文章编号: 1001-0920 (2013) 09-1485-06

一种基于最大边界投影和 $l_{2,1}$ 范数正则化的属性选择算法

夏建明, 杨俊安

(电子工程学院 a. 通信对抗系，b. 电子制约技术安徽省重点实验室，合肥 230037)

摘 要: 当数据含有噪声或标签错误时, 传统的属性选择方法 (如粗糙集) 无法得到正确结果, 为此提出一种针对含噪、标签错误数据的属性选择方法. 首先用最大边界投影方法获得数据的最佳投影; 然后通过对投影矩阵进行 $l_{2,1}$ 范数正则化操作, 进而获得行稀疏的投影矩阵, 据此获得对关键属性的挖掘; 最后给出方法的收敛性和针对标签错误数据的有效性证明. 实验结果表明, 所提出的算法克服了噪声和标签错误的影响, 较好地实现了针对含噪、标签错误数据的属性选择.

关键词: 属性选择；最大边界投影；$l_{2,1}$ 范数；噪声数据；标签错误

中图分类号: TN911.5 **文献标志码:** A

A novel attribute reduction algorithm based on maximum margin projection and $l_{2,1}$ norm regularization

XIA Jian-ming, *YANG Jun-an*

(a. Department of Communication Countermeasures，b. Key Laboratory of Electronic Restriction of Anhui Province, Electronic Engineering Institute，Hefei 230037，China．Correspondent：XIA Jian-ming，E-mail：jianmingeei@163.com)

Abstract: The traditional attribute reduction algorithms such as rough set will fail to get accurate results when deal with the data sets which have noise or labeling errors. Therefore, this paper proposes an attribute reduction algorithm which can analyze this kind of data effectively. Firstly, the best projection of the data sets is obtained by using the maximum margin projection(MMP) method. Then $l_{2,1}$-norm on the projection matrix is used to achieve row-sparsity, which leads to selecting relevant features. Finally, the proof of the algorithm's convergence and validity to the data sets with errors is given. The result of experiments on the UCI data sets show the effectiveness of the proposed algorithm.

Key words: attribute reduction；maximum margin projection；$l_{2,1}$ norm；noise data；labeling error

0 引 言

随着信息时代的到来, 海量复杂的数据在各个领域中不断涌现, 人们希望自动地从数据中获取其潜在的知识模型. 在知识挖掘过程中, 首先需要对大量的属性进行选择, 通过属性选择, 去除冗余属性, 获得关键属性, 最终获得对规则的挖掘. 大批学者对这一领域进行研究, 取得了丰硕的成果, 但这些挖掘工具的前提是数据是正确的, 当数据中各属性中含有噪声或数据标签发生错误时, 其选择结果将发生错误. 其中经典方法如粗糙集, 由于其无需任何先验知识, 能从大量含糊和不确定的数据中发现有用信息, 在属性选择领域得到了广泛的应用. 但是由于其对不可分辨性的严格规定, 无法应对噪声和标签错误数据. 之后学者们又从 3 个方面对该问题进行研究, 一是从提高算法的鲁棒性着手, 如徐怡等[1] 引入正确分类率 β, 提出了可变精度粗糙集模型, 通过允许一定的错误分类率存在来完善近似空间; 但是在应用过程中参数 β 往往仅能通过领域专家按照知识和经验指定, 从而失去了最初粗糙集不需要任何先验知识、仅从数据本身出发的初衷, 且针对不同数据需要确定不同的错误分类率, 一旦错误分类率给定后, 在数据本身变化的情况下将获得错误的结果, 这将制约属性选择的应用. 变精度粗糙集的一系列改进方法也尚未给出正确分类率的自适应求解方法[2]. 二是针对数据清洗、过滤进行研究, 如建立多个分类器, 使得分类器预测错误最多的样本作为发生错误的数据过滤掉[3]. 三是在预先

收稿日期: 2012-05-17；**修回日期:** 2012-08-30.

基金项目: 安徽省自然科学基金项目(1208085MF94).

作者简介: 夏建明(1982−), 男, 博士生, 从事数据挖掘、机器学习的研究；杨俊安(1965−), 男, 教授, 博士生导师, 从事信号处理、智能计算等研究.

获得错误数据的统计信息的基础上构建模型, 如根据错误分布概率模型设计 NB (Naïve Bayes) 分类器来捕获错误样本的方法[4]. 这些方法要么仅仅利用投票原则进行选择, 要么需要对样本中的错误信息有足够的先验信息, 这些特点都或多或少限制了相关方法的应用范围.

