主动协同半监督粗糙集分类模型*

高 灿 苗夺谦 张志飞 刘财辉

(同济大学 电子与信息工程学院 计算机科学与技术系 上海 201804)
(同济大学 嵌入式系统与服务计算教育部重点实验室 上海 201804)

摘 要 粗糙集理论是一种有监督学习模型,一般需要适量有标记的数据来训练分类器. 但现实一些问题往往存在大量无标记的数据,而有标记数据由于标记代价过大较为稀少. 文中结合主动学习和协同训练理论,提出一种可有效利用无标记数据提升分类性能的半监督粗糙集模型. 该模型利用半监督属性约简算法提取两个差异性较大的约简构造基分类器,然后基于主动学习思想在无标记数据中选择两分类器分歧较大的样本进行人工标注,并将更新后的分类器交互协同学习. UCI 数据集实验对比分析表明,该模型能明显提高分类学习性能,甚至能达到数据集的最优值.

关键词 粗糙集,差别矩阵,半监督约简,主动学习,协同训练
中图法分类号 TP 181

A Semi-Supervised Rough Set Model for Classification Based on Active Learning and Co-Training

GAO Can, MIAO Duo-Qian, ZHANG Zhi-Fei, LIU Cai-Hui
(*Department of Computer Science and Technology, College of Electronics and Information Engineering, Tongji University, Shanghai* 201804)
(*Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai* 201804)

ABSTRACT

Rough set theory, as an effective supervised learning model, usually relies on the availability of an amount of labeled data to train the classifier. Howerer, in many practical problems, large amount of unlabeled data are readily available, and labeled ones are fairly expensive to obtain because of high cost. In this paper, a semi-supervised rough set model is proposed to deal with the partially labeled data. The proposed model firstly employs two diverse semi-supervised reducts to train its base classifiers on labeled data. The unlabeled ramified samples for two base classifiers are selected to be labeled based on the principle of active learning, and then the updated classifiers learn from each other by labeling confident unlabeled samples to its concomitant. The experimental results on selected UCI datasets show that the proposed model greatly improves the classification performance of partially labeled data, and even the best

* 国家自然科学基金项目(No. 60970061,61075056,61103067)、中国博士后科学基金项目(No. 2011M500626, 2011M500815)、上海市重点学科建设项目(No. B004)资助
收稿日期: 2011-06-20; 修回日期: 2012-04-20
作者简介 高灿,男,1983 年生,博士研究生,主要研究方向为粒计算、机器学习等. E-mail: 2005gaocan@163.com. 苗夺谦,男,1964 年生,教授,博士生导师,主要研究方向为粒计算、Web 智能、机器学习等. 张志飞,男,1986 年生,博士研究生,主要研究方向为文本挖掘,机器学习等. 刘财辉,男,1979 年生,博士研究生,主要研究方向为粒计算、机器学习.

performance of dataset is obtained.

Key Words Rough Set, Discernibility Matrix, Semi-Supervised Reduction, Active Learning, Co-Training

1 引 言

粗糙集理论是一种处理不精确、不一致和不完备信息的数学工具[1],自波兰学者 Pawlak 教授提出以来,在机器学习、数据挖掘和人工智能等领域得到广泛应用[2-3].在传统粗糙集分类应用中,为了训练较好的分类器往往需要大量有标记数据.而在较多现实问题中(如垃圾邮件处理、网页分类和入侵检测等),由于获取数据标记的代价昂贵,以致有类别标记数据较为稀少,而无标记的数据获取则相对容易,往往有大量无类别信息的数据可供使用.如果仅在有标记数据上通过约简而产生相应的分类器,其分类预测效果不理想.因此,研究如何有效利用大量的无标记数据提升粗糙集的分类性能具有重要意义.

文献[3]将相容粗糙集模型引入文本分类中,利用相容类的概念重复从无标记数据中抽取可信度较高的潜在正例和负例文本对象,以增加有标记文本对象的方式提高文本分类的准确度.文献[4]将 Yao 的决策论粗糙集模型从两类问题扩展到多类情况,给出半监督粗糙风险/代价的计算方法,并成功应用于指导现实零售商店的优惠活动.针对基于反馈神经网络(Back Propagation Neural Network, BPNN)的瞬态稳定性评估方法中由于特征选择方法不足而造成误分率较高的问题,文献[10]运用粗糙集上下近似概念将系统的瞬态稳定性评估状态分为稳定类、不稳定类和不可确定的边界类,并在不可确定边界类上利用半监督反馈算法训练反向传播神经网络进一步进行类别划分,以减少稳定性评估中的误分率.文献[11]为了在无线多媒体传感器网络实现有效的可视目标分类,提出协同统计在线支持向量机学习算法.算法首先利用粗糙集理论对多源传感器数据约简降维,同时分类器在不同环境下利用新的无标记数据更新学习,达到较高的可视目标分类效果.以上类方法仅将粗糙集理论引入部分标记数据分类应用中,而从理论模型上利用无标记数据提升有监督粗糙集知识获取及分类性能的研究还鲜有报道.

粗糙集的研究对象一般是有标记对象集合(相对约简)或无标记对象集合(约简),对于部分标记数据暂无较好的方法获取其约简.本文通过数据变换,将部分标记数据转换为特殊决策表,给出其相应的差别矩阵及约简算法以解决部分标记数据属性约简问题.对于该约简算法的有效性,相关命题亦进行详细论证.其次,对于粗糙集不能有效处理部分标记数据分类的缺陷,提出主动协同半监督粗糙集模型.该模型利用所提出的部分标记数据属性约简算法生成两个具有差异性的约简,并在有标记数据上分别训练两初始基分类器.然后在无标记数据上,以人工标注方式消除两分类器的分歧.模型最后将提升后的分类器重复协同训练,以增加各分类器较大信度标记样本的方式进一步提升各分类器的分类性能.在理论上,本文分析主动协同半监督粗糙集模型的有效性,给出模型分类性能提升的主要原因.UCI 数据集实验仿真结果表明,该模型的分类性能优于已有同类学习算法,进一步说明模型的有效性.

