)})^{\mathrm{T}}\|_{\mathrm{F}}^{2}\end{aligned}\tag{20}$$

上述不等式(20)表明算法 2 在每次迭代中,目标函数都是单调递减的。因此,算法 2 可以得到收敛的结果。

3 实验与结果分析

3.1 实验数据集和对比算法

本文用 6 个数据集测试 RS_FS 属性选择算法的性能,其中数据集 ecoli-uni 和 CNAE-9 来自 UCI[19],Smk-can,PCMAC,BASEHOCK 和 warpPIE10 来自属性选择数据集[20],实验数据集信息统计结果见表 1。

表 1 数据集信息统计

数据集	样本数	属性数	类别数	数据类型
BASEHOCK	1 993	1 500	2	text
PCMAC	1 080	856	2	text
Smk-can	187	19 993	2	multivariate,bio
ecoli-uni	336	343	8	multivariate
CNAE-9	1 080	856	9	multivariate,text
warpPIE10	210	2 420	10	image,face

实验运行在 Window 7 系统上，使用 Matlab 2014a 软件进行编程实现。为了更有效地反映算法的有效性，实验选用 4 种具有代表性的对比算法与 RS_FS 算法进行比较。各算法均在同一实验环境下进行。对比算法具体情况如下:

1) NFC 算法: 作为一种基准对比算法，未对原始数据做任何预处理，直接使用 LIBSVM 工具箱[21] 进行 SVM 分类，因此，相比 RS_FS 算法不但数据处理量大，且容易受到噪声数据和冗余数据的影响，时间复杂度为 $O(n^3)$。

2) LDA 算法[6]: 在数据处理过程中，仅考虑数据之间的全局结构。相比 RS_FS 算法，单一的 LDA 算法未对原始数据做相应预处理，且在进行空间投影时，数据局部结构容易被忽略。因此，该算法容易被离群数据和噪声数据干扰，时间复杂度为 $O(n^3)$。

3) RFS 算法[22]: 通过权值大小(赋值大权重给重要的属性，赋值小权重给不重要的属性) 来保证属性的重要性强弱，而且采用 $l_{2,1}$-范数来约束损失函数和正则化项。该算法对离群样本具有鲁棒性。相比 RS_FS 算法，RFS 算法对原始数据做了相应的稀疏处理，去除了相应的噪声数据，但是该算法未能充分地考虑数据与数据之间的相关性，尤其对于多类数据集，需考虑数据之间的相关结构，时间复杂度为 $O(n^2)$。

4) RSR 算法[23]: 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表示对应属性重要性的强弱。与 RS_FS 算法相比，RSR 只是一种无监督的属性选择算法，虽然对原始数据通过稀疏处理去除了相应的冗余数据，但是该算法既没有利用类标签进行有效的有监督选择，也没有充分考虑所有数据之间的关联结构，时间复杂度为 $O(n^2)$。

对于 RS_FS 算法，由于需要进行特征值分解，因此时间复杂度为 $O(n^3)$。

3.2　结果分析

本文参数设置如下: r 是由 m 和 k 约束得到，即 $r \leqslant \min(m,k)$，$\alpha,\beta \in \{10^{-3},10^{-1},10^{0},10^{1},10^{3}\}$，聚类中心数 c 默认设置为 10，控制关联结构度的参数 p 取值为 $0<p<2$。本文采用十折交叉验证方法将原始数据分成训练集和测试集，再利用 SVM 进行分类，同时把分类准确率作为评价指标，即分类准确率越高表示算法效果越好。所有算法均保证在同一实验环境下进行，最后提取 10 次运行的实验结果来评估各算法的性能。在每个数据集上运行 10 次减少实验误差，以此来增加实验的稳定性，各算法在6 个数据集上的实验结果对比如图 1 ~ 图 6 所示。

