基于三支标签传播的半监督属性约简

胡声丹[1),2)] 苗夺谦[1),2)] 姚一豫[3)]

[1)](同济大学计算机科学与技术系 上海 201804)

[2)](嵌入式系统与服务计算教育部重点实验室（同济大学） 上海 201804)

[3)](里贾纳大学计算机科学系 里贾纳 S4S 0A2 加拿大)

摘 要 属性约简是粗糙集理论的重要应用之一. 为了对部分标记的数据进行属性约简，一些基于粗糙集的半监督属性约简方法相继被提出，但这些方法在数据信息利用、运行代价、约简质量等方面仍然存在挑战. 本文针对混合型分类数据，提出了一种新的基于三支标签传播的半监督属性约简（3WLPME）方法. 该方法包括两个过程：三支标签传播（3WLP）和基于混合熵的启发式属性约简（MEHAR）. 其中，3WLP 在经典标签传播算法的基础上，结合三支决策和主动学习思想，对无标签数据进行标注，并更新有标签集和无标签集. 迭代执行上述过程直至收敛，可以提升最终的伪标签准确率. 在 MEHAR 中，属性重要度由混合熵度量. 基于依赖度和条件熵定义的混合熵，融合了粗糙集的代数表示和信息表示，能更深刻地反映属性的分类能力. 本文对 3WLP 算法和 MEHAR 算法的有效性进行了理论分析. 在 UCI 数据集上进行了以下仿真实验：3WLP 与随机标签传播在伪标签准确率上的对比；不同属性约简算法在约简质量上的对比；3WLPME 与其他基于粗糙集的半监督属性约简方法，在约简质量上的对比. 实验结果验证了 3WLP 能获得较高的伪标签准确率；MEHAR 在不降低分类准确率的前提下，能获得较小的约简；3WLPME 在半监督约简过程中具有更高的效率和稳定性，说明本文所提方法是有效的.

关键词 半监督学习；属性约简；三支决策；标签传播；邻域粗糙集；混合熵

中图法分类号 TP18 **DOI 号** 10. 11897/SP. J. 1016. 2021.02332

Three-Way Label Propagation Based Semi-Supervised Attribute Reduction

HU Sheng-Dan[1),2)] MIAO Duo-Qian[1),2)] YAO Yi-Yu[3)]

[1)](*Department of Computer Science and Technology, Tongji University, Shanghai* 201804)

[2)](*Key Laboratory of Embedded System and Service Computing (Tongji University), Ministry of Education, Shanghai* 201804)

[3)](*Department of Computer Science, University of Regina, Regina, Saskatchewan, S4S 0A2 Canada*)

Abstract Attribute reduction is one of the most important applications of rough set theory. It obtains attribute subsets by selecting relevant attributes of data. It can reduce the dimension of data and reserve the discernible ability of original attributes. In many real-world applications, data are easily obtained, but annotating all of them is laborious. In order to exploit few labeled and large unlabeled data in attribute reduction, several rough set-based semi-supervised attribute reduction methods have been proposed. However, there still exist some challenges: discretization of numerical attributes may cause data information loss, some methods may have high computational complexity as data volumes increase, the information of both labeled and unlabeled data are not used to the full, etc. Therefore, the reduction

收稿日期：2020-03-09；在线发布日期：2020-08-21. 本课题得到国家自然科学基金项目(61976158，61673301)资助. **胡声丹**，博士研究生，中国计算机学会(CCF)学生会员（69242G），主要研究领域为粗糙集理论、粒计算、机器学习. E-mail: hushengdan@163. com. **苗夺谦（通信作者）**，博士，教授，博士生导师，中国计算机学会(CCF)杰出会员（09187D），主要研究领域为粗糙集、粒计算、模式识别、人工智能等. E-mail: dqmiao@tongji. edu. cn. **姚一豫**，博士，教授，主要研究领域为三支决策、粗糙集、区间集、粒计算、信息检索、Web智能和数据挖掘等.

efficiency needs to be further improved. In this paper, we propose a novel semi-supervised attribute reduction method (3WLPME) for classification problem with mixed (categorical and numerical) attributes. The proposed method contains two procedures: label propagation and then attribute reduction. As for label propagation, we extend the classical label propagation algorithm to three-way label propagation (3WLP). In 3WLP, the pseudo labels of unlabeled data are firstly learned by two classical label propagation processes with diverse neighborhood size *k*. Then three-way decision acts as a decision strategy to choose unlabeled data with different pseudo labels for annotating. Finally, the labeled data subset and unlabeled data subset are updated. All the three processes mentioned above are executed iteratively until convergency. After the procedure of 3WLP, all the data are labeled and the ultimate pseudo label accuracy of unlabeled data may be higher. In term of attribute reduction, a heuristic attribute reduction algorithm called MEHAR based on mix entropy is proposed. It is applied to all the labeled data (true labels and pseudo labels) to reduce their dimension. In MEHAR, the heuristic information is attribute significance, and which is defined by mix entropy. Mix entropy is an integration of dependency degree and conditional entropy. As dependency degree measures the classification ability of an attribute from the algebra view of rough set, and conditional entropy measures the average uncertainty of a decision system from the information view of rough set, combination of them can reflect classification ability of attributes more deeply. Effectiveness of the proposed 3WLP algorithm and MEHAR algorithm are discussed theoretically. Several simulation experiments have also been conducted on eight selected UCI datasets, and they include: (1) the comparisons of the pseudo label accuracy between 3WLP and random label propagation, (2) the comparisons of the reduction quality between different attribute reduction algorithms, namely, positive field-based, conditional entropy-based, discernibility matrix-based and mix entropy-based (MEHAR), (3) the comparisons of the reduction quality between the proposed method 3WLPME and three other rough set-based semi-supervised attribute reduction methods. The experimental results of (1) show the superiority of 3WLP and indicate that it can be a novel way to annotate data. The experimental results of (2) show that the number of attributes in MEHAR-based reduction is less without reducing the classification accuracy. The experimental results of (3) show the efficiency and stability of 3WLPME. So, the proposed semi-supervised attribute reduction method is effective.

Keywords　semi-supervised learning; attribute reduction; three-way decision; label propagation; neighborhood rough set; mix entropy

1　引　言

特征选择（feature selection）是从数据的原始特征中删除无关、冗余的特征，选取最有效的特征子集以降低数据的维度，既可以减少后续存储、计算开销，又可以防止模型过拟合、提高模型泛化能力，是机器学习、模式识别、数据挖掘、大数据等领域中关键的数据预处理步骤[1-3]. 根据特征选择过程是否依赖于学习器，通常将特征选择方法分为三类：过滤式（filter）[4-6]、包裹式（wrapper）[7]和嵌入式（embedded）[8]，其中过滤式方法先依据评价准则选取特征子集，再训练学习器，即特征选择过程独立于学习器；包裹式方法直接使用学习器的性能来评价选取的特征；嵌入式方法将特征选择嵌入到学习器构造的目标函数中，特征选择与学习器在同一优化过程中完成. Pawlak 经典粗糙集（Rough Sets）[9]中的属性约简（attribute reduction）能在保持决策表条件属性与决策属性间依赖关系的条件下，删除条件属性集中冗余的属性或属性值，是一种过滤式特征选择方法. 自粗糙集理论提出以来，属性约简已得到众多学者的研究[10-12].

此外，在许多机器学习任务中，很容易获取大量无标签样本数据，而对所有数据进行标注，需要的人力、时间等成本较高，半监督学习（semi-supervised learning）旨在使用有标签数据和大量的无标签数据构建更优的学习器[13]. 半监督学习的学

习场景有半监督分类、半监督聚类、半监督回归和半监督降维等[14]，半监督特征选择作为一类半监督降维方法，是在特征选择过程中，既使用有标签数据的标签信息，又使用所有有标签数据和无标签数据的结构信息来评价特征的相关性[15].