近几年来, 利用研究对象的稀疏特性建立稀疏表示的模型来解决问题已成为研究热点, 促进了压缩感知、模式识别、计算机视觉、图像处理等众多领域的新一轮发展[5-6]. 然而这些基于稀疏表示的机器学习算法基本上都是研究如何用更少的元素更好地表达原数据, 而应用在数据挖掘、知识发现方面的研究较少, 缺乏适合于数据挖掘中规则与数据之间相对稀疏关系的稀疏表示学习模型. 与数据挖掘目标方向相近的应用有, 将稀疏表示与以往属性抽取的方法结合到一起应用到模式识别上, 如稀疏主分量分析 (SPCA)[7]、稀疏线性鉴别分析 (SLDA)[8]、稀疏子空间学习算法 (SSL)[9]等. 这类方法都只是致力于解决属性抽取的问题, 目的是使得到的新属性更加具有鉴别性, 其不同点仅在于稀疏性和稀疏估计参数的构建方式不同. 如稀疏子空间学习算法基于对投影矩阵中的行向量进行 L_1 范数正则化运算, 但并没有成功地实现属性选择; 最为接近的是线性判别特征选择算法[10], 但其仅研究模式识别问题, 且计算复杂度太大, 只适用于小规模的数据. 文献 [11] 针对图像重构问题提出了最大边界投影方法, 通过构建类内紧凑图和类间分离图, 进而挖掘数据样本集中多个子流形之间的联系来得到最佳投影矩阵, 通过这种属性抽取的方法完成了高维图像到低维元素的映射, 较好地反映了样本集的数据结构. 属性抽取与属性选择这两类方法虽然最终结果不一样, 但它们都致力于获得更具鉴别性的效果, 解空间应该是相同的. 以知识发现为目标, 将属性抽取的方法与稀疏表示相结合, 综合利用属性抽取方法的解空间和稀疏表示方法的特性及对噪声的鲁棒性优点, 使算法在获得正确关键属性的同时具有对噪声的鲁棒性.

本文首先通过最大边界投影方法获得最优的投影方向, 然后据此进行 $l_{2,1}$ 范数最小化运算求解得到行稀疏的投影矩阵, 其中不为零的行向量所对应的属性即为所求的关键属性, 理论证明和实验分析表明本文方法针对含噪、标签错误数据具有较好的效果.

1 最大边界投影方法

1.1 最大边界投影方法的目标函数

最大边界投影方法认为数据是嵌入在多个低维子流形中, 因而致力于挖掘出既具有鉴别特点又反映数据几何结构的投影矩阵[11]. 设 m 个数据点 $X = (x_1, x_2, \cdots, x_m) \in R^n$ 嵌入在一个流形 M 中, 对于每个数据点 x_i, 可以发现它的 k 个近邻 $N(x_i) = x_i^1, x_i^2, \cdots, x_i^k$, 为了同时获得数据流形的几何结构和鉴别信息, 采用最大边界投影方法构建类内紧凑图 G_w 和类间分离图 G_b. 将 x_i 的近邻 $N(x_i)$ 分为两个子集: $N_w(x_i)$ 和 $N_b(x_i)$. $N_w(x_i)$ 含有近邻中与 x_i 相同类别的数据, 而 $N_b(x_i)$ 则是近邻中与 x_i 不同的类别, 即

$$N_w(x_i) = \{x_i^j | l(x_i^j) \neq l(x_i),\ 1 \leqslant j \leqslant k\}, \tag{1}$$

$$N_b(x_i) = N(x_i) - N_w(x_i), \tag{2}$$

其中用 $l(x_i)$ 表示 x_i 的类别. 设 W_w 和 W_b 分别为两个图 G_w、G_b 的权重矩阵, 则定义为

$$W_{w,ij} = \begin{cases} 1,\ x_i, x_j\text{同类且在邻域内;} \\ 0,\ x_i, x_j\text{异类且在邻域内;} \end{cases} \tag{3}$$