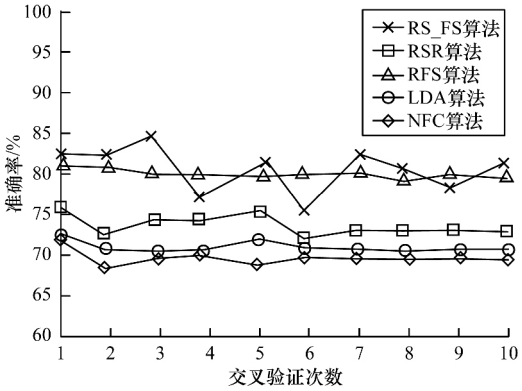

图 1　BASEHOCK 数据集上的分类结果

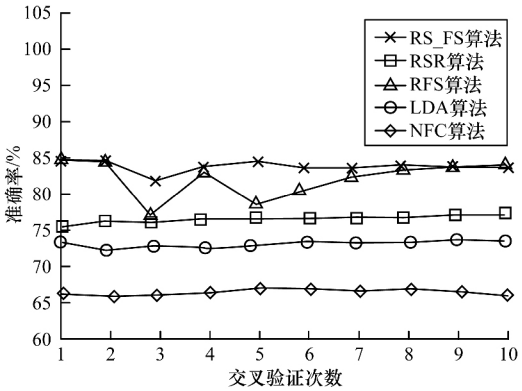

图 2　PCMAC 数据集上的分类结果

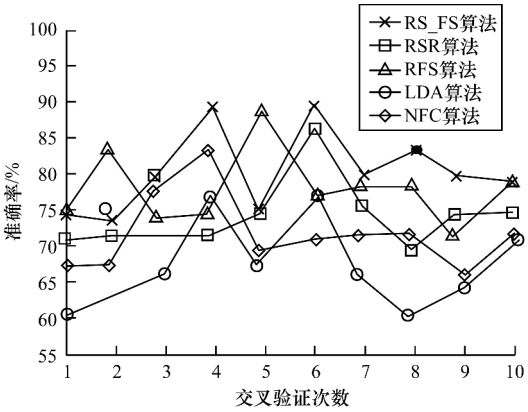

图 3　Smk-can 数据集上的分类结果

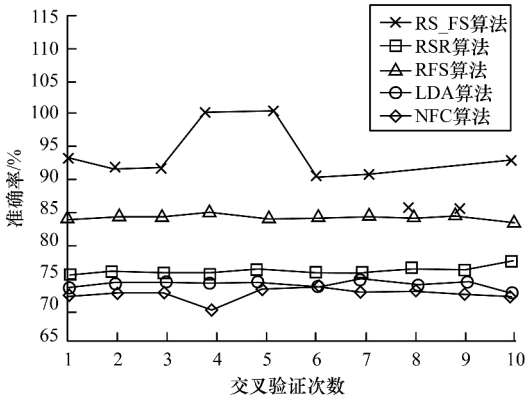

图 4　ecoli_uni 数据集上的分类结果

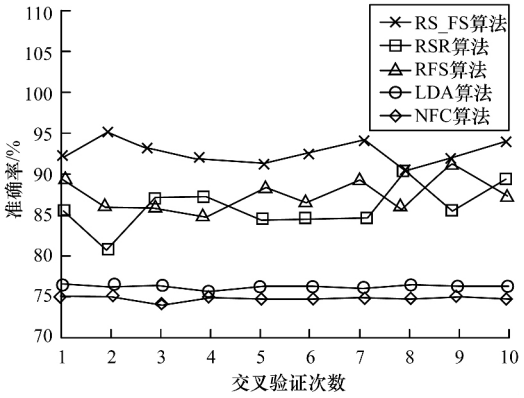

图 5 CNAE-9 数据集上的分类结果

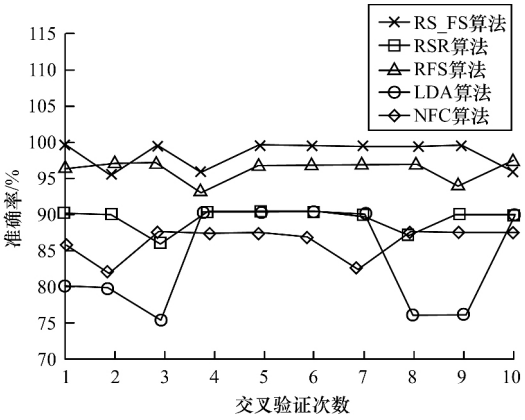

图 6 warPIE10 数据集上的分类结果

通过图 1 ~ 图 6 可以看出，RS_FS 算法在 6 个数据集上的分类准确率由于十折交叉验证的随机性，因此不能保证每次结果都是最好的，但在所有数据集上 10 次实验平均结果大部分比对比算法都好，最终的平均分类准确率也是最高的。

通过分析表 2 的分类准确率(均值 ± 方差) 可以得出，在 6 个数据集上，RS_FS 算法分类准确率均好于其他 4 种对比算法。从分类准确率的角度看，与 NFC 算法相比平均提高了 15.02%，与 RSR 和 RFS 算法相比平均提高了8.34% 和 3.36%。在多类数据集上，RS_FS 算法分类效果最好。与 RFS 对比算法相比，尤其在 ecoli_uni 数据集上平均提高了8.89%。从方差的角度，RS_FS 在大多数数据集上效果都很好，而在数据集 CNAE-9 上，由于 RS_FS 运用了十折交叉验证的方法，该算法的随机性难以保证每次的效果都达到最佳，但 RS_FS 算法只比 RSR 算法的方差小0.28%，依然具有较好的稳定性。由于 RFS 算法是有监督的属性选择，因此利用现有的类标签可以在不同类的分类上均呈现较好的效果，在各数据集上的平均分类准确率为 84.17%。