2 基本概念

为便于描述,以下给出与本文相关的基本概念,相关理论的详细介绍请分别参阅文献[1] ~ [7]和文献[12] ~ [16].

2.1 粗糙集

定义 1 信息系统可表示为 $S=(U, A, V, f)$,其中,U 是对象集合;A 是属性非空集合;$V=\cup V_a$,V_a 表示属性 a 的值域;$f: U\times A\to V$ 是信息函数,指定 U 中每个对象 x 的属性值,即对 $x\in U$,$a\in A$,有 $f(x, a)\in V_a$.

如果属性集 A 可分为条件属性集 C 和决策属性集 D,即 $A=C\cup D$,$C\cap D=\varnothing$,则该信息系统称为决策信息系统或决策表.

定义 2 给定信息系统 $S=(U,A,V,f)$,对于任意属性子集 $B\subseteq A$,可以定义不可分辨关系:

$$IND(B)=\{(x, y)\in U\times U\mid \forall a\in B, f(x, a)=f(y, a)\}.$$

$IND(B)$ 是一个等价关系,构成 U 的一个划分,用 $U/IND(B)$ 表示,简记为 U/B.

定义 3 给定决策表 $S=(U, A=C\cup D, V, f)$,条件属性 C 和决策属性 D 导出的不可分辨关系:

$$U/C=\{C_1, C_2, \cdots, C_{|U/C|}\},$$

$$U/D=\{D_1, D_2, \cdots, D_{|U/D|}\},$$

其中,U/C 中的元素 C_i 称为条件类,U/D 中的元素 D_j 称决策类.

定义4　给定决策表 $S=(U, A=C\cup D, V, f)$，对任意属性子集 $B\subseteq A$，概念 $X\subseteq U$ 关于 B 的下近似 $\underline{B}(X)$ 及上近似 $\overline{B}(X)$：

$$\underline{B}(X)=\cup\{Y\in U/B \mid Y\subseteq X\},$$

$$\overline{B}(X)=\cup\{Y\in U/B \mid Y\cap X\neq\varnothing\}.$$

定义5　给定决策表 $S=(U, A=C\cup D, V, f)$，则决策表的正域 $POS_C(D)$ 及边界域 $BND_C(D)$：

$$POS_C(D)=\bigcup_{X\in U/D}\underline{C}(X),$$

$$BND_C(D)=\bigcup_{X\in U/D}\overline{C}(X)-\bigcup_{X\in U/D}\underline{C}(X).$$

定义6　给定决策表 $S=(U, A=C\cup D, V, f)$，则决策表 S 的差别矩阵 $\boldsymbol{M}$ 的元素项：

$$m_{ij}=\begin{cases}\{a_k \mid a_k\in C\wedge a_k(x_i)\neq a_k(x_j)\}, & d(x_i)\neq d(x_j)\\ \varnothing, & \text{otherwise}\end{cases}$$

不同类型决策表，相容或不相容，其差别矩阵构造方式不同，在此不作详细探讨. 若无特殊说明，本文研究对象均为相容决策表，也即决策表中不存在条件属性完全相同而决策不同的对象.

定义7　给定决策表 $S=(U, A=C\cup D, V, f)$，$\boldsymbol{M}$ 为决策表 S 的差别矩阵，$P\subseteq C$，若 P 满足

1) $\forall m\in\boldsymbol{M}$, $m\cap P\neq\varnothing$,

2) $\forall a\in P, P'=P-\{a\}$, $\exists m\in\boldsymbol{M}, m\cap P'=\varnothing$,

则称 P 是决策表的一个约简.

定义8　给定决策表 $S=(U, A=C\cup D, V, f)$，其所有约简表示为 $Red(C)$，而所有约简的交集称为核属性，记为 $Core(C)$，即 $Core(C)=\cap Red(C)$.

2.2　主动学习与协同训练

主动学习[12-13]是机器学习的重要研究领域之一，其主要思想是通过人工(专家)标注少量学习模型最不确定的样本达到提升学习性能的效果. 在部分标记数据中，主动学习先利用有标记数据训练其初始学习模型，其次在无标记数据上选择有助于提升当前学习模型性能的关键样本进行标注，从而在较小的标注代价下获得较好的学习性能. 根据无标记数据出现方式，主动学习大体可分为[12]，基于样本空间、流和池的主动学习. 第一类方法的无标记数据为特征空间中所有样本，主动学习可能选择任一无标记样本进行标注; 基于流的学习方式中，学习模型逐次对顺序生成的无标记样本进行标记或丢弃; 而基于池的主动学习通常有一个无标记样本集合，学习模型可依据策略选择某些样本进行标注. 根据无标记样本选择策略的不同，主动学习又可分为[12]，基于不确定度、模型空间缩减、期望模型变化，泛化误差等方法. 虽然对无标记样本度量方式不同，但都是选择最具信息量的无标记样本进行标注进而提升学习模型的性能.