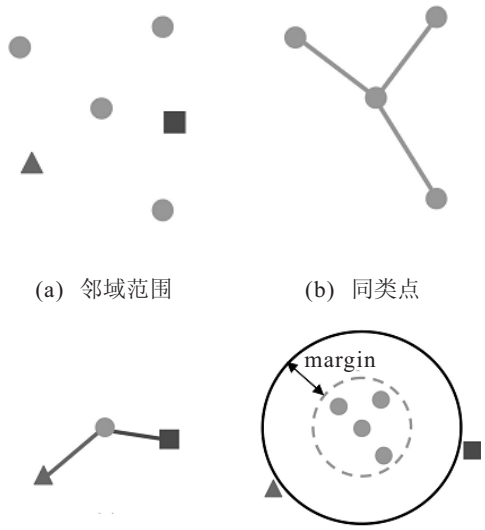
$$W_{b,ij} = \begin{cases} 0,\ x_i, x_j\text{同类且在邻域内;} \\ 1,\ x_i, x_j\text{异类且在邻域内.} \end{cases} \tag{4}$$

假设投影值为 $Y = X^{\mathrm{T}}A$, $Y = (y_1, y_2, \cdots, y_m)^{\mathrm{T}}$, 由图嵌入框架可知, 一个好的投影值需要使得类内更加紧凑、类间更加离散, 即满足

$$\min \sum_{ij} (y_i - y_j)^2 w_{w,ij}, \tag{5}$$

$$\max \sum_{ij} (y_i - y_j)^2 w_{b,ij}. \tag{6}$$

其思想如图 1 所示, 在目标点的邻域范围内 (图 1(a)) 有 3 个同类点 (图 1(b)) 和 2 个异类点 (图 1(c)), 最大边界投影方法旨在使邻域中目标点的同类点之间更加紧凑, 而邻域范围内同类点与异类点之间的边界最大 (图 1(d)).



(a) 邻域范围 (b) 同类点

(c) 异类点 (d) 方法效果

图 1 最大边界投影方法思想示意图

1.2 最优线性嵌入的获得

由目标函数 (5) 可以简化得到

$$\frac{1}{2}\sum_{ij}(y_i-y_j)^2w_{w,ij}=$$
$$\frac{1}{2}\sum_{ij}(x_i^{\mathrm{T}}a-x_j^{\mathrm{T}}a)^2w_{w,ij}=$$
$$\sum_i x_i^{\mathrm{T}}aD_{w,ii}a^{\mathrm{T}}x_i-\sum_{ij}x_i^{\mathrm{T}}aw_{w,ii}a^{\mathrm{T}}x_j=$$
$$X^{\mathrm{T}}AD_wA^{\mathrm{T}}X-X^{\mathrm{T}}AW_wA^{\mathrm{T}}X. \tag{7}$$

同理, 目标函数 (6) 可简化为

$$X^{\mathrm{T}}A(D_b-W_b)A^{\mathrm{T}}X. \tag{8}$$

其中: D_w、D_b 均为对角阵, 其元素 $D_{w,ii}=\sum_j W_{w,ij}$, $D_{b,ii}=\sum_j W_{b,ij}$. 对 D_w 施加一个约束 $X^{\mathrm{T}}AD_wA^{\mathrm{T}}X=1$, 则目标函数 (7) 可进一步简化为

$$\min_A 1-X^{\mathrm{T}}AW_wA^{\mathrm{T}}X=\max_A X^{\mathrm{T}}AW_wA^{\mathrm{T}}X. \tag{9}$$

两个目标函数可以综合为一个优化函数

$$\arg\max_{X^{\mathrm{T}}AD_wA^{\mathrm{T}}X=1} X^{\mathrm{T}}A[\alpha(D_b-W_b)+(1-\alpha)W_w]A^{\mathrm{T}}X. \tag{10}$$

求解式 (10) 便可获得最优投影矩阵 A, 这个问题可以通过广义特征值分解来计算. 作为属性抽取的方法, 最大边界投影方法已经成功地将原高维数据投影到低维空间中, 但无法从这样的 A 中获得利于属性选择的信息.

2 基于最大边界投影和 $l_{2,1}$ 范数正则化的属性选择算法

假设解空间中良好的分类器为 $Y=X^{\mathrm{T}}A$ (数据集 X 为 $n\times m$ 的矩阵, Y 为 $m\times d$, 投影矩阵 A 为 $n\times d$, 其中 n 为属性个数, m 为数据个数, d 为投影空间的维数), 若能通过稀疏表示的方法在保证分类精度的前提下获得行稀疏的矩阵 A, 则使得对应于冗余属性的行向量数值趋近于 0, 而对应于关键属性的行向量数值得以保存. 通过这种矩阵的行稀疏模式可得到关键的属性, 实现属性选择.