近年来，基于粗糙集理论研究半监督属性约简得到了关注[16-18]. Zhang[16]提出的半监督属性约简方法结合了协同学习理论，在有标签数据集上构造两个差异性较大的属性约简，利用约简后的有标签数据构造两个基分类器，再在无标签数据上进行交互协同学习，增量式扩大有标签数据集并获得新的约简，直到所有无标签数据被加入到有标签数据集. Dai[17]提出的半监督属性选择方法给出了辨识对的概念，并针对部分标记数据，定义了基于辨识对的属性重要度，利用该属性重要度作为启发信息进行属性选择. Liu[18]提出的半监督特征选择方法包含了标签传播和特征选择两个过程，在特征选择时，利用类局部邻域决策误差率构造多个属性适应函数，并结合集成学习思想，基于属性适应函数评价属性重要性，进而进行启发式特征选择. 但是，现有的方法在进行属性选择时，或多或少地面临以下这些挑战：连续属性的离散化或删除会丢失数据的结构信息[16-17]，基于辨识矩阵[16]、辨识对[17]的属性约简在数据量增加时运行代价显著提高，约简质量较低[18]等.

针对半监督属性约简问题，本文基于混合型分类数据，提出基于三支标签传播的半监督属性约简方法. 首先利用三支标签传播算法（3WLP）为无标签数据打上伪标签，得到包含伪标签的数据集，再对包含伪标签的数据集以邻域混合熵作为启发条件进行属性约简（MEHAR），得到只包含特征子集的数据集. 该方法的主要贡献有：

（1）使用 Gower 距离度量混合型数据间的距离. 首先统计数据各属性的分布信息，再在 k 近邻和邻域粒计算过程中采用 Gower 距离度量数据间的距离.

（2）提出三支标签传播算法（3WLP）. 该算法结合三支决策和主动学习思想，基于两个不同 k 值的相似度矩阵，使用两个 kNN 标签传播算法对无标签数据分别进行标注，对两算法标注不一致的无标签数据进行人工标注，并将其加到原来的有标签数据集中，迭代执行该过程直至收敛，能有效提高伪标签准确率.

（3）提出基于邻域混合熵的启发式属性约简算法（MEHAR）. 该算法中属性重要度由混合熵度量，而混合熵的定义融合了依赖度和条件熵. 依赖度从粗糙集的代数观来度量属性的分类能力，条件熵从粗糙集的信息观来度量决策系统的平均不确定性，因此，融合两者的混合熵可以更深刻地反映属性的分类能力.

本文第 2 节简单介绍三支决策、标签传播、邻域粗糙集等背景知识，第 3 节详细介绍本文所提的基于三支标签传播的半监督属性约简方法，第 4 节通过实验对本文所提算法及方法进行测试并对结果进行比较分析，最后第 5 节对本文的工作进行总结并指出进一步研究的方向.

2 基本概念

2.1 三支决策

三支决策[19]是加拿大学者 Yao 教授在研究概率粗糙集的概率阈值（α, β）的估值时所提出的，其基本思想是将整体划分成三个独立的部分，针对不同的部分采取不同的处理策略. 三支决策反映了人类处理信息的模式，是一种求解复杂问题的有效策略和方法. 随着越来越多学者对三支决策在理论、方法、应用等方面的研究，三支决策理论已然成为一个更普适的理论框架[20]，并扩展到三支概念分析、三支聚类、三支分类、三支属性约简、三支决策支持等领域[21].

三支决策形式化描述如下：

设 U 为有限、非空数据集，在条件集 C 下，对 $\forall x \in U$，基于评价函数 $v(x)$和阈值对 (α,β) $(0 \leqslant \beta \leqslant \alpha \leqslant 1)$，将 U 划分为三个两两不相交的域，即正域、负域和边界域，记为 $POS_{(\alpha,\beta)}(C)$、$NEG_{(\alpha,\beta)}(C)$、$BND_{(\alpha,\beta)}(C)$，相对应的三支决策规则如下：

（1）若 $v(x) \geqslant \alpha$，则 $x \in POS_{(\alpha,\beta)}(C)$，接受 $x \in C$.

（2）若 $v(x) \leqslant \beta$，则 $x \in NEG_{(\alpha,\beta)}(C)$，拒绝 $x \in C$.

（3）若 $\beta < v(x) < \alpha$，则 $x \in BND_{(\alpha,\beta)}(C)$，不承诺 $x \in C$.

2.2 标签传播算法

标签传播算法是一种基于图的半监督学习方法[22]，该算法利用数据间的关系建立图模型，图中的节点表示所有数据，包括有标签数据和无标签数据；图中的边表示节点间的相似度，相似度越大，标签越容易传播. 节点的标签定义为类上的概率分布，标签传播过程中，固定有标签节点的标签不变，利用其来预测无标签节点的标签信息，通过迭代，图中所有节点的标签趋于稳定.

给定数据集 $U=\{x_1,x_2,\cdots,x_n\}$，该数据集分为有标签数据集 $U_L=\{x_1,x_2,\cdots,x_l\}$ 和无标签数据集 $U_U=\{x_{l+1},x_{l+2},\cdots,x_{l+u}\}$，通常 $l\ll u$，且 $l+u=n$. 设数据集的类别标签为 $y_i\in\{1,2,\cdots,c\},i=1,2,\cdots,l$，其中类别数 c 已知，并假定有标签数据集的标签包括所有 c 类标签. 标签传播算法的目的是通过数据集 $\{<x_1,y_1>,<x_2,y_2>,\cdots,<x_l,y_l>\}$ 学习无标签数据的标签 $\{y_{l+1},y_{l+2},\cdots,y_{l+u}\}$.

建立一个全连接图 $G=\{V,E,W\}$，其中 $V=U=\{x_1,x_2,\cdots,x_n\}$ 为顶点集，$E\subseteq V\times V$ 为边集，相似矩阵 $\boldsymbol{W}=(w_{ij})_{n\times n}$ 度量图中任意两节点之间的相似度，且

$$w_{ij}=\exp\left(-\frac{(\|x_i-x_j\|_2)^2}{\sigma^2}\right) \tag{1}$$

$\|x_i-x_j\|_2$ 为节点 x_i 和 x_j 对应的向量间的 2 范数，σ 为超参数.

根据相似矩阵 $\boldsymbol{W}$ 定义概率转移矩阵 $\boldsymbol{P}=(P_{ij})_{n\times n}$，且

$$P_{ij}=\boldsymbol{P}(j\to i)=\frac{w_{ij}}{\sum_{k=1}^{n}w_{kj}} \tag{2}$$

其中 P_{ij} 表示节点 x_j 到 x_i 的传播概率.

定义一个标签矩阵 $\boldsymbol{Y}=[\boldsymbol{Y}_L;\boldsymbol{Y}_U]=(y_{ij})_{n\times c}$，其中 $y_{ij}=\delta(y_i,j)$ 表示节点 x_i 的标签为 j 的概率，$\boldsymbol{Y}_L$、$\boldsymbol{Y}_U$ 分别表示有标签数据集、无标签数据集的标签矩阵. 对于 $\boldsymbol{Y}_L$，

$$y_{ij}=\begin{cases}1, & y_i=j\\ 0, & \text{others}\end{cases}\quad(i=1,2,\cdots,l) \tag{3}$$

而 $\boldsymbol{Y}_U$ 的初始值可取任意值，但需满足 $\sum_{j=1}^{c}\delta(y_i,j)=1,i\in\{l+1,l+2,\cdots l+u\}$.

依据 $\boldsymbol{Y}=\boldsymbol{P}\boldsymbol{Y}$ 进行标签传播，并迭代直至 $\boldsymbol{Y}$ 收敛. 在标签传播过程中，需要保持有标签数据的标签源 $\boldsymbol{Y}_L$ 不变.

基于收敛后的标签矩阵 $\boldsymbol{Y}$，获得无标签数据 x_i 的标签为

$$y_i^*=\arg\max_j(y_{ij}) \tag{4}$$

其中 $i\in\{l+1,l+2,\cdots,l+u\},j\in\{1,2,\cdots,c\}$.

2.3　邻域粗糙集

基于邻域模型[23]提出的邻域粗糙集模型[24]使用实数空间中点的δ邻域来粒化论域，进而基于邻域关系粒处理连续性数据.