Blum 和 Mitchell[14]提出的协同训练 Co-Training 是一种处理部分标记数据[15-18]的经典算法. 算法假设部分标记数据的条件属性能自然分割成两个充分且冗余的视图(属性子集)，在两个视图上利用有标记数据分别训练初始分类器，然后在无标记数据上相互标记一些置信度较高的样本作为另一分类器的训练集，重复迭代直到满足某个停止条件. 然而，在现实问题中，充分冗余视图这一要求往往很难得到满足. 文献[19] ~ [22]通过采用不同分类器或重采样技术来训练多个具有差异性的分类器协同训练，取得一定的性能提升效果. 但以上方法一定程度上放松协同训练的约束条件，从而失去一些较好的性质. 如何将不存在自然分割的属性集划分成两个充分且冗余视图还是一个没有完全解决的问题.

3　基于粗糙集的半监督属性约简

原有粗糙集方法一般处理单种类型的数据，要么是有标记的决策表，要么是无标记的信息系统. 而对于部分标记数据，目前没有较好的方法获取其约简. 为方便粗糙集处理，对于部分标记数据可考虑以下两种数据转换策略: 1) 将有标记数据去掉其标记，部分标记数据转变为信息系统; 2) 在条件属性集下将有标记和无标记对象等价类划分，对于包含有标记对象的等价类，将等价类中无标记对象的决策值标记为有标记对象的决策值，其它各等价类标记为不同的特殊决策值，部分标记数据转变为决策表. 策略 1 没有利用重要的有标记数据决策信息，转换后的信息系统约简能保持原部分标记数据的分辨能力，但其约简包含的冗余属性可能较多，约简质量较差. 策略 2 首先在所有条件属性上对有标记和无标记数据进行等价类划分，由于等价类中各对象是不可区分的，如果等价类中包含有标记对象，则可将其决策标记传播至等价类中其它无标记对象. 而对于全部由无标记对象构成的等价类，由于无决策标记信息可利用，所以标记为与其它已知标记类都不同的特殊决策值(文中以“* ”表示). 各等价类对象标记决策值后，部分标记数据转换为决策表. 该思想充分利用有标记数据的决策信息和无标记数据的区分信息，保证约简前后的分类能力不变，并能有效去

除冗余属性. 为了方便描述,在此引入一些新的符号及定义.

一般地,部分标记数据表示为

$$PS = (U = L \cup N, A = C \cup D, V', f),$$

其中,U 为全体对象集合,包括有标记数据集合 L 和无标记数据集合 N,且决策属性 D 的值域 V_D 可取空值. 而通过策略 2 转换后的数据可表示为

$$TS = (U', A = C \cup D, V'', f),$$

决策属性 D 的值域 V_D 不包含空值,但存在特殊决策值"* ". 将部分标记数据中所有无标记对象标注正确的决策值后形成的决策表称为潜在决策表,仍以

$$S = (U, A = C \cup D, V, f)$$

表示. 根据策略 2 的思想,可定义以下差别矩阵处理部分标记数据.

定义 9 给定部分标记数据

$$PS = (U = L \cup N, A = C \cup D, V', f),$$
$$TS = (U', A = C \cup D, V'', f)$$

为转换后的决策表,则部分标记数据的差别矩阵 M' 元素项:

$$m'_{ij} = \begin{cases} \{a_k \mid a_k \in C \wedge a_k(x_i) \neq a_k(x_j)\}, \\ \quad d(x_i) \neq d(x_j) \vee d(x_i) = * \vee d(x_j) = * \\ \varnothing, \quad \text{otherwise} \end{cases}$$

定义的差别矩阵元素项包括三种不同对象之间产生的分辨信息,即不同决策的有标记对象,有标记对象与无标记对象,以及两无标记对象. 对于无标记对象,由于其决策值可能与有标记对象和其它无标记对象不同,所以需保留其分辨信息,以保证约简后属性集的分辨能力. 为证明该差别矩阵生成的约简与潜在决策表约简的关系,现给出以下命题.

命题 1 假设部分标记数据

$$PS = (U = L \cup N, A = C \cup D, V', f)$$

的核属性集为 $Core_1$,将无标记数据 N 标记正确的决策值后形成的潜在决策表

$$S = (U, A = C \cup D, V, f)$$

的核属性集为 $Core_2$,则 $Core_2 \subseteq Core_1$.

证明 反证法. 假设 S 中存在核属性 $a \in Core_2$,但 $a \notin Core_1$. 因为 $a \in Core_2$,则必定存在两对象 X_i, X_j,其决策值不同且在条件属性上仅有 a 能区分两对象. 如果对象 X_i, X_j 同属于 PS 的有标记对象,则 PS 的差别矩阵必定会产生区分 X_i, X_j 的区分属性信息 a,所以 a 必定属于 PS 的核属性集. 如果 X_i, X_j 有其一或两者都属于无标记对象,两不同等价类的对象至少有其一会标记为特殊决策值. 根据差别矩阵的定义(定义 9),PS 的差别矩阵会产生区分对象 X_i, X_j 的元素项 a,即 $a \in Core_1$. 假设矛盾,证毕.

命题 2 假设部分标记数据

$$PS = (U = L \cup N, A = C \cup D, V', f)$$

的差别矩阵为 M',将无标记数据 N 标记正确的决策值后形成的潜在决策表

$$S = (U, A = C \cup D, V, f)$$

的差别矩阵为 M,则 $M \subseteq M'$.

证明 在有标记数据 L 上,差别矩阵 M' 和 M 的元素项完全相同. 而在无标记数据 N 上,两差别矩阵有所差异. 对于无标记与有标记对象,由于在部分标记数据变换中无标记对象将标记为特殊决策值"* ",所以任意无标记对象都将与有标记对象进行对比,生成分辨信息以区分两对象. 而在潜在决策表 S 中,无标记数据 N 已标记正确的决策值,其值可能与原有标记对象的决策值相同,所以潜在决策表 S 将不会产生相关的差别矩阵元素项. 对于两无标记对象,也存在类似情况. 所以潜在决策表 S 的任意差别矩阵元素项都存在于部分标记数据 PS 的差别矩阵中,也即 $M \subseteq M'$.