秉持这种思想, 可通过先求解 Y 再获得行稀疏的 A 的方式来获得属性选择:

1) 对 Y 的求解. 由 $Y=X^{\mathrm{T}}A$ 和式 (10) 可得优化函数为

$$\arg\max_{X^{\mathrm{T}}AD_wA^{\mathrm{T}}X=1} Y[\alpha(D_b-W_b)+(1-\alpha)W_w]Y^{\mathrm{T}}. \tag{11}$$

其中: D_b、W_b、W_w 均可从数据集中获得, α 预先设定. 则 Y 可以通过对 $\alpha(D_b-W_b)+(1-\alpha)W_w$ 进行特征值分解获得.

2) 求解行稀疏的 A. 获得 Y 后, 将属性选择问题转化为如下优化问题:

$$\arg\min(||A||_{2,1}),\ \text{s.t. } X^{\mathrm{T}}A=Y. \tag{12}$$

其中 $||A||_{2,1}=\sum_i^n||a_i||_2$, a_i 为投影矩阵 A 中第 i 个行向量. 考虑到数据集中可能存在噪声, 进一步将优化函数确定为

$$\arg\min(||A||_{2,1}),\ \text{s.t. } ||X^{\mathrm{T}}A-Y||_F\leqslant\delta. \tag{13}$$

即 $\arg\min(||A||_{2,1}+\beta||X^{\mathrm{T}}A-Y||_F)$, 其中 β 为参数. 然而有

$$||X^{\mathrm{T}}A-Y||_F=$$
$$\mathrm{tr}(Y^{\mathrm{T}}Y)+\mathrm{tr}(X^{\mathrm{T}}AA^{\mathrm{T}}X)-2\mathrm{tr}(X^{\mathrm{T}}AY), \tag{14}$$

则式 (13) 的拉格朗日函数为

$$L(A)=||A||_{2,1}+\beta(\mathrm{tr}(Y^{\mathrm{T}}Y)+\mathrm{tr}(X^{\mathrm{T}}AA^{\mathrm{T}}X)-2\mathrm{tr}(X^{\mathrm{T}}AY)). \tag{15}$$

对其进行求导, 可得

$$\frac{\partial(L)}{\partial(A)}=GA+2\beta(XX^{\mathrm{T}}A-XY^{\mathrm{T}})=0. \tag{16}$$

由于 $||A||_{2,1}$ 不是光滑的, 可计算它的子梯度代替梯度. 其中 G 是对角矩阵, 且其中元素为

$$g_{i,i}=\begin{cases}0,\ A^i=0;\\ 1/||A^i||_2,\ \text{otherwise.}\end{cases} \tag{17}$$

A^i 表示 A 中第 i 个行向量. 最终解得

$$A=2\beta(G+2\beta XX^{\mathrm{T}})^{-1}XY^{\mathrm{T}}. \tag{18}$$

由于 G 的取值取决于每次迭代运算后的投影矩阵 A, 而数据本身 X 已知, Y 已由第 1 步算出, 最终可通过固定点算法迭代将行稀疏的投影矩阵 A 获得. 最终算法步骤如下.

输入: 数据集 $X=(x_1,x_2,\cdots,x_m)\in R^n$;

输出: 关键属性.

Step 1: 求解数据集 $X=(x_1,x_2,\cdots,x_m)\in R^n$ 中每个点的 K 最近邻点, 构成各个点的近邻域.

Step 2: 根据每个点及其邻域内点的类别构造类内紧凑图 G_w 和类间分离图 G_b 及相应的权重矩阵 W_w 和 W_b, 进而得到 D_w 和 D_b.

Step 3: 设定参数 α 的大小, 对 $\alpha(D_b-W_b)+(1-\alpha)W_w$ 进行广义特征值分解, 获得矩阵 Y.

Step 4: 根据式 (18) 迭代获得行稀疏的投影矩阵 A.

Step 5: 根据 A 中不为零向量的行所对应的属性确定最终挖掘的关键属性.