定义 1. 一个信息系统可以表示为一个四元组 $IS=(U,A,V,f)$，其中：

（1）$U=\{x_1,x_2,\cdots,x_n\}$ 是非空数据对象的集合，称为论域，且 $|U|=n$.

（2）$A=\{a_1,a_2,\cdots,a_m\}$ 是非空属性的集合，称为属性集，且 $|A|=m$.

（3）$V=\bigcup_{a\in A}V_a$ 是所有属性可能取值的集合，称为值域.

（4）$f:U\times A\to V$ 是信息函数的集合，使得对 $\forall x\in U$，$a\in A$，$f(x,a)\in V_a$.

若信息系统 IS 中的属性集 $A=C\cup D$，其中 C 为条件属性集，D 为决策属性集，且 $C\cap D=\varnothing$，则称 $DS=(U,A=C\cup D,V,f)$ 为决策信息系统.

定义 2. 设 $DS=(U,A=C\cup D,V,f)$ 是一个决策系统，对 $\forall x_i\in U$，其关于 R（$R\subseteq C$）的 δ 邻域定义为

$$N_R^{\delta}(x_i)=\{x_j\in U\mid\Delta_R(x_i,x_j)\leqslant\delta\} \tag{5}$$

其中 $\delta\geqslant 0$，$\Delta_R(x_i,x_j)$ 定义为对象 x_i 和 x_j 在属性集 R 上的距离测度，且

$$\Delta_R(x_i,x_j)=\left(\sum_{a\in R}|f(x_i,a)-f(x_j,a)|^p\right)^{1/p} \tag{6}$$

公式(6)中，p=1时，Δ为曼哈顿距离，p=2时，Δ为欧氏距离，$p\to\infty$时，Δ为切比雪夫距离.

定义 3. 决策系统 $DS=(U,A=C\cup D,V,f)$ 中，由 R（$R\subseteq C$）和参数 δ（$\delta\geqslant 0$）形成的邻域关系定义为

$$NR_R^{\delta}=\{(x_i,x_j)\in U\times U\mid\Delta_R(x_i,x_j)\leqslant\delta\} \tag{7}$$

邻域决策系统为 $NDS=(U,C\cup D,V,f,\delta)$.

定义 4. 设 $NDS=(U,C\cup D,V,f,\delta)$ 是一个邻域决策系统，对 $\forall X\subseteq U$，其关于 R（$R\subseteq C$）的 δ 邻域下近似、上近似分别定义为

$$\underline{N_R^{\delta}}(X)=\{x\in U\mid N_R^{\delta}(x)\subseteq X\} \tag{8}$$

$$\overline{N_R^{\delta}}(X)=\{x\in U\mid N_R^{\delta}(x)\cap X\neq\varnothing\} \tag{9}$$

若在决策属性集 D 下，对象集被划分为 r 个互不相交的子集 $\{D_1,D_2,\cdots,D_r\}$，则 D 关于 R 的 δ 邻域下近似、上近似分别为

$$\underline{N_R^{\delta}}(D)=\bigcup_{i=1}^{r}\underline{N_R^{\delta}}(D_i) \tag{10}$$

$$\overline{N_R^{\delta}}(D)=\bigcup_{i=1}^{r}\overline{N_R^{\delta}}(D_i) \tag{11}$$

X 关于 R 的邻域下近似 $\underline{N_R^{\delta}}(X)$ 也称为正域 $POS_R^{\delta}(X)$，由那些根据属性 R 判断肯定属于 X 的邻域粒组成；负域 $NEG_R^{\delta}(X)=U-\overline{N_R^{\delta}}(X)$，由那些根据属性 R 判断肯定不属于 X 的邻域粒组成；边界域 $BN_R^{\delta}(X)=\overline{N_R^{\delta}}(X)-\underline{N_R^{\delta}}(X)$，由那些根据属性 R 判断不能肯定属于 X，也不能肯定不属于 X 的邻域粒组成.

3 半监督属性约简方法

针对部分有标签、混合型分类数据的约简，本文提出基于三支标签传播的半监督属性约简（3WLPME）方法，主要包括两个过程：（1）利用三支标签传播算法（3WLP）为无标签数据打上伪标签，得到包含伪标签的数据集；（2）以基于混合熵的属性重要度作为启发条件，对包含伪标签的数据简（MEHAR），得到只包含特征子集的数据集. 3WLPME 流程框架如图 1 所示：

3.1 混合数据距离度量

混合型数据包括连续型数据、名义型数据和顺序型数据等，为充分利用所有数据的统计信息，对名义型数据采用 one-hot 编码，并对数据进行标准化处理. 样本数据距离计算采用 Gower 距离[25]，其定义如下：

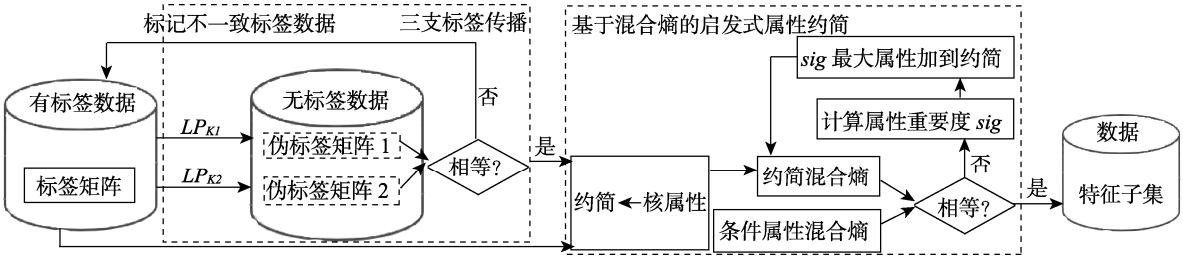

图 1 基于三支标签传播的半监督属性约简流程

$$\Delta(x_i,x_j)=\frac{\sum_{k=1}^{m}w_{ijk}\Delta(x_i^k,x_j^k)}{\sum_{k=1}^{m}w_{ijk}} \tag{12}$$

其中 $\Delta(x_i^k,x_j^k)$ 为数据对象 x_i 和 x_j 在第 k 个属性上的距离测度，w_{ijk} 为第 k 个属性的权重.

假设数据集中所有条件属性权重相等，则

$$\begin{aligned}\Delta(x_i,x_j)&=\frac{1}{m}\sum_{k=1}^{m}\Delta(x_i^k,x_j^k)\\&=\frac{1}{m}\Big(\sum_{a\in C}|f(x_i,a)-f(x_j,a)|^p\Big)^{1/p}\end{aligned} \tag{13}$$

通常 p 取 1 或 2.

3.2 三支标签传播算法

基于全连接图的标签传播算法需要存储和计算所有节点对间的相似度，算法收敛速度较慢，因此常采用 kNN 图[26]构建相似度的稀疏矩阵，进而进行标签传播，如算法 1 所示.

算法 1. kNN 标签传播算法（LP_k）.

输入：l 个带标签数据，u 个无标签数据，近邻系数 k

输出：无标签数据的伪标签

1. 依据公式(13)和近邻系数 k 构建概率转移矩阵 $\boldsymbol{P}$；
2. 设置标签矩阵 $\boldsymbol{Y}^0=[\boldsymbol{Y}_L;\boldsymbol{1}/\boldsymbol{c}]$；
3. 设置收敛条件 ε，$t_{\max}$；
4. 初始迭代次数 $t=0$；
5. 执行 $\boldsymbol{Y}^{t+1}=\boldsymbol{P}\boldsymbol{Y}^t$；
6. 重置已标记数据的标签 $\boldsymbol{Y}^{t+1}\leftarrow[\boldsymbol{Y}_L;\boldsymbol{Y}_U^{t+1}]$；
7. 更新迭代次数 t=t+1；
8. 判断收敛条件，当 $\|\boldsymbol{Y}^{t+1}-\boldsymbol{Y}^t\|_2>\varepsilon$ 且 $t<t_{\max}$，转步骤 5；
9. 依据公式⑷返回无标签数据的伪标签 $Y_U^{LP_k}$，算法结束.