命题 3 假设部分标记数据

$$PS = (U = L \cup N, A = C \cup D, V', f)$$

的差别矩阵为 M',将无标记数据 N 标记正确的决策值后形成的潜在决策表

$$S = (U, A = C \cup D, V, f)$$

的差别矩阵为 M,则在 M' 上生成的约简至少包含一 M 上生成的约简,也即对任意的 PS 约简 RED_1 至少存在一 S 的约简 RED_2,使 $RED_2 \subseteq RED_1$.

证明 根据命题 2 的结论,部分标记数据的差别矩阵 M' 将包含潜在决策表的差别矩阵 M. 不失一般性,假设两差别矩阵的差集 DM 中存在一元素项 m_1,则差集元素项 m_1 与差别矩阵 M 存在以下关系.

1) $\exists m_2 \in M$ 且 $m_1 \subseteq m_2$. 根据约简的定义(定义 7),约简与任一非空元素项都存在交集. 由于差集元素项 m_1 包含于 M 中一元素项,所以差别矩阵 M 与 m_1 产生的约简必定与 M 的任意元素项都存在非空交集,也即 M 与 m_1 产生的任意约简都能分辨 S 的所有对象. 如果差集 DM 全由此类元素项构成,则部分标记数据 PS 的约简与潜在决策表 S 的约简相同.

2) $\exists m_2 \in M$ 且 $m_1 \supset m_2$. 根据吸收律,M 与 m_1 的约简分辨信息实际上等价于 M,所以对于任意 M 与 m_1 形成的约简,必定存在一 M 的约简且两约简具有相同的属性.

3) $\forall m_2 \in M$,$m_1 \not\subset m_2$ 且 $m_2 \not\subset m_1$. 由于差集元素项与任意 M 元素项都不存在包含关系,则 M 形成的约简可能不能完全区分部分标记数据 PS 上所有对

象,所以需加入相关属性至 M 的约简. 因此,M 和差集元素项 m_1 产生的约简必定包含一 M 产生的约简.

综合以上情况,部分标记数据 PS 产生的任意约简都包含潜在决策表 S 一约简,命题得证.

以上命题分析部分标记数据 PS 在核属性和约简上与潜在决策表 S 的关系. 根据命题 3 的结论,部分标记数据 PS 产生的约简将包含或等于数据的真实约简,说明差别矩阵及其约简的有效性. 因此,可运用原有差别矩阵启发式算法进行属性约简,在此不再详细描述.

4　主动协同半监督粗糙集模型

4.1　模型基本思想及框架

针对部分标记数据,原有粗糙集方法仅在少量有标记数据上训练单个分类器,大量无标记数据得不到有效利用,因此其分类效果可能不太理想. 属性约简是粗糙集理论重要研究内容之一,能有效将高维特征数据降至低维而不造成分类信息的损失,所以决策表的每个属性约简都能训练出一个较好的分类器. 一般来说,决策表的属性约简不是唯一的,即同一个决策表可能存在多个属性约简. 而每个属性约简都可训练较好的分类器,所以可运用半监督属性约简算法对属性空间进行分割,寻找两个最具差异性的约简进行有效的主动协同训练.

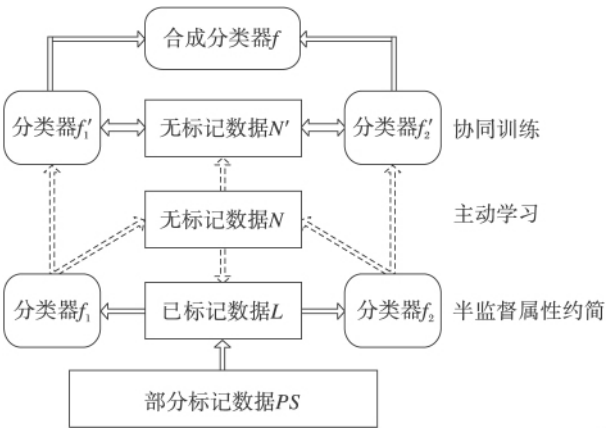


图 1　主动协同半监督粗糙集模型框架

Fig. 1　Framework of active learning and co-training-based semi-supervised rough set model

主动协同半监督粗糙集模型主要包括: 半监督属性约简、主动学习和协同训练三个步骤. 半监督属性约简过程利用定义的差别矩阵求取部分标记数据的差别信息,运用启发式算法获取最优半监督约简作为主动协同粗糙集学习模型的一个视图,而另一个视图则可选择与最优半监督约简存在最大差异性的其它部分标记数据约简. 其次,在两差异性约简视图上,主动学习过程利用有标记数据分别构造初始分类器,对两分类器分歧较大的无标记样本进行人工标注,将该无标记样本及其标记加入各分类器的训练集,并重新训练各分类器. 然后,协同训练过程将各分类器置信度较高的无标记样本标记决策值加入另一分类器的训练集,重复在无标记数据上进行交互协同训练,直至无有效的无标记数据可利用. 最后,对学习后的分类器进行合并,形成最终的分类器. 模型的框架图如图 1.