但其由于在多类数据集情况未能够考虑到整体数据之间的相关结构，而 RS_FS 算法不但可以利用建立伪类标签的方法进行分类，还利用线性判别分析考虑数据之间的相关性，因此具有更好的分类效果。RSR 算法利用了自表达的方法，通过原始数据本身得到一个自表达系数矩阵，但由于是无监督的属性选择，因此实际效果不理想，在各数据集上的平均分类准确率仅为 79.19%。与其相比，RS_FS 算法不仅运用自表达的方法，还结合K 均值聚类算法得到伪类标签，同时，利用 $l_{2,p}$-范数考虑样本与伪类标签之间的关联结构和低秩约束来描述所有类别之间的关联性，以此选取更具代表性的属性集合，最后利用线性判别分析调整属性选择的结果，确保提取后的属性最大程度地代表原始数据，因此，可显著提高模型性能。

表 2 分类准确率统计结果 %

数据集	NFC 算法	LDA 算法	RFS 算法	RSR 算法	RS_FS 算法
BASEHOCK	69.95 ± 6.75	71.79 ± 4.98	80.21 ± 4.29	74.51 ± 4.25	81.04 ± 2.59
PCMAC	66.85 ± 5.42	72.88 ± 4.29	81.28 ± 2.24	76.11 ± 3.75	82.36 ± 2.17
Smk-can	67.71 ± 6.40	70.91 ± 5.87	77.53 ± 5.27	74.32 ± 5.53	79.59 ± 5.19
ecoli_uni	72.31 ± 5.28	73.92 ± 5.67	83.24 ± 5.52	76.16 ± 5.18	92.13 ± 4.98
CNAE-9	74.76 ± 2.79	76.22 ± 3.40	86.24 ± 2.27	84.79 ± 1.66	91.48 ± 1.94
warpPIE10	83.50 ± 4.12	85.78 ± 3.44	96.50 ± 2.50	89.25 ± 2.71	98.57 ± 2.30
平均值	72.51	75.25	84.17	79.19	87.53