但算法 1 中近邻系数 k 的选取对学习效果影响较大[27]：当 k 值过小时，可能不能充分反映流行的局部几何结构，如图 2(a)所示，带圈节点的近邻节点定义为距离其最近的两个节点，此时不能充分反映数据的局部分布信息；而当 k 值过大时又可能会引入流行结构以外的边，如图 2(b)所示，带圈节点的近邻节点定义为距离其最近的六个节点，错误的类别信息会通过标签传播算法传播过来. 另一方面，当图中存在桥节点（bridge points）时，如图 2(c)中的带圈节点，会引起标签的错误传播[28]. 因此标签传播算法的准确率有待进一步提高.

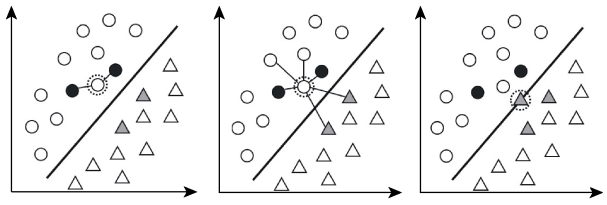

图 2 不同近邻系数及桥节点示例

三支标签传播算法结合三支决策和主动学习思想，使用两个不同 k 值的 kNN 标签传播算法协同工作，若两个算法对某一无标签数据给出的伪标签不一致，说明两者对该数据对象的所属类别意见不一，则人工给出该对象的正确标签，并将其加到原来的有标签数据集中，迭代执行该过程，直到两个标签传播算法为所有数据对象给出的伪标签信息一致.

设两个 kNN 标签传播算法分别为 LP_{k1} 和 LP_{k2}

（$k1\neq k2$），依据 LP_{k1} 和 LP_{k2} 对无标签数据集 U_U 给出的伪标签，将 U_U 划分成三个互不相交的域 U_{U1}、U_{U2}、U_{U3}，且：

$$U_{U1}=\{x_i\in U_U \mid y_i^{LP_{k1}}=y_i^{LP_{k2}}=y_i\} \tag{14}$$

$$U_{U2}=\{x_i\in U_U \mid y_i^{LP_{k1}}=y_i^{LP_{k2}}\neq y_i\} \tag{15}$$

$$U_{U3}=\{x_i\in U_U \mid y_i^{LP_{k1}}\neq y_i^{LP_{k2}}\} \tag{16}$$

其中 $y_i^{LP_{k1}}$、$y_i^{LP_{k2}}$ 分别表示标签传播算法 LP_{k1} 和 LP_{k2} 为 x_i 打上的伪标签，y_i 表示 x_i 的真实标签.

直观上解释为：基于两个标签传播算法对无标签数据集分类，U_{U1} 包括同时分对的数据对象，U_{U2} 包括同时分错且错误相同的数据对象，而 U_{U3} 包括两算法分的结果不一致的对象.

三支标签传播算法不区分 U_{U1} 和 U_{U2}，但对于最不确定的 U_{U3}，人工给出该部分对象的正确标签，并将其加到原来的标签数据集 U_L 中，迭代执行该过程，直到 LP_{k1} 和 LP_{k2} 为所有无标签数据对象给出的伪标签信息一致，从而可以提高伪标签的准确率.

算法 2.　三支标签传播算法（3WLP）.

输入：l 个带标签数据，u 个无标签数据

输出：无标签数据的伪标签

1. 设置近邻系数 $k1, k2$.

2. 分别执行 kNN 标签传播算法 LP_{k1}、LP_{k2} 得到 $Y_U^{LP_{k1}}$、$Y_U^{LP_{k2}}$.

3. 判断收敛条件，当 $Y_U^{LP_{k1}}=Y_U^{LP_{k2}}$ 时，转步骤 5.

4. 对 $\forall x_i\in U_U$，如果 $y_i^{LP_{k1}}\neq y_i^{LP_{k2}}$，标注 x_i，并执行 $U_L=U_L\bigcup\{x_i\}$，$U_U=U_U-\{x_i\}$，转步骤 2.

5. 返回无标签数据的伪标签 Y_U，算法结束.

定理 1.　基于三支标签传播算法，无标签数据集的伪标签矩阵收敛.

证明.　首先证明基于 kNN 标签传播算法，无标签数据集的伪标签矩阵收敛：

将概率转移矩阵依据有标签数据集和无标签数据集分成 4 个子矩阵：

$$\boldsymbol{P}=\begin{bmatrix}\boldsymbol{P}_{LL} & \boldsymbol{P}_{LU}\\ \boldsymbol{P}_{UL} & \boldsymbol{P}_{UU}\end{bmatrix},$$

则

$$\begin{aligned}\boldsymbol{Y}_U^{t+1}&=\boldsymbol{P}_{UL}\boldsymbol{Y}_L+\boldsymbol{P}_{UU}\boldsymbol{Y}_U^t\\&=(\mathrm{I}+\boldsymbol{P}_{UU}+\cdots+(\boldsymbol{P}_{UU})^t)\boldsymbol{P}_{UL}\boldsymbol{Y}_L+(\boldsymbol{P}_{UU})^{t+1}\boldsymbol{Y}_U^0.\end{aligned}$$

因为 $\boldsymbol{P}$ 为行归一化矩阵，$\boldsymbol{P}_{UU}$ 为 $\boldsymbol{P}$ 的子矩阵，所以

$$\lim_{t\to\infty}(\boldsymbol{P}_{UU})^t=0,$$

进而

$$\lim_{t\to\infty}\boldsymbol{Y}_U^t=(\mathrm{I}-\boldsymbol{P}_{UU})^{-1}\boldsymbol{P}_{UL}\boldsymbol{Y}_L,$$

即无标签数据集的标签矩阵收敛于 $(\boldsymbol{I}-\boldsymbol{P}_{UU})^{-1}\boldsymbol{P}_{UL}\boldsymbol{Y}_L$.

其次，基于两个 kNN 标签传播算法 LP_{k1} 和 LP_{k2} 实现的三支标签传播算法中，考虑最不确定情况，即对 $\forall x_i\in U_U$，都有 $y_i^{LP_{k1}}\neq y_i^{LP_{k2}}$，则该算法的迭代次数最大，为 $|U_U|$，此时所有无标签数据的标签均由人工给出，算法收敛，所以无标签数据集的伪标签矩阵收敛.

证毕.

3.3　基于混合熵的启发式属性约简算法

粗糙集中属性约简的常用方法有：基于正域的方法[9]、基于辨识矩阵的方法[29]和基于信息论的方法[10]. 由于在决策表中，求所有约简和最小约简是 NP-hard 问题，通常以属性重要度作为启发信息进行启发式属性约简. 其中，基于正域的启发式约简算法通常使用属性依赖度计算属性重要度，基于信息论的启发式约简算法通常使用熵计算属性重要度. 前者从知识的分类保持能力上，度量属性对论域对象正确分类的影响；后者从系统平均不确定性上，度量属性对论域对象不确定分类的影响，因此两者在度量属性的重要性上具有较强的互补性[11]，结合两者可以更深入地反映属性的分类能力[12].

下面给出邻域决策系统中混合熵的定义，及基于混合熵的属性约简.

定义 5.　设 $NDS=(U,C\bigcup D,V,f,\delta)$ 是一个邻域决策系统，在决策属性集 D 下，对象集被划分为 m 个互不相交的子集 $\{D_1,D_2,\cdots,D_m\}$，在条件属性集 B（$B\subseteq C$）和参数 δ 下，对象集形成 n（$|U|=n$）个邻域子集 $\left\{B_1^\delta,B_2^\delta,\cdots,B_n^\delta\right\}$，$D$ 相对于 B 的混合熵定义为

$$ME^\delta(D\mid B)=-\log_2^{\gamma_B^\delta(D)}H^\delta(D\mid B) \tag{17}$$

其中

$$\gamma_B^\delta(D)=\frac{\sum_{j=1}^{m}|\underline{N_B^\delta}(D_j)|}{|U|} \tag{18}$$

为决策属性 D 对条件属性 B 的依赖度，

$$H^\delta(D\mid B)=-\sum_{i=1}^{n}\frac{|B_i^\delta|}{|U|}\sum_{j=1}^{m}\frac{|B_i^\delta\cap D_j|}{|B_i^\delta|}\log_2\frac{|B_i^\delta\cap D_j|}{|B_i^\delta|} \tag{19}$$

为决策属性 D 相对条件属性 B 的条件熵.