4.2　主动协同半监督粗糙集模型分类算法

要实现主动协同半监督粗糙集模型,首先需对属性空间进行分割,以获取两差异性的约简进行主动协同训练. 最优约简具有属性少、规则抽象程度较高的优点,可选为模型的一约简. 对于另一约简,理论最优方法是计算部分标记数据的所有约简,再选择与最优约简差异度最大的一约简. 但部分标记数据属性过多时,计算所有约简代价非常大,所以应考虑运用启发式思想求得另一约简. 约简的差异性有多种度量方法,如分类性能,规则空间,属性空间等. 由于本文处理对象是部分标记数据,所以采用属性空间差异性度量方法,也即构成约简的属性越不同,约简之间的差异性越大. 根据属性空间差异性度量标准,主动协同半监督粗糙集模型两约简应具有较少的共同属性,因此启发式算法应尽量避免选择出现在最优约简中的属性. 换句话说,算法在求取另一约简过程中应优先考虑未出现在最优约简中的属性,则生成的约简与最优约简具有较少的共同属性,而两约简的差异性则较大. 依据以上思想,则可构建如下启发式算法求取部分标记数据的两差异性约简.

算法 1　部分标记数据差异性约简获取算法

输入　部分标记数据

$$PS = (U = L \cup N, A = C \cup D, V', f)$$

输出　差异约简 Red_1 和 Red_2

step 1　将部分标记数据

$$PS = (U = L \cup N, A = C \cup D, V', f)$$

行转换;

step 2　根据定义 9 计算差别矩阵 M';

step 3　计算核属性 $Core$,$Red_1 = Red_2 = Core$;

step 4　如果 M' 为空,转 step 7;

step 5　设置约简 Red_1 的优先候选集 At 为 C-$Core$,在 M' 上调用约简算法得约简 Red_1;

step 6 设置约简 Red_2 的优先候选集 At 为 $C-Red_1$,在 M' 上调用约简算法得约简 Red_2;

step 7 输出 Red_1 和 Red_2,算法结束.

算法 1 首先根据定义 9 计算变换后数据的差别矩阵. 其差别矩阵元素项信息可分核属性和候选属性信息两部分. 核属性是所有约简的交集,所以当部分标记数据存在核属性时,算法 1 生成的两约简的交集将包含核属性. 第一个约简(最优约简)的生成是在核属性基础上,逐次选择除核属性以外的最大频率候选属性而获得. 要保证第二个属性约简与第一个的差异性,则需加入控制策略使两约简的属性交集较小. 算法 1 的 step 6 将约简的优先候选属性集限制在第一约简属性以外的条件属性,只有当优先候选属性不能形成约简时,再考虑加入第一约简中的属性,所以算法生成的两次约简将具有较少的相同属性. 由于算法仅需二次调用已有启发式约简算法,时间和空间复杂度的数量级与一般的启发式算法相当,在此不再详细描述.

依据算法 1 的思想,可将部分标记数据条件属性集分割成两个具有较少共同属性的差异约简,而每个约简都是保持部分标记数据分类能力的较优属性子集,因此其约简可构造好的分类器. 而且各约简从不同层次和角度描述数据的结构信息,所以两个具有差异性的约简能形成较好的分类器进行主动协同训练. 实现主动协同半监督粗糙集模型的算法描述如下.

算法 2 主动协同半监督粗糙集分类算法

输入 部分标记数据

$$PS=(U=L\cup N, A=C\cup D, V', f)$$

输出 分类器 f

step 1 运用算法 1 将部分标记数据条件属性分割成充分且具差异性的子集 Red_1 和 Red_2;

step 2 置两分类器的训练集 $L_1=L_2=L$,基于属性子集 Red_1 和 Red_2 构造分类器 f_1 和 f_2;

step 3 选择部分两分类器分歧较大的无标记样本进行标注,并加入各分类器训练集,更新无标记样本集 N 和分类器 f_1,f_2;

step 4 将 f_1 和 f_2 的非较大信度无标记样本分别加入集合 N_1 和 N_2;

step 5 如果 $N_1\cup N_2$ 等于 $N_1\cap N_2$,转 step 6,否则重复以下步骤:

step 5.1 将分类器 f_1 的较大信度预测样本 $N_2-(N_1\cap N_2)$ 标记决策值加入 f_2 的训练集 L_2,更新分类器 f_2 及非较大信度无标记样本集 $N_2=U-L_2$ 或 $N_2=N_2-(N_2-(N_1\cap N_2))$;

step 5.2 将分类器 f_2 的较大信度预测样本 $N_1-(N_1\cap N_2)$ 标记决策值加入 f_1 的训练集 L_1,更新分类器 f_1 及非较大信度无标记样本集 $N_1=U-L_1$ 或 $N_1=N_1-(N_1-(N_1\cap N_2))$;

step 6 输出合成分类器 f.

算法 2 首先通过算法 1 进行属性分割,在各属性子集上以有标记数据为训练集构造两个基分类器. 两分类器在无标记数据上可能出现以下 3 种情况:两分类器都以较大信度预测、两分类器有其一能以较大信度预测和两分类器都不能以较大信度预测. 对于第一类无标记样本,两分类器可能出现预测值不同的情形,此时两分类器必有其一预测错误. step 3 的主动学习主要选择以上类型无标记样本进行标注以降低各分类器的错误率. 而对于两分类器只有其一能以较大信度预测的无标记样本,较大信度分类器可将此样本标记其预测值加入另一分类器的训练集,以提高各分类器较大信度预测样本的数量. step 5 通过将分类器较大信度预测样本的决策值传播至另一分类器,以协同学习的方式达到提高各分类器性能的目的. 在理想情况下,两分类器最终都能以较大信度预测无标记数据样本,分类器的性能得到较大提升.