本文利用召回率(Recall) 和 F-值(F-score) 进行二分类数据集的分析比较。如表 3 所示，在所有的二分类数据集上，本文提出的 RS_FS 算法均具有较高的召回率和 F-值。在 PCMAC 数据集上，直接利用 SVM 进行分类验证效果较好; 在 BASEHOCK 和 Smk-can 数据集上结果也较为接近。因此，通过该实验可表明 RS_FS 算法利用伪类标签对无监督学习的分类，能够接近甚至略微地超过经典的二分类 SVM 方法。

表 3 二分类数据集召回率和 F-值统计结果 %

数据集	性能指标	NFC 算法	LDA 算法	RFS 算法	RSR 算法	RS_FS 算法
BASEHOCK	召回率	89.80	82.34	89.40	89.62	93.69
	F-值	90.34	85.71	73.39	90.71	91.75
PCMAC	召回率	91.43	78.24	83.51	84.02	90.02
	F-值	91.10	77.84	82.17	84.25	89.84
Smk-can	召回率	78.89	74.44	76.67	68.89	84.44
	F-值	76.87	73.85	76.58	78.94	86.82

由于不同数据集有不同的数据分布且含有不同的干扰因素,因此对于多类数据集,如 ecoli_uni,CNAE-9 和 warpPIE10 数据集,可以通过控制秩的大小来得到不同的分类结果,如图 7 ~ 图 9 所示。

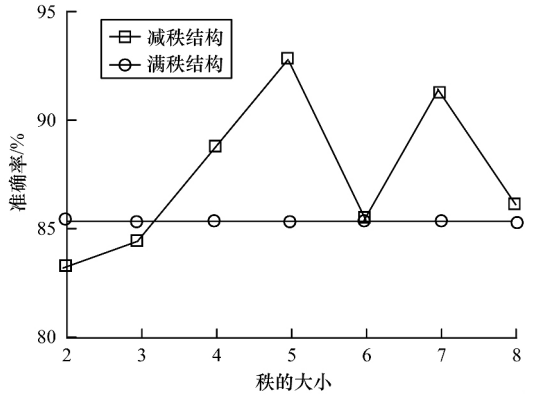

图 7　ecoli_uni 数据集上低秩和满秩比较结果

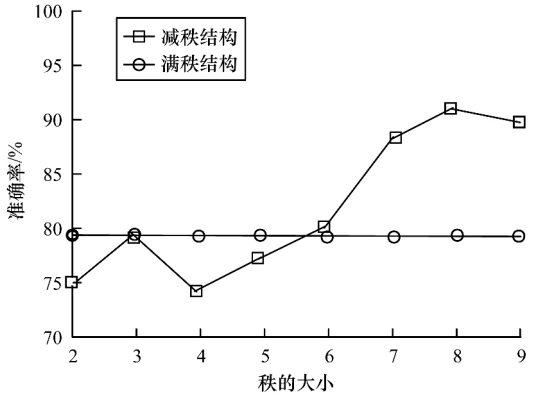

图 8　CNAE-9 数据集上低秩和满秩比较结果

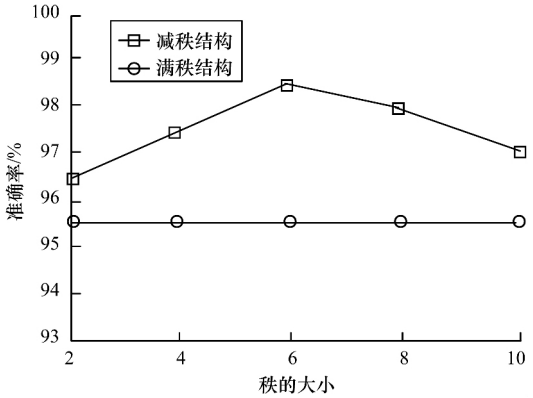

图 9　warPIE10 数据集上低秩和满秩比较结果

具有低秩约束的 RS_FS 算法比满秩的效果更佳。实验结果显示,与其他算法相比,RS_FS 算法在多类数据集上均能取得较好的效果。由此可知,RS_FS 算法不仅能保证重要属性的提取,而且可利用子空间学习对获得的重要属性进一步微调,相比单一的属性选择算法能更好地保证数据自身的结构。