3 对算法收敛性、含错数据有效性的相关证明

3.1 算法的收敛性证明

由于 $\frac{\partial(||A||_{2,1})}{\partial(A)}=GA=\frac{1}{2}\frac{\partial(\mathrm{tr}(A^{\mathrm{T}}GA))}{\partial(A)}$, 求解

式(13)等同于求解以下问题:

$$\arg\min(\mathrm{tr}(A^{\mathrm{T}}GA)),\ \mathrm{s.t}\,\|A^{\mathrm{T}}X-Y\|_F\leqslant\delta. \tag{19}$$

在第t次迭代时有

$$A_{t+1}=\underset{A^{\mathrm{T}}\bar{X}=Y}{\arg\min}(\mathrm{tr}(A_t^{\mathrm{T}}G_tA_t)), \tag{20}$$

则有

$$\mathrm{tr}(A_{t+1}^{\mathrm{T}}G_tA_{t+1})\leqslant\mathrm{tr}(A_t^{\mathrm{T}}G_tA_t), \tag{21}$$

代入G_t的表达式有

$$\sum_{i=1}^{m}\frac{||A_{t+1}^i||_2^2}{2||A_t^i||_2}\leqslant\sum_{i=1}^{m}\frac{||A_t^i||_2^2}{2||A_t^i||_2}. \tag{22}$$

此外, 由$(||A_{t+1}^i||_2-||A_t^i||_2)^2\geqslant 0$可以得到

$$||A_{t+1}^i||_2^2+||A_t^i||_2^2-2||A_{t+1}^i||_2||A_t^i||_2\geqslant 0\Rightarrow$$

$$||A_{t+1}^i||_2-\frac{||A_{t+1}^i||_2^2}{2||A_t^i||_2}\leqslant\frac{||A_t^i||_2}{2}, \tag{23}$$

$$||A_{t+1}^i||_2-\frac{||A_{t+1}^i||_2^2}{2||A_t^i||_2}\leqslant\frac{||A_t^i||_2}{2}\Rightarrow$$

$$||A_{t+1}^i||_2-\frac{||A_{t+1}^i||_2^2}{2||A_t^i||_2}\leqslant||A_t^i||_2-\frac{||A_t^i||_2^2}{2||A_t^i||_2}. \tag{24}$$

进而有

$$\sum_{i}^{m}||A_{t+1}^i||_2-\frac{||A_{t+1}^i||_2^2}{2||A_t^i||_2}\leqslant\sum_{i}^{m}||A_t^i||_2-\frac{||A_t^i||_2^2}{2||A_t^i||_2}. \tag{25}$$

联合式(22)可知$\sum_{i}^{m}||A_{t+1}^i||_2\leqslant\sum_{i}^{m}||A_t^i||_2$, 即

$$||A_{t+1}||_2^1\leqslant||A_t||_2^1. \tag{26}$$

算法每次迭代的结果是递减的, 必然会收敛到一个合适的值; 同时, 由结果表达式$A=2\beta(G+2\beta\bar{X}\bar{X}^{\mathrm{T}})^{-1}\bar{X}Y^{\mathrm{T}}$可知, 由于$\bar{X}$为已知数据, Y^{T}计算得出后也相当于已知量, 不需要在迭代过程中进行运算; 而G为对角矩阵, A的表达式为线性, 算法可以较快的速度和较少的迭代步数达到收敛.

3.2 对标签错误数据进行属性选择的有效性证明

由A的表达式可以知道, Y是由W特征值分解得到的, 当数据发生标签错误时, Y肯定要发生某种变化, 设变化后的为Y', 则A相应地变为

$$A'=2\beta(G+2\beta\bar{X}\bar{X}^{\mathrm{T}})^{-1}\bar{X}(Y')^{\mathrm{T}}. \tag{27}$$

设$Z=Y^{\mathrm{T}}$, 相应地$Z'=(Y')^{\mathrm{T}}$.

由极式分解定理可知, 若$Z^{\mathrm{T}}Z'=U\sum V^{\mathrm{T}}$是矩阵乘积$Z^{\mathrm{T}}Z'$的奇异值分解, 则$Q=UV^{\mathrm{T}}$是$\min||Z'-ZQ||_F$的解. 即可以找到一个正交矩阵$Q$, 使得$Z'=ZQ$, 进而可以得到当数据发生标签错误时的最终结果为

$$A'=2\beta(G+2\beta\bar{X}\bar{X}^{\mathrm{T}})^{-1}\bar{X}(Y')^{\mathrm{T}}a=$$
$$2\beta(G+2\beta\bar{X}\bar{X}^{\mathrm{T}})^{-1}\bar{X}ZQ=AQ. \tag{28}$$

由式(28)可知, 若数据发生标签错误, 则投影矩阵A的变化即相当于右乘正交矩阵Q, 而行稀疏矩阵右乘正交矩阵的结果仍然是个行稀疏矩阵, 这种变化不会改变投影矩阵的行稀疏以及矩阵中哪一行稀疏的性质, 亦不会改变所选择的关键属性.