设 $|L|=l$,$|N|=n$,$|C|=m$,单分类器训练时间为 k. 算法 2 主动学习过程的复杂度主要取决于选择标注样本的数量. 假设主动学习标记样本数量为 t,$t<<n$,则 step 3 的复杂度为 $O(tk)$. 算法 2 循环协同训练过程中,两分类器通过相互学习消除其分类器的非较大信度预测样本,新的学习对象只能从两分类器相同的非较大信度预测样本中产生,所以算法循环次数一般较少. 最坏情况下,模型每次只能在一无标记样本上进行协同学习,而重新训练分类器后又有新的可利用无标记对象,此时最大循环次数为 $n-t$. 所以算法 2 在两差异性属性约简上进行主动协同学习的最坏时间复杂度为 $O(nk)$.

4.3 有效性分析

在协同训练的有效性方面,相关文献进行详细的研究分析. Blum 和 Mitchell [14] 证明,当充分冗余视图假设条件成立时,协同训练算法可通过利用无标记数据将一个从有标记数据学得的弱分类器提升到任意精度. Balcan 等 [23] 研究发现,只要数据分布满足"扩张性"假设,协同训练算法就可奏效. Wang 和 Zhou [24] 进一步证明只要两个分类器有较大的差异,就可通过利用无标记数据进行协同训练来提高分类性能. 主动协同半监督粗糙集模型的视图选用两个具有较少共同属性的部分标记数据约简,而各

约简实质上从不同视角描述数据结构特性，必定使其构造的分类器之间具有较大的差异性. 同时，约简的性质使模型满足协同训练的充分性假设条件，保证各分类器的有效性. 而主动学习过程有效提升各初始分类器的性能，提高后续协同训练的质量. 因此，模型应能有效处理部分标记数据.

假设部分标记数据包括有标记数据 L 和无标记数据 N，其中 $|L| = l$，$|N| = n$. 各样本在两约简空间上可表示为 $X = X_1 \times X_2$，其中，X_1 是属性约简 1 下的数据，X_2 是属性约简 2 下的数据. 在有标记数据 L 上，可分别训练两个分类器 f_1 和 f_2. 两分类器对无标记数据 N 的预测结果可分 3 种情况: 较大信度预测正确(Correct) ，较大信度预测错误(Incorrect) 和非较大信度预测(Unconfident) ，则两分类器在 n 个无标记数据上的差异性可表示为如下矩阵:

$$\begin{array}{l} \\ f_1\ correct(c) \\ f_1\ incorrect(i) \\ f_1\ unconfident(u) \end{array} \begin{array}{c} \begin{array}{ccc} f_2\ correct(c) & f_2\ incorrect(i) & f_2\ unconfident(u) \end{array} \\ \begin{bmatrix} n_{cc} & n_{ci} & n_{cu} \\ n_{ic} & n_{ii} & n_{iu} \\ n_{uc} & n_{ui} & n_{uu} \end{bmatrix} \end{array},$$

其中，n_{cc}，n_{ii} 和 n_{uu} 分别表示两分类器都以较大信度预测正确、错误以及非较大信度预测的样本数目; n_{ci} 和 n_{ic} 表示一分类器以较大信度预测正确而另一分类器以较大信度预测错误的对象数目; n_{cu} 和 n_{uc} 表示一分类器较大信度预测正确另一分类器不能以较大信度预测的样本数目; 一分类器较大信度预测错误而另一分类器不能以较大信度预测的样本分别以 n_{iu} 和 n_{ui} 表示. 在主动学习之前，两分类器的较大信度预测正确率:

$$\frac{(n_{cc} + n_{ci} + n_{cu})}{n}, \frac{(n_{cc} + n_{ic} + n_{uc})}{n},$$

而两分类器的差异性主要体现在除 n_{cc}、n_{ii} 和 n_{uu} 以外的样本. 通过主动学习标注分歧样本，各分类器的较大信度预测正确率最大提升度为 n_{ci}/n 和 n_{ic}/n. 在第一次协同训练过程中，分类器 f_2 将 n_{uc} 个无标记数据标记正确的预测值加入 f_1 的训练集，f_1 的较大信度预测正确的样本增加 n_{uc}. 与 f_1 类似，f_2 的较大信度预测正确的样本将增加 n_{cu}. 更新各分类器后，原有 n_{uu} 个非较大信度预测的样本对象可能出现两分类器有其一能以较大信度预测的情况，此时可进行第二次协同训练. 假设给定充足的无标记数据，在理想情况下，分类器 f_1 和 f_2 通过多次协同训练后都能以较大信度预测其属性空间任意样本. 通过将分类器的训练集从有标记数据扩充至较大信度预测正确的无标记数据，各分类器的性能得到较大提升.

5　实验仿真

5.1　数据集

实验选用 4 个 UCI 标准数据集作为实验对象，数据集详细信息见表 1. 表 1 中，Wine 和 Ionosphere 的连续型数据运用等频方法(三等分) 进行离散化预处理，而数据集 Lymphography 仅选取类别数据较平衡的两类进行实验.

表 1　UCI 数据集

Table 1　UCI datasets

数据集	属性数	样本数	类别数
Tic-Tac-Toe (TTT)	9	958	2
Wine recognition (Wine)	13	178	3
Lymphography-2(Lymp2)	18	142	2
Ionosphere (Iono)	34	351	2

实验采用 10 重交叉验证方法划分训练集和测试集，每重验证先按标记率将训练集随机划分为有标记和无标记集. 为验证算法在不同初始标记样本和主动标记样本比率下的效果，进一步对有标记样本集按初标率抽取部分有标记样本训练初始分类器，其它训练样本构成主动学习和协同训练的无标记样本集. 假设一数据集有 1 000 个样本，标记率为 10%，初标率为 10%，则 10% 的数据将会选为测试集，训练集先按标记率划分为 90 有标记样本集和 810 无标记样本集，进而按照初标率选 9 个样本作为初始标记集训练分类器，再运用主动学习思想从剩余 891 个样本中主动标记 81 个构成整体的有标记集进行协同训练. 由于 10 重交叉验证方法受数据集样本次序的影响，实验打乱样本进行 10 次随机 10 重交叉验证，以保证结果的有效性.