4　结束语

本文提出一种新的属性选择算法,通过使用原始数据构建自表达矩阵,并整合 K 均值聚类算法、低秩属性选择和子空间方法,较好地扩展了无监督学习的应用范围。实验结果表明,本文算法具有较高的分类准确率和稳定性。下一步将尝试在半监督属性选择领域应用并验证本文算法,在减少标签信息开销的同时进一步扩展属性选择算法的应用范围。

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 Shicaho, JIN Zhi, et al. Missing Value Estimation for Mixed-attribute Data Sets [J]. IEEE Transactions on Knowledge & Data Engineering, 2010, 23(1): 110-121.

[3] ZHU X, SUK H I, SHEN D. Matrix-similarity Based Loss Function and Feature Selection for Alzheimer's Disease Diagnosis [C]//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2014: 3089-3096.

[4] ZHU X, LI X, ZHANG S. Block-row Sparse Multiview Multilabel Learning for Image Classification [J]. IEEE Transactions on Cybernetics, 2016, 46(2): 450-461.

[5] ZHU Xiaofeng, SUK H I, SHEN D. A Novel Matrix-similarity Based Loss Function for Joint Regression and Classification in AD Diagnosis [J]. NeuroImage, 2014, 100: 91-105.

[6] ZHU X, HUANG Z, CHENG H, et al. Sparse Hashing for Fast Multimedia Search [J]. ACM Transactions on Information Systems, 2013, 31(2): 595-605.

[7] 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.

[8] FAN Zizhu, XU Yong, ZHANG D. Local Linear Discriminant Analysis Framework Using Sample Neighbors [J]. IEEE Transactions on Neural Networks, 2011, 22(7): 1119-1132.

[9] GUI Jie, SUN Zhenan, JIA Wei, et al. Discriminant Sparse Neighborhood Preserving Embedding for Face Recognition [J]. Pattern Recognition, 2012, 45(8): 2884-2893.

[10] MUSORO J Z, ZWINDERMAN A H, PUHAN M A, et al. Validation of Prediction Models Based on Lasso Regression with Multiply Imputed Data [J]. BMC Medical Research Methodology, 2014, 14(1): 116.

[11] ABDI H, WILLIAMS L J. Partial Least Squares Methods: Partial Least Squares Correlation and Partial Least Square Regression [J]. Methods in Molecular Biology, 2013, 930: 549-579.

[12] ZHU Xiaofeng,HUANG Zi,SHEN Hengtao,et al. Linear Cross-modal Hashing for Efficient Multimedia Search[C]//Proceedings of the 21st ACM International Conference on Multimedia. New York,USA: ACM Press,2013:143-152.

[13] ZHU Xiaofeng,HUANG Zi,CUI Jiangtao,et al. Video-to-shot Tag Propagation by Graph Sparse Group Lasso[J]. IEEE Transactions on Multimedia,2013,15(3):633-646.

[14] ZHU X,ZHANG L,HUANG Z. A Sparse Embedding and Least Variance Encoding Approach to Hashing[J]. IEEE Transactions on Image Processing,2014,23(9):3737-3750.

[15] 钟 智,胡荣耀,何 威,等. 基于图稀疏的自表达属性选择算法[J]. 计算机工程与设计,2016,37(6):1643-1648.

[16] 程德波,苏毅娟,宗 鸣,等. 基于稀疏学习的自适应近邻分类算法[J]. 计算机工程与设计,2015,36(7):1912-1916.