4 实验验证

实验数据集来自UCI数据库, 共有Iris、Glass、Diabets和Wine四个数据集, 表1为各个数据集的相关情况. 为力求实验准确性, 污染数据的获得均重复100次, 通过平均效果进行验证. 变精度粗糙集中错误分类率设定为0.2. 由于根据错误数据建模的方法需要预先知道噪声的概率分布[2], 而通过数据清洗、通过集成分类器学习的方法适用于数据挖掘无法实现的属性选择[3], 故选择将本文方法与邻域粗糙集、变精度粗糙集效果进行对比, 分3个实验进行验证.

表1 实验数据描述

数据集	样本数	属性数	类别数
Iris	150	4	3
Glass	683	9	2
Diabets	768	8	2
Wine	178	13	3

4.1 原始数据属性选择效果验证

基于本文方法与邻域粗糙集、变精度粗糙集3种方法对5个数据集选择属性. 引入SVM分类学习算法, 以10折交叉验证的分类精度来检验属性选择的效果.

如表2所示, 在不添加噪声的条件下, 邻域粗糙集、变精度粗糙集和本文方法均取得了较好的属性选择效果. 总之, 邻域粗糙集和变精度粗糙集选择的属性一样, 分类正确率较好, 基于稀疏表示的属性选择方法选择的属性略有不同, 分类效果稍差.

表2 3种方法对原始数据选择的属性及分类精度

数据集	邻域粗糙集		变精度粗糙集		本文方法	
	选择属性	分类精度/%	选择属性	分类精度/%	选择属性	分类精度/%
Iris	4	96	4	96	4	96
Glass	2 3 4 7 9	72	2 3 4 7 9	72	3 4 5 7 9	69
Diabets	2 6 3	90.62	2 6 3	90.62	1 4 3 5	89.45
Wine	2 5 13	100	2 5 13	100	2 7 10 13	98.31

4.2 噪声污染数据属性选择效果验证

1) 噪声污染数据的产生方式.

方式1: 针对4个数据集, 按照受污染样本个数占整体数据集规模的1%~100%的概率依次添加高斯噪声, 噪声均值为0, 方差为所干扰数值的30%, 获得100组共400个数据集.

方式 2: 现实生活中噪声往往不是稀疏的, 故对各个数据全体添加高斯噪声, 噪声强度按照信噪比从 40 dB∼0 dB 进行添加 (即每个噪声数据方差按照从所干扰数值的 1%∼100% 顺序增强), 获得 100 组共 400 个数据集.

方式 3: 在没有任何干扰的情况下, 本文方法和变精度粗糙集方法针对 Iris 数据选择的关键属性均为属性 4, 因此选择不同程度噪声干扰下各方法赋予 Iris 数据中属性 4 的权重进行对比. 针对 Iris 数据集, 按照受污染样本个数占整体数据集规模的 1%∼100%的概率依次添加噪声, 噪声幅值按照相对于所干扰数值的 0.01∼1 依次递增, 获得 10 000 组数据集.

2) 实验结果.

采用本文方法和变精度粗糙集两种方法对方式 1、方式 2 噪声数据集选择属性并记录, 其结果对比如表 3 所示.

表 3　含噪数据属性选择效果对比

数据集	干扰方式	变精度粗糙集		本文方法	
		起始出错点	噪声极限点	起始出错点	噪声极限点
Iris	方式 1	3%	8%	无	无
	方式 2	0.01	0.09	0.92	无
Glass	方式 1	6%	无	48%	无
	方式 2	0.03	0.42	0.13	无
Diabets	方式 1	4%	55%	无	无
	方式 2	0.12	0.23	0.16	0.25
Wine	方式 1	6%	19%	无	无
	方式 2	0.04	0.09	0.06	无