由定义 5 可知，混合熵的定义结合了依赖度和条件熵，其中依赖度 $\gamma_B^\delta(D)$ 反映了基于条件属性 B，能将对象正确分到依据 D 的分类的对象比例；条件熵 $H^\delta(D\mid B)$ 反映了基于条件属性 B，依据 D 的分类的不确定性.

定理 2.　设 $NDS=(U,C\bigcup D,V,f,\delta)$ 是一个邻域决策系统，若 $\boldsymbol{P}\subseteq Q\subseteq C$，则有 $ME^\delta(D\mid\boldsymbol{P})\geqslant ME^\delta(D\mid Q)$.

证明.　当 $\boldsymbol{P}\subseteq Q$ 时，由定义 2 可得出 $N_{\boldsymbol{P}}^\delta(x)\supseteq$

$N_Q^\delta(x)$，则 $\underline{N_{\boldsymbol{P}}^\delta}(D)\subseteq\underline{N_Q^\delta}(D)$，$0\leqslant\gamma_{\boldsymbol{P}}^\delta(D)\leqslant\gamma_Q^\delta(D)\leqslant 1$，所以 $-\log_2^{\gamma_{\boldsymbol{P}}^\delta(D)}\geqslant-\log_2^{\gamma_Q^\delta(D)}\geqslant 0$.

又因为 P 和 Q 形成的邻域子集满足 $P_i^\delta\supseteq Q_i^\delta$，根据文献[30]中的定理 2 可得：

$$\begin{aligned}&H^\delta(D|\boldsymbol{P})\text{-}H^\delta(D|Q)\\&=-\sum_{i=1}^{n}\frac{|P_i^\delta|}{|U|}\sum_{j=1}^{m}\frac{|P_i^\delta\cap D_j|}{|P_i^\delta|}\log_2\frac{|P_i^\delta\cap D_j|}{|P_i^\delta|}\\&\quad+\sum_{i=1}^{n}\frac{|Q_i^\delta|}{|U|}\sum_{j=1}^{m}\frac{|Q_i^\delta\cap D_j|}{|Q_i^\delta|}\log_2\frac{|Q_i^\delta\cap D_j|}{|Q_i^\delta|}\\&\geqslant 0,\end{aligned}$$

即 $H^\delta(D|\boldsymbol{P})\geqslant H^\delta(D|Q)$.

所以 $-\log_2^{\gamma_{\boldsymbol{P}}^\delta(D)}H^\delta(D|\boldsymbol{P})\geqslant-\log_2^{\gamma_Q^\delta(D)}H^\delta(D|Q)$，即 $ME^\delta(D|\boldsymbol{P})\geqslant ME^\delta(D|Q)$.

证毕.

依据依赖度和条件熵的单调性，定理 2 的证明过程说明混合熵具有单调性，即随着条件属性的增加，基于条件属性集的混合熵会减小. 显然，混合熵具有非负性，因此可以利用混合熵计算属性重要度，既能度量属性对论域对象正确分类的影响，又能度量属性对论域对象不确定分类的影响，从更深层次上刻画属性的分类能力.

定义 6. 设 $NDS=(U,C\cup D,V,f,\delta)$ 是一个邻域决策系统且 $B\subseteq C$，对 $\forall a\in C-B$，a 相对于 B 和 D 的属性重要度定义为：

$$sig^\delta(a,B,D)=ME^\delta(D|B)-ME^\delta(D|B\cup\{a\})\quad(20)$$

基于混合熵的相对属性约简融合了属性约简的代数表示和信息表示，其定义为：

定义 7. 设 $NDS=(U,C\cup D,V,f,\delta)$ 是一个邻域决策系统，若 $ME^\delta(D|B)=ME^\delta(D|C)$（$B\subseteq C$），且对 $\forall a\in B$，$ME^\delta(D|B-\{a\})>ME^\delta(D|C)$，称 B 为邻域决策系统的一个相对属性约简.

基于混合熵的属性约简算法具体步骤如下：

算法 3. 基于混合熵的启发式属性约简算法（MEHAR）.

输入：邻域决策系统 $NDS=(U,C\cup D,V,f,\delta)$

输出：约简 B

1. 初始化，令 $B=\varnothing$；

2. 对 $\forall a\in C$，计算 $POS_{C-\{a\}}^\delta(D)$，当 $POS_{C-\{a\}}^\delta(D)\neq POS_C^\delta(D)$ 时，$B\leftarrow B\cup\{a\}$；

3. 依公式(17)分别计算 $ME^\delta(D|C)$ 和 $ME^\delta(D|B)$，若 $B=\varnothing$，$ME^\delta(D|B)\leftarrow+\infty$；

4. 判断收敛条件，当 $ME^\delta(D|B)-ME^\delta(D|C)=0$，转步骤 7；

5. 对 $\forall a\in C\text{-}B$，依据公式(20)计算属性重要度 $sig^\delta(a,B,D)$；

6. $B=B\cup\{\arg\max_a(sig^\delta(a,B,D))\}$，转步骤 4；

7. 返回约简 B，算法结束.

4 实验与分析

实验环境为 Intel 酷睿 i5 四核处理器、8GB 内存、Ubuntu 16. 04 LTS 操作系统，编程语言及环境为 MATLAB 2017a. 实验将从以下三个方面进行：（1）在伪标签准确率上，对三支标签传播算法和随机标签传播算法进行比较分析；（2）在约简大小、运行时间、分类准确率等方面，对基于混合熵的属性约简算法与基于正域、基于条件熵、基于辨识矩阵的属性约简算法进行比较分析；（3）在约简大小、运行时间、分类准确率等方面，对本文提出的半监督属性约简方法与其他 3 个基于粗糙集的半监督属性约简方法进行比较分析.

4.1 数据集及预处理

实验选用了 UCI 机器学习数据库中的 8 个数据集①，对数据集进行下列预处理操作：删除存在缺失数据的数据对象，对离散属性数据进行 one-hot 编码，得到的数据集详细信息如表 1 所示.

表 1 实验数据集

数据集	对象数	条件属性数	决策属性值个数
zoo	101	31	7
lymphography	148	60	4
dermatology	358	131	6
credit	653	46	2
yeast	1484	8	10
steel	1941	27	7
image	2310	19	7
seismic	2584	24	2

分别采用 z-score、min-max 标准化方法对数据进行标准化处理，在计算对象距离时依据公式(13)，分别取 p=1 和 p=2，即曼哈顿距离（M）和欧氏距离（E），采用算法 1（k=5）对标记率为 0.1 的数据集进行标签传播，伪标签的准确率如表 2 所示. 实验中初始有标签数据为随机选取，并保证每类有数据被标注，结果取 10 次的平均值.

① The UC Irvine Machine Learning Repository (UCI), http://archive.ics. uci. edu/ml/index. php

表 2　不同标准化及距离测度方式下伪标签准确率（ratio=0.1）

数据集	z-score/M	z-score/E	min-max/M	min-max/E
zoo	**0.7934±0.0811**	0.7538±0.0819	0.7825±0.0593	0.7912±0.0817
lymphography	0.6073±0.1119	0.5925±0.0887	0.6107±0.1092	**0.6157±0.0611**
dermatology	**0.8654±0.0368**	0.8000±0.0599	0.8461±0.0443	0.8393±0.0466
credit	0.6817±0.0433	0.6372±0.0474	**0.7026±0.0349**	0.7003±0.0334
yeast	**0.4794±0.0309**	0.4781±0.0335	0.4763±0.0316	0.4754±0.0255
steel	**0.6126±0.0178**	0.6104±0.0173	0.6042±0.0147	0.5890±0.0233
image	0.8359±0.0152	0.8289±0.0164	**0.8490±0.0210**	0.8253±0.0188
seismic	**0.8613±0.0176**	0.8352±0.0201	0.8445±0.0186	0.8328±0.0134

表 2 表明有 5 个数据集在采用 z-score/M 方式，即 z-score 标准化和曼哈顿距离时，平均伪标签准确率高于其它三种方式,本文后续实验均采用 z-score/M 方式.