5.2　实验对比分析

为验证算法的有效性，实验计算原有粗糙集分类算法(算法 A1) ，即仅利用有标记数据训练分类

器的性能. 其次选用自训练[19](算法 A2)、随机协同[19](算法 A3)两种传统的半监督学习算法和协同测试主动学习算法[25](算法 A4)进行对比分析. 协同测试主动学习算法也需要两个视图训练分类器,为进行算法有效对比,协同测试主动学习算法亦采用本文算法 1 进行视图分割. 实验过程中,各算法将采用 ID3 决策树作为分类器. 主动协同半监督粗糙集模型(本文算法)将选择两分类器预测结果不同的无标记数据作为主动学习的样本,一分类器可预测另一分类器不可预测的样本则进行交互协同训练,本文算法的性能取主动协同训练后两分类器的平均值. 部分实验结果如表 2 所示.

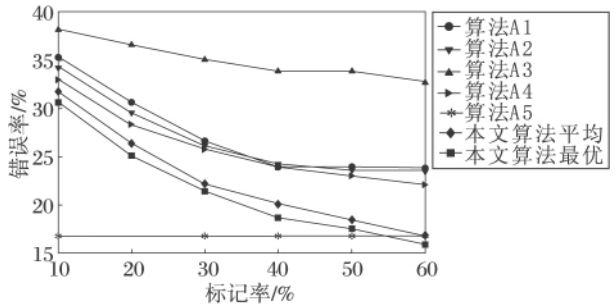
从表 2 可看出,各算法在所选数据集上性能表现不尽相同. 由于自训练方法(算法 A2)采用单分类器模式,无标记数据只能通过自我标注方式进行利用,其初始分类器的错误可能会通过自学习传播加强,所以自训练方法能利用无标记数据提升其学习性能,但同时也可能出现性能降低的情况,如数据集 TTT (40%),Wine(40%),Lymp2(40%)和 Iono (20%,40%). 随机协同训练方法(算法 A3)是双分类器模型,分类器可通过相互标注无标记样本提升其性能. 但两分类器的属性空间都是随机生成,很可能训练出较差的分类器,这就违反协同训练属性子集的充分性假设条件. 如果分类器协同训练的提升性能不能补尝其初始分类器的错误率,则随机协同训练的结果较差,甚至劣于原有粗糙集分类方法(算法 A1),如数据集 TTT 和 Wine 在各标记率下的实验结果. 协同测试算法主要通过人工标注两分类器产生分歧的无标记样本,各分类器的样本质量较高,因此一般比原粗糙集方法更优. 但相对于自训练和随机协同学习,协同测试仅能主动标记少量无标记样本,无标记数据的利用相当有限,其性能可能弱于半监督学习方法,如数据集 Wine(20%)和 Lymp2 (20%,40%,60%). 本文算法运用粗糙集理论生成两差异性属性约简,保证两分类器充分且具有差异性. 而主动学习过程进一步提升分类器协同训练前的分类性能,因此充分且差异的两分类器能通过协同训练获得较好的学习性能. 根据不同的初始标记率,本文算法的分类性能有所差异,表中加粗标记的结果为最优性能. 从表 2 可看出,初始标记样本过少或过多时,本文算法的性能一般相对较低. 当初始标记样本较少时,其初始分类器错率较高,两分类器在无标记数据上分歧样本过多或过少,因此主动学习的质量较低. 而当初始标记样本较多时,主动学习的次数将减少,两分类器的分歧样本得不到足够的标注,限制协同学习前各分类器的性能. 在标记率为 60% 时,由于各初标率下初始标记样本都较多,主动学习过程对本文算法的性能影响减少,数据集的平

表 2 标记率为 20%,40% 和 60% 下算法错误率对比

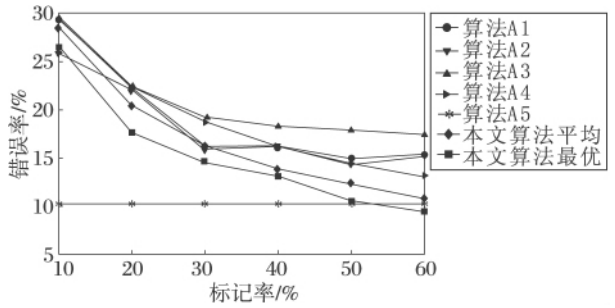
Table 2 Comparison of error rates among selected algorithms under label rates 20%, 40% and 60%