[17] 邓振云,龚永红,孙 可,等. 基于局部相关性的 kNN 分类算法[J]. 广西师范大学学报(自然科学版),2016,34(1):52-58.

[18] XIANG Shuo,ZHU Yunzhang,SHEN Xiaotong,et al. Optimal Exact Least Squares Rank Minimization[C]//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New,York,USA: ACM Press,2012:480-488.

[19] UCI Repository of Machine Learning Datasets[EB/OL]. [2015-04-10]. http://archive.ics.uci.edu/ml/.

[20] Feature Selection Datasets[EB/OL]. [2015-04-10]. http://featureselection.asu.edu/datasets.php.

[21] LIBSVM——A Library for Support Vector Machines[EB/OL]. [2015-04-10]. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[22] NIE Feiping,HUANG Heng,CAI Xiao,et al. Efficient and Robust Feature Selection via Joint $l_{2,1}$-norms Minimization[C]//Proceedings of the 23rd International Conference on Neural Information Processing Systems. New York,USA: ACM Press,2010:1813-1821.

[23] ZHU Pengfei,ZUO Wangmeng,ZHANG Lei,et al. Unsupervised Feature Selection by Regularized Self-representation[J]. Pattern Recognition,2015,48(2):438-446.

编辑 陆燕菲

(上接第 42 页)

参考文献

[1] 杨义先,姚文斌,陈 钊. 信息系统灾备技术综论[J]. 北京邮电大学学报,2010,33(2):1-6.

[2] 马 强. 铁路客票系统异地灾备中心方案设计[D]. 北京: 中国铁道科学研究院,2015.

[3] 肖良华. 从灾备到双活[J]. 金融电子化,2013(11):55-56.

[4] KRASKA T,PANG G,FRANKLIN M J,et al. MDCC: Multi-data Center Consistency[C]//Proceedings of the 8th ACM European Conference on Computer Systems. New York,USA: ACM Press,2013:113-126.

[5] WILKINSON J P. Managing Remote Data Replication: U.S. Patent 8,868,874[P]. 2014-10-21.

[6] HRLE N,MARTIN D,MOHAN C,et al. Data Replication in a Database Management System: U.S. Patent Application 14/592,165[P]. 2015-01-08.

[7] JUNG H,HAN H,FEKETE A,et al. Serializable Snapshot Isolation for Replicated Databases in High-update Scenarios[EB/OL]. (2011-08-10). http://www.vldb.org/pvldb/vol4/p783-jung.pdf.

[8] 王 珏. 基于快照隔离的分布式数据库同步技术研究与应用[D]. 郑州: 解放军信息工程大学,2012.

[9] 陈广明. 基于 SQLite 的分布式数据同步技术研究与实现[D]. 广州: 华南理工大学,2013.

[10] 曹雪春. 灾难备份系统的数据同步技术研究及应用[D]. 上海: 上海交通大学,2012.

[11] 彭远浩,潘久辉. 基于日志分析的增量数据捕获方法研究[J]. 计算机工程,2015,41(6):56-60,65.

[12] OSMANALIHEGAZI M. An Approach for Designing and Implementing Eager and Lazy Data Replication[J]. International Journal of Computer Applications,2015,110(6):30-33.

[13] 张大朋,陈 驰,徐 震. 异构数据库复制技术的研究与实现[J]. 中国科学院大学学报,2012,29(1):101-108.

[14] VARLEY I,HANSMA S,BURSTEIN P. Multi-master Data Replication in a Distributed Multi-tenant System: U.S. Patent Application 13/252,214[P]. 2012-10-11.

[15] HAMILTON G,CATTELL R,FISHER M. JDBC Database Access with Java[M]. Boston,USA: Addison Wesley,1997.

[16] Oracle Corporation. mysql-connector-java-5.1.37[EB/OL]. [2016-04-20]. https://dev.mysql.com/doc/relnotes/connector-j/5.1/en/news-5-1-37.html.

编辑 金胡考