4.2　标签传播对比实验

在三支标签传播算法（3WLP）中，依据三支决策思想，对两个 kNN 标签传播算法 LP_{k1} 和 LP_{k2} 标记不一致的对象进行了标注，为验证三支标签传播算法的有效性，实验比较了采用三支决策标注无标签数据（$3WLP_{k1}$、$3WLP_{k2}$）和随机标注无标签数据（$RANLP_{k1}$、$RANLP_{k2}$）的伪标签准确率，取近邻系数 $k1$=5、$k2$=9，标记率为 0.1，在 8 个数据集上按标记率随机选取初始数据并进行标签传播，伪标签准确率如图 3 所示.

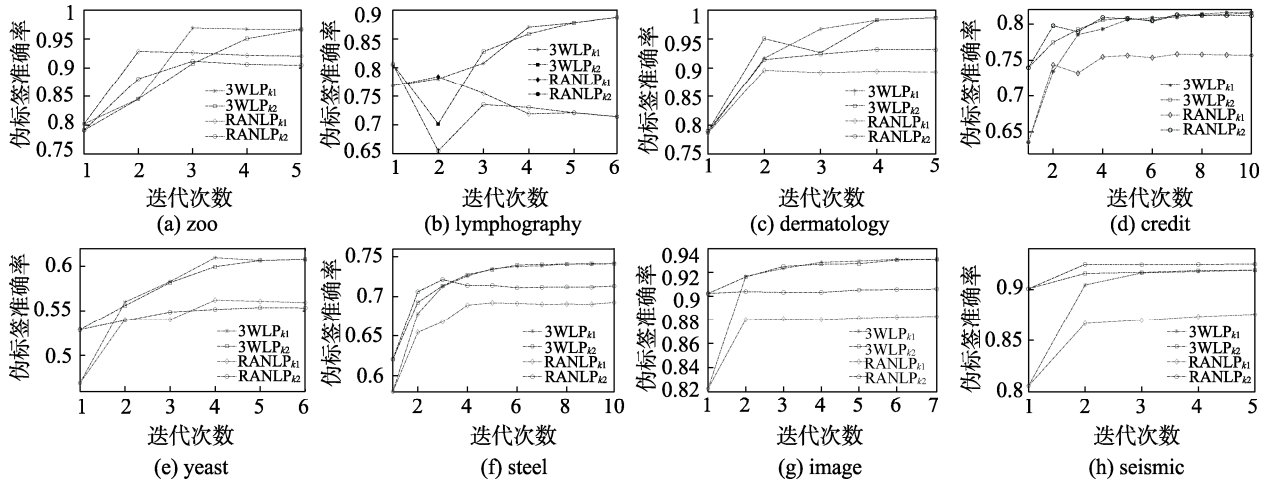

图 3　三支决策标注无标签数据和随机标注无标签数据的伪标签准确率对比

从图 3 中可以看出，随着迭代次数的增加，利用三支标签传播算法标注数据，伪标签准确率在 credit、steel、image 和 seismic 数据集上是单调递增的，在另外四个数据集上存在个别降低的情况，如 zoo 数据集上 $3WLP_{k1}$ 的第 4 次迭代，lymphography 数据集上 $3WLP_{k2}$ 的第 2 次迭代，dermatology 数据集上 $3WLP_{k2}$ 的第 3 次迭代. 最终收敛时，除 seismic 数据集外，三支标签传播算法标注的伪标签准确率均大于随机标签传播算法. 此外，随机标签传播算法的伪标签准确率在迭代过程中会出现单调递减的情况，如数据集 lymphography 和 steel 中，说明在随机标注过程中，可能聚集越来越多的处在分类边界附近或异常的数据点，这类数据的标签在传播过程中会造成其近邻数据点标签的误差，随着标签传播迭代的进行，误差逐渐累积从而影响最终伪标签准确率.

4.3　属性约简对比实验

4.3.1　基于混合熵的属性约简

为验证基于混合熵的属性约简算法的有效性，在约简后属性个数、算法运行时间、基于约简数据的分类准确率三方面，比较其与基于正域、基于条件熵、基于辨识矩阵的属性约简算法的结果.

实验中，设置参数λ=2，各属性的邻域半径定义为δ=std(data)/λ，为加快算法的收敛，设置条件熵和混合熵控制参数为 0.05，取数据全集进行属性约简，得到的约简大小如表 3 所示，运行时间如表 4 所示.

表 3 不同约简算法约简大小

数据集	正域	条件熵	辨识矩阵	混合熵
zoo	**1**	16	13	13
lymphography	**1**	35	20	28
dermatology	**1**	51	12	34
credit	46	41	**24**	37
yeast	**8**	**8**	**8**	**8**
steel	27	16	**14**	**14**
image	18	18	14	**11**
seismic	18	18	17	**15**
平均	**15**	26	16	20

表 4 不同约简算法运行时间（单位：s）

数据集	正域	条件熵	辨识矩阵	混合熵
zoo	**0.83**	5.38	1.41	7.42
lymphography	**5.14**	32.97	16.07	38.29
dermatology	**57.53**	649.42	1280.07	618.94
credit	**54.89**	95.46	3027.85	101.51
yeast	**5.22**	7.16	60.82	7.29
steel	67.95	51.29	4812.07	**20.09**
image	37.92	61.48	3021.02	**18.92**
seismic	**301.31**	334.20	327.02	528.42

由表 3 可以看出，基于混合熵的约简算法得到的约简大小，在 zoo、lymphography 和 dermatology 数据集上，介于基于正域和基于条件熵的约简算法之间；在 credit、steel、image 和 seismic 数据集上，少于基于正域和基于条件熵的约简算法；在全部 8 个数据集上，均少于基于条件熵的约简算法. 就平均约简大小而言，基于混合熵的算法介于基于正域和基于条件熵的算法之间，且多于基于辨识矩阵的算法. 由表 4 可以看出，基于混合熵的约简算法的运行时间，在 zoo、lymphography、credit、yeast、seismic 数据集上，多于基于条件熵的约简算法；在 steel、image 数据集上，少于基于正域和基于条件熵的约简算法；在 dermatology 数据集上，介于基于正域和基于条件熵的约简算法之间. 基于正域的约简算法的运行时间，在 6 个数据集上最短；基于辨识矩阵的属性约简算法的运行时间，在 5 个数据集上运行时间最长，且在稍大数据集上运行时间增长显著.

基于上述各算法得到的约简数据，使用决策树算法训练分类模型，采用十折交叉算法划分训练集与测试集并进行十次随机实验，得到的平均分类准确率如图 4 所示.

其中，由基于混合熵的约简数据训练的模型，其分类准确率在 dermatology、image 数据集上，略低于基于条件熵的分类准确率；在其它 6 个数据集上，高于基于条件熵的分类准确率；在全部 8 个数据集上，均高于基于正域的分类准确率；在除 lymphography 和 dermatology 以外的 6 个数据集上，高于基于辨识矩阵的分类准确率.