表 3 为含噪数据属性选择效果对比. 方式 1 产生的噪声数据的起始出错点指的是开始出错的含噪数据比率; 方式 2 中数据则指的是开始出错的噪声强度. 噪声极限点指的是由于受到噪声影响, 使得属性选择结果近似于随机选择时的噪声强弱程度 (本文中, 噪声数据属性选择结果与正常数据结果不同的属性个数占正常结果 50% 以上时, 视为近似于随机选择). 如表 3 所示, 无论在方式 1 还是方式 2 的噪声干扰下, 本文方法均好于变精度粗糙集方法, 其起始出错点均高于变精度粗糙集, 且在绝大部分含噪数据上没有发生随机选择属性的现象, 抗噪性能明显. 如在方式 1 干扰下, 当 Iris 数据中含有 3% 的噪声点时, 邻域粗糙集开始出错, 上升到 8% 时开始近似于随机选择属性, 而本文方法未出错; 方式 2 产生的噪声明显对属性选择的结果有更强的干扰, 当噪声强度为 0.01 时变精度粗糙集开始出错, 到 0.09 时开始近似于随机选择属性; 而本文方法在噪声强度达到 0.92 时开始出错, 属性选择结果始终没有达到随机选择的错误程度.

图 2 和图 3 为方式 3 噪声数据的实验结果, 即在含噪点的数目、噪声强度两个因素影响下本文方法、变精度粗糙集方法赋予 Iris 数据关键属性 4 的权重变化对比 (粗糙集中属性权重用依赖度衡量, 本文方法中用投影矩阵中对应属性的行向量的 l_2 范数衡量).

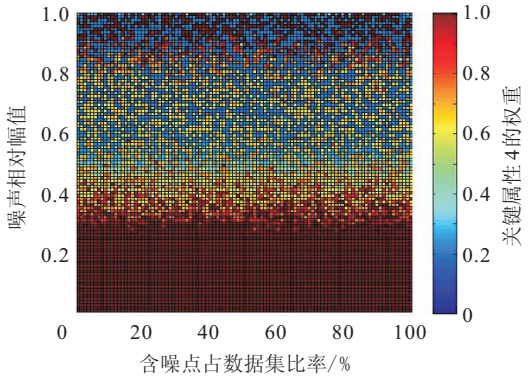


图 2　本文方法获得 Iris 数据关键属性权重变化

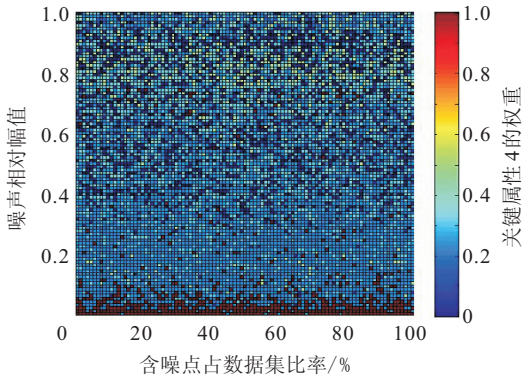


图 3　变精度粗糙集方法获得 Iris 数据关键属性权重变化

图 2 和图 3 中, 横坐标为含噪点数目占整个数据集规模的比例, 其变化范围为 1%∼100%; 纵坐标为噪声的相对强度, 即噪声幅值相对于数值大小的比值, 变化范围为 0.01∼1; 用颜色的冷暖程度来表征赋给关键属性 4 的权重, 变化范围为 0∼1. 由图 2 和图 3 可以看出, 变精度粗糙集在含噪点个数较少及噪声强度较低的情况下尚能选择正确的属性, 在噪声点较少的情况下尽管噪声强度较大也能获得较好的的结果, 但随着干扰强度的加大, 正确关键属性的权重迅速降低; 而本文方法针对噪声的鲁棒性显然要好于变精度粗糙集方法, 在噪声强度小于 0.3 的情况下, 噪声点的数目对关键属性的权重不构成影响, 随着干扰强度的加大, 关键属性的权重始终保持较高.

由以上结果可知, 基于稀疏表示的属性选择方法可以选择出正确的关键属性, 且针对含噪错误数据的鲁棒性要明显好于变精度粗糙集.

4.3　标签错误数据属性选择效果验证

将 4 个数据集按从 1%∼100% 的概率依次随机更改标签, 得到 100 组共 400 个数据集.

用本文方法和变精度粗糙集方法对上述污染数

据集选择属性并进行相关对比. 实验结果如表4和图4、图5所示.