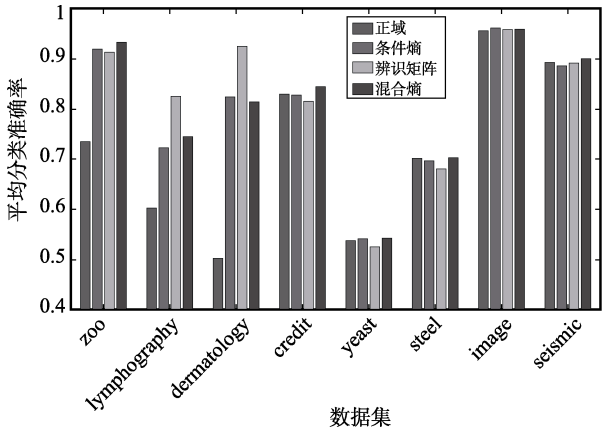

图 4 基于约简数据的平均分类准确率对比

综上所述，针对实验中的 8 个混合型数据，基于混合熵的属性约简算法，在约简大小上，介于基于正域和基于条件熵的属性约简算法之间；在算法运行时间上，与基于条件熵的属性约简算法大致相当；在约简后模型分类准确率上，高于基于正域的属性约简算法，与基于条件熵的属性约简算法大致相当. 以上结果说明了基于混合熵的属性约简算法的有效性及优势.

4.3.2 半监督属性约简

为验证本文所提半监督属性约简方法（3WLPME）的有效性，比较其与方法 COIN[16]、RSFS[17] 和 LPEN[18]，在不同标记率数据上，属性约简大小、运行时间、以及基于约简数据的分类准确率.

实验中，采用十折交叉算法将数据集划分为训练集与测试集，再将训练集按标记率随机划分为有标签集和无标签集，使用不同半监督属性约简方法对训练集进行属性约简，并训练分类器. 实验结果为十次实验的平均值.

各数据集约简后属性个数取不同标记率下约简的平均值，不同半监督约简方法的约简大小如表 5 所示，平均运行时间如表 6 所示. 其中，基于辨识对的 RSFS 方法在数据量增加时运行时间增长显著，未能获取其在 steel、image 和 seismic 数据集上的实验数据.

由表 5 可以看出，3WLPME 的约简大小，在 steel、image、seismic 数据集上最小；在 zoo、lymphography、dermatology 数据集上，比 COIN 大，

表 5　不同半监督属性约简方法约简大小

数据集	COIN	RSFS	LPEN	3WLPME
zoo	**15**	28	30	16
lymphography	**23**	59	59	27
dermatology	**30**	130	130	34
credit	23	**15**	38	36
yeast	**8**	**8**	**8**	**8**
steel	16	–	25	**14**
image	15	–	18	**13**
seismic	16	–	21	**15**
平均	**19**	–	42	21

比 RSFS、LPEN 小；在 8 个数据集上的平均约简大小比 COIN 大，比 LPEN 小.

表 6 的数据表明，3WLPME 的运行时间，在除 zoo、lymphography、seismic 外的 5 个数据集上最短，在所有 8 个数据集上运行时间比算法 RSFS、LPEN 短. 说明该方法运行效率的优势.

基于约简后的数据，用决策树分类器进行模型训练，分类准确率与标记率之间的关系如图 5 所示. 其中 RAWDATA 表示使用训练集全集训练模型，LABELEDDATA 表示仅使用有标签集约简后，再训练模型.

表 6　不同半监督约简方法的平均运行时间（单位：s）

数据集	COIN	RSFS	LPEN	3WLPME
zoo	**1.6545**	155.4713	36.5641	7.4161
lymphography	**32.0853**	1318.4948	69.1131	38.6474
dermatology	2036.9210	23498.8572	2386.2684	**622.2288**
credit	6692.7153	1810.8065	469.6257	**161.5791**
yeast	334.7282	3101.5585	523.2795	**28.2203**
steel	4573.5784	–	3991.6776	**79.6061**
image	5042.8877	–	4319.4109	**226.2888**
seismic	**860.1784**	–	3529.9003	1044.9542

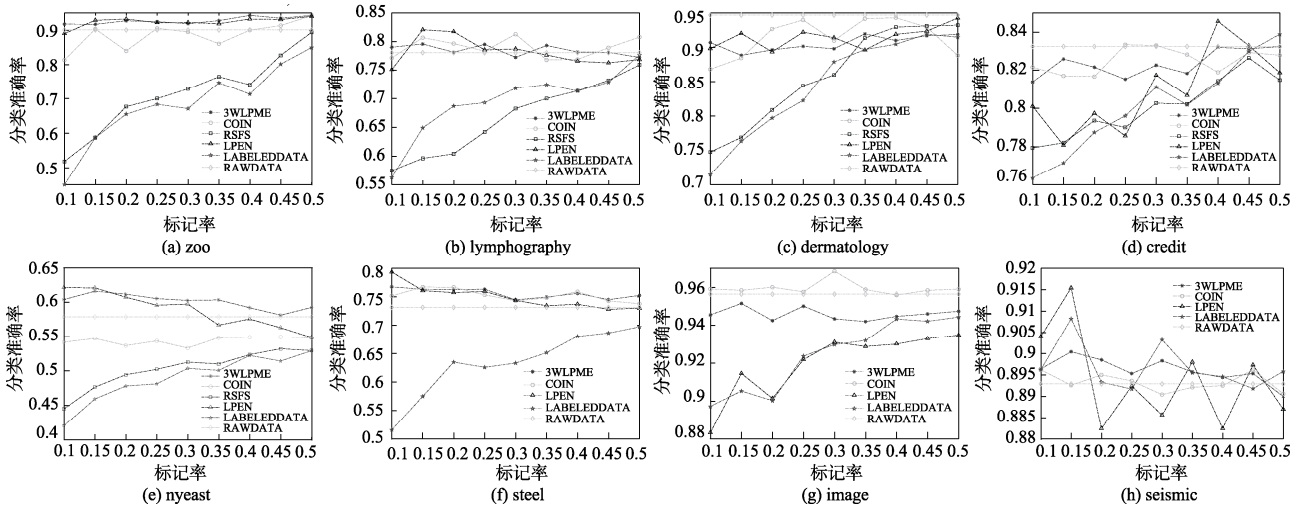

图 5　不同标记率下基于半监督属性约简的分类准确率

从图 5 可以看出，使用 3WLPME 方法后训练模型，其分类准确率在除 seismic 外的 7 个数据集上，比仅使用有标签集约简（LABELEDDATA）后的分类准确率高，说明通过 3WLP 中主动学习挑选的样本能提升约简的有效性. 在数据标记率较低时，基于半监督属性约简后的分类准确率，通常比仅使用有标签集约简后的分类准确率高. 在数据集 zoo、lymphography、dermatology、steel 上，基于 3WLPME 的分类准确率与基于 COIN、LPEN 的相当；在数据集 credit、yeast、image 上，基于 3WLPME 的分类准确率高于基于 LPEN 的分类准确率. 此外，每次实验中，有标签集和无标签集是按标记率随机选取的，分类准确率存在波动现象，但方法 3WLPME 对应的波动较小，说明了该方法的稳定性.

综上，使用本文提出的基于三支标签传播的半监督属性约简方法，在保证数据分类准确率的同时，相比其它算法得到的约简属性个数较少，运行时间较短，特别是相比于基于辨识矩阵和辨识对的半监督属性约简方法，说明本文方法是有效的.

5　结　论

针对部分标记的混合型分类数据，本文提出了基于三支标签传播的半监督属性约简方法，该方法主要分为两个阶段：首先使用三支标签传播算法获得无标签数据集的伪标签，接着基于数据全集，使用基于邻域混合熵的属性约简算法得到特征子集.

其中，三支标签传播算法结合了三支决策和主动学习思想；基于混合熵的属性重要度，可以同时度量属性分类能力的保持，和对系统平均不确定性的影响. 理论分析和实验结果说明了本文所提方法的有效性及优势. 下一步，可对本文方法进行改进，以验证其在更大数据集或不平衡数据集上的有效性.

参 考 文 献

[1] Girish Chandrashekar, Ferat Sahin, A survey on feature selection methods. Computers & Electrical Engineering, 2014, 40(1): 16-28

[2] Zhang R, Nie F, Li X, et al. Feature selection with multi-view data: A survey[J]. Information Fusion, 2019, 50: 158-167

[3] Hu X, Zhou P, Li P, Wang J, Wu X. A survey on online feature selection with streaming features. Frontiers of Computer Science, 2018, 12(3): 479-493

[4] Lewis D D. Feature selection and feature extraction for text categorization//Proceedings of the workshop on Speech and Natural Language. New York, USA: Association for Computational Linguistics, 1992: 212-217

[5] Kira K, Rendell L A. The feature selection problem: traditional methods and a new algorithm//Proceedings of the 10th National Conference on Artificial Intelligence. Menlo Park, USA: AAAI Press, 1992: 129-134

[6] He Xiaofei, Deng Cai, and Partha Niyogi. Laplacian score for feature selection//Proceedings of the 18th International Conference on Neural Information Processing Systems (NIPS'05). Cambridge, USA: MIT Press, 2005: 507-514.

[7] Kohavi R, John G H. Wrappers for feature subset selection. Artificial Intelligence, 1997, 97(1-2): 273-324

[8] Weston J, Elisseeff, André, Schölkopf, Bernhard, et al. Use of the zero-norm with linear models and kernel methods. Journal of Machine Learning Research, 2003, 3:1439-1461

[9] Pawlak Z. Rough sets: theoretical aspects of reasoning about data. Boston, USA: Kluwer Academic Publishers, 1991

[10] Miao D Q, Hu G R. A heuristic algorithm for reduction of knowledge. Journal of Computer Research & Development, 1999, 36(6): 681-684 (in Chinese)
(苗夺谦, 胡桂荣. 知识约简的一种启发式算法. 计算机研究与发展, 1999, 36(6): 681-684)

[11] Wang G Y, Yu H, Yang D. Decision table reduction based on conditional information entropy. Chinese Journal of Computers, 2002, 25(7): 759-766 (in Chinese)
（王国胤, 于洪, 杨大春. 基于条件信息熵的决策表约简. 计算机学报, 2002, 25(7): 759-766）

[12] Luo S, Miao D Q, Zhang Z F, Zhang Y J, Hu S D. A neighborhood rough set model with nominal metric embedding. Information Sciences, 2020, 520: 373-388

[13] Zhou Z H. A brief introduction to weakly supervised learning. National Science Review, 2018, 5(01): 48-57

[14] Liu J W, Liu Y, Luo X L. Semi-supervised learning methods. Chinese Journal of Computers, 2015, 38(8): 1592-1617 (in Chinese)
(刘建伟, 刘媛, 罗雄麟. 半监督学习方法. 计算机学报, 2015, 38(8): 1592-1617)

[15] Sheikhpour, R, Sarram, M A, Gharaghani, S, Chahooki, M A Z. A Survey on semi-supervised feature selection methods. Pattern Recognition, 2017, 64: 141-158

[16] Zhang W, Miao D Q, Gao C, Li F. Rough set attribute reduction algorithm for partially labeled data. Computer Science, 2017, 44(1): 25-31 (in Chinese)
(张维, 苗夺谦, 高灿, 李峰. 一种处理部分标记数据的粗糙集属性约简算法. 计算机科学, 2017, 44(1): 25-31)

[17] Dai J, Hu Q, Zhang J, et al. Attribute selection for partially labeled categorical data by rough set approach. IEEE Transactions on Cybernetics, 2017, 47(9): 2460-2471

[18] Liu K Y, Yang X B, Yu H L, Mi J S, Wang P X, Chen X J. Rough set based semi-supervised feature selection via ensemble selector. Knowledge-Based Systems, 2019, 165: 282-296

[19] Yao Y Y. Three-way decisions with probabilistic rough sets. Information Sciences, 2010, 180(3): 341-353

[20] Yao Y Y. Tri-level thinking: models of three-way decision. International Journal of Machine Learning and Cybernetics, published online, 2019, doi:10. 1007/s13042-019-01040-2

[21] Yao Y Y. Three-way granular computing, rough sets, and formal concept analysis. International Journal of Approximate Reasoning, 2020, 116: 106-125

[22] Zhu X, Ghahramani Z. Learning from labeled and unlabeled data with label propagation. Pennsylvania, American: Carnegie Mellon University, Technical Report: CMU-CALD-02-107, 2002

[23] Lin T Y. Neighborhood systems and approximation in relational databases and knowledge bases//Proceedings of the 4th International Symposium on Methodologies of Intelligent Systems. Charlotte, USA, 1988: 75-86

[24] Hu Q, Yu D, Xie Z. Neighborhood classifiers. Expert Systems with Applications, 2008, 34(2): 866-876

[25] Gower J C. A general coefficient of similarity and some of its properties. Biometrics, 1971, 27(4): 857-871

[26] Zhu X, Semi-supervised learning literature survey. Madison, WI, American: University of Wisconsin - Madison, Technical Report: 1530, 2008

[27] Szummer M, Jaakkola T. Partially labeled classification with markov random walks//Proceedings of the Advances in Neural Information Processing Systems. British Columbia, Canada: MIT Press, 2001: 945-952

[28] Wang F, Zhang C. Label propagation through linear neighborhoods. IEEE Transactions on Knowledge & Data Engineering, 2008, 20(1): 55-67

[29] Skowron A, Rauszer C. The discernibility matrices and functions in information systems//Proceedings of the Intelligent Decision Support-handbook of Applications and Advances of the Rough Sets theory. Dordrecht, Netherlands, 1991: 331-362

[30] Miao D Q, Wang J. On the relationships between information entropy and roughness of knowledge in rough set theory. Pattern Recognition and Artificial Intelligence, 1998, 11(1): 34-40 (in Chinese)
(苗夺谦, 王珏. 粗糙集理论中知识粗糙性与信息熵关系的讨论. 模式识别与人工智能, 1998, 11(1): 34-40)

HU Sheng-Dan, Ph. D. candidate. Her main research interests include rough sets, granular computing, and machine learning.

MIAO Duo-Qian, Ph. D. , professor, Ph. D. supervisor. His main research interests include rough sets, granular computing, pattern recognition and artificial intelligence, etc.

YAO Yi-Yu, Ph. D. , professor. His main research interests include three-way decision, rough sets, interval sets, granular computing, information retrieval, Web intelligence, data mining etc.

Background

Semi-supervised feature selection is a hot topic in machine learning community due to the benefits it provides for reducing the computational cost and improving the learning performance in diverse fields, such as machine learning, pattern recognition, data mining and big data. Many methods have hitherto been proposed, and they can be categorized into three groups: filter-based methods, wrapper-based methods and embedded-based methods. Attribute reduction is an important application of rough set theory and can be reviewed as a filter-based method. It has attracted many researchers since the day rough set theory was put forward. But there are only a few studies about semi-supervised attribute reduction in rough set community, at the same time there exist some challenges: such as the data type should be extended, the reduction quality and the reduction efficiency need to be improved.

For the partially labeled data with mixed categorical and continuous attributes, we propose a semi-supervised attribute reduction method. There are two algorithms in this method, namely the three-way label propagation algorithm (3WLP) and the heuristic attribute reduction algorithm based on mix entropy. The uniqueness of the three-way label propagation algorithm is obtaining the pseudo labels of unlabeled data by integrating the three-way decision theory and active leaning into *k*NN label propagation algorithm. Comparative experiments show that the accuracy of pseudo labels in this algorithm is higher than the one in random label propagation algorithm. In the heuristic attribute reduction algorithm, the heuristic information is also attribute significant, but it is designed by mix entropy, which can reflect the classification ability of an attribute and reveal the average uncertainty of the decision system at the same time. Comparative experiments indicate that, in this algorithm, the reduction quality is between the ones of positive-based algorithm and conditional entropy-based algorithm, and the classification accuracy of reduced data is higher than the ones of positive-based algorithm and conditional entropy-based algorithm. At last, various semi-supervised attribute reduction methods are also compared. The experimental results reflect the effectiveness of the proposed method.

The research is supported by the National Natural Science Foundation of China under Grant Nos. 61976158, 61673301.

This paper presents a new method to handle the semi-supervised attribute reduction, which uses three-way decision theory into label propagation algorithm to obtain more accurate pseudo labels and uses mix entropy to measure the significance of attribute. It extends the rough set theory and the three-way decision theory.