表 4 标签错误数据属性选择效果对比 %

数据集	变精度粗糙集		本文方法	
	起始出错点	噪声极限点	起始出错点	噪声极限点
Iris	8	11	无	无
Glass	6	12.5	17	78
Diabets	8	54	5	无
Wine	13	42	无	无

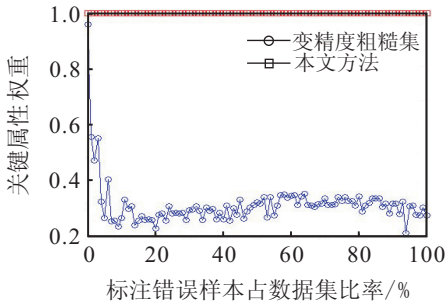


图 4 正确关键属性权重对比 (Iris 数据)

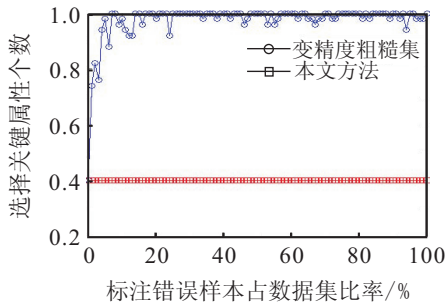


图 5 选择关键属性个数对比 (Iris 数据)

由表4可知, 不管是起始出错点还是噪声极限点标准上, 本文方法均好于变精度粗糙集方法. 如当Iris数据发生8%的标签错误时变精度粗糙集开始出错, 且在11%的标签错误时开始随机选择属性, 而本文方法选择结果一直未出错. 图4和图5为针对Iris数据的属性选择效果对比图. 由图4和图5可知, 随着标签错误数据个数的增加, 变精度粗糙集赋予属性4的属性重要度迅速降低到0.4以下, 选择的关键属性个数随之上升到3个以上, 而本文方法属性重要度及选择属性个数则始终保持稳定. 由结果可知, 本文方法针对标签错误数据的能力也明显好于变精度粗糙集.

5 结 论

为了应对含噪或标签错误数据, 本文提出一种基于最大边界投影和$l_{2,1}$范数正则化的属性选择算法, 该算法主要利用了最大边界投影方法的解空间和稀疏表示的稀疏结构把握能力, 可以较好地解决含噪、标签错误的数据挖掘问题. 但利用$l_{2,1}$范数正则化的方法来获得行稀疏的矩阵仅相当于对单个属性的选择, 未来将考虑到数据中属性组合稀疏以及属性之间结构稀疏的特点, 进行属性组合的选择.

参考文献(References)

[1] 徐怡, 李龙澍. 变精度集对势粗糙集模型[J]. 控制与决策, 2010, 25(11): 1732-1736.
(Xu Y,Li L S. Variable precision rough set model based on set pair situation[J]. Control and Decesion, 2010, 25(11): 1732-1736.)

[2] Wang Jianping, Zhang Damin. Handwritten Chinese character recognition with variable precision rough set approach[C]. Proc of the 2010 Int Conf on Electrical and Control Engineering. Wuhan, 2010: 1108-1111.

[3] Shi Zhong, Wei Tang, Taghi M. Khoshgoftaar. Boosted noise filters for identifying mislabeled data[R]. Boca Raton: Florida Atlantic University, 2005: 383-401.

[4] Wu Xindong, Zhu Xingquan. Mining with noise knowledge: Error-aware data mining[J]. IEEE Trans on Systems, Man, and Cybernetics, Part A: Systems and Humans, 2008, 38(4): 917-932.

[5] Wu T T, Chen Y F, Hastie T, et al. Genome-wide association analysis by Lasso penalized logistic regression[J]. Bioinformatics, 2009, 25(6): 714-721.

[6] Wright J, Ganesh A, Allen Y, et al. Robust face recognition via sparse representation[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2009, 31(2): 210-227.

[7] Zhou H, Hastie T, Tibshirani R. Sparse principle component analysis[R]. Stanford: Stanford University, 2004.

[8] Zhihua Qiao, Lan Zhou, Jianhua Zhuang. Sparse linear discriminant analysis with applications to high dimensional low sample size data[J]. Int J of Applied Mathematics, 2009, 39(1): 6-29.

[9] Deng Cai, Xiaofei He, Jiawei Han. Spectral regression: A unified approach for sparse sub-space learning[C]. Proc of the 7th IEEE Int Conf on Data Mining. Omaha, 2007: 73-82.

[10] Mahdokht Masaeli, Glenn Fung, Jennifer G Dy. From transformation-based dimensionality reduction to feature selection[C]. Proc of the 27th Int Conf on Machine Learning. Haifa, 2010: 751-758.

[11] Xiaofei He, Deng Cai, Jiawei Han. Learning a maximum margin subspace for image retrieval[J]. IEEE Trans on Knowledge and Data Engineering, 2008, 20(2): 189-201.