%

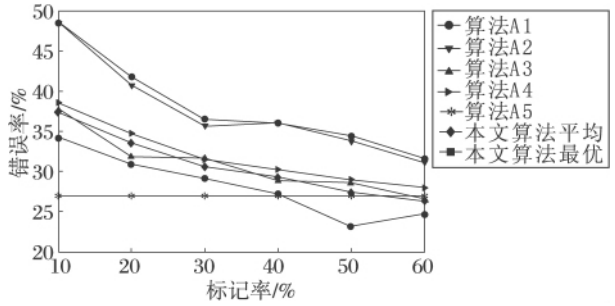
标记率	数据集	本文算法									算法 A1 (100)	算法 A2 (100)	算法 A3 (100)	算法 A4 (平均)
		10	20	30	40	50	60	70	80	90				
20	TTT	25.73	26.25	26.12	27.26	**25.12**	25.89	26.57	26.25	27.07	30.62	29.55	36.52	28.21
	Wine	22.67	22.34	20.41	19.26	21.47	19.88	**17.54**	18.68	21.14	21.92	21.92	22.42	22.26
	Lymp2	33.98	33.25	33.31	36.06	32.10	**30.87**	32.37	35.53	34.13	41.82	40.83	31.87	34.67
	Iono	21.82	20.40	21.18	21.85	**19.87**	21.16	21.60	20.86	22.98	25.21	25.78	22.41	21.96
	平均值	26.05	25.56	25.26	26.11	24.64	**24.45**	24.52	25.33	26.33	29.89	29.52	28.31	26.78
40	TTT	20.54	20.07	**18.54**	20.59	20.03	20.31	19.91	19.88	19.88	23.83	24.12	33.83	23.91
	Wine	15.72	13.99	**13.05**	13.41	13.09	13.77	13.67	13.93	13.35	16.10	16.16	18.25	16.05
	Lymp2	30.89	30.13	**27.14**	30.27	28.39	29.70	27.80	29.95	28.84	35.99	36.07	28.82	30.22
	Iono	18.30	18.29	**17.58**	19.03	18.56	18.04	18.56	18.82	18.55	22.06	22.83	20.09	18.99
	平均值	21.36	20.62	**19.08**	20.83	20.02	20.46	19.99	20.65	20.16	24.50	24.80	25.25	22.29
60	TTT	16.77	16.57	**15.85**	16.58	16.66	16.85	16.90	17.24	17.21	23.66	23.53	32.74	21.98
	Wine	12.10	10.97	11.40	9.77	**9.40**	10.30	9.53	10.90	11.04	15.21	15.16	17.30	13.03
	Lymp2	25.27	26.27	28.41	26.24	28.10	25.75	26.19	25.43	**24.63**	31.51	31.24	26.61	27.86
	Iono	16.50	16.58	16.70	16.46	17.07	15.79	16.64	**15.52**	17.52	20.68	20.28	17.90	17.24
	平均值	17.66	17.60	18.09	17.26	17.81	**17.17**	17.32	17.27	17.60	22.77	22.55	23.64	20.03



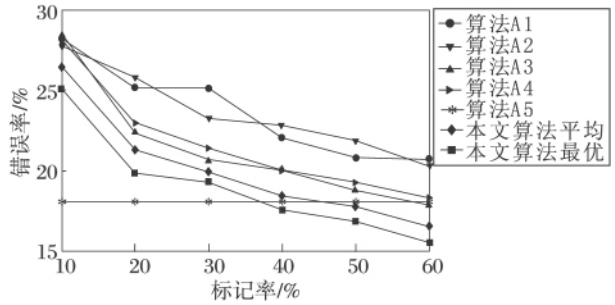
(a) TTT



(b) Wine



(c) Lymp2



(d) Iono

图 2　不同标记率下算法性能对比

Fig. 2　Comparison of performance among selected algorithms under different label rates

均值验证该结论. 综合所选数据集实验结果,当初始标记样本数量为训练集的 10% ~15% 时,主动学习质量较好,本文算法的分类性能亦较优.

为进一步比较算法的效率,各算法在其它标记率下也进行实验,结果如图 2 所示. 算法 A5 表示将所有训练集样本标注正确标记后的分类性能. 本文算法平均和最优是指本文算法在不同初标率下的平均性能和最优性能.

从图 2 可看出,除 10% 标记率下的 Wine 数据集外,本文算法的分类性能一般都优于其它算法. 表 1 中的 Wine 数据集包含 3 个不同类别的 178 样本,在 10% 标记率下,本文算法初始标记样本和主动学习样本的数量仅为 16,所以在较低初标率下初始分类器可能不包含某些类别的分类信息,以致本文算法出现较大的错误率. 同时从图 2 可看到,本文算法在某些标记率下的平均和最小错误率甚至优于数据集的最优性能(算法 A5),如 Lymp2 和 Iono(标记率为 40%). 这种现象可归结于本文算法采用双分类器,而多分类器模型的分类性能通常优于单分类器. 以上实验结果说明无标记数据能有效提升分类学习的性能,同时也显示本文算法的有效性.

6　结 束 语

在现实世界中,一些问题通常存在大量的无标记数据,而有标记数据由于标记代价过大则相对较少. 如果仅利用少量的有标记数据,其分类性能可能不理想. 本文通过对原有监督粗糙集模型进行扩展,提出可有效利用无标记数据提升分类性能的主动协同半监督粗糙集模型,解决部分标记数据的属性约简和分类学习问题. 实验仿真结果表明,本文算法不仅能显著提高部分标记数据的分类学习性能,而且在某些标记率下,本文算法甚至达到或优于数据集的最优性能. 因此,可将本文算法应用于现实部分标记数据问题,以减少无标记数据手工标注代价. 下一步将考虑研究高效的半监督属性约简算法,以提高本文算法的学习效率. 同时将本文算法应用于实际领域,进一步评估其有效性.

参 考 文 献

[1] Pawlak Z. Rough Sets. International Journal of Computer and Information Science, 1982, 11(5): 341-356

[2] Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning about Data. Dordrecht, Netherlands: Kluwer Academic Publishers, 1991

[3] Liu Qing. Rough Sets and Rough Reasoning. Beijing, China: Science Press, 2001 (in Chinese)
(刘 清. Rough 集及 Rough 推理. 北京: 科学出版社, 2001)

[4] Wang Guoyin. Rough Sets Theory and Knowledge Acquisition. Xi'an, China: Xi'an Jiaotong University Press, 2001 (in Chinese)
(王国胤. Rough 集理论与知识获取. 西安: 西安交通大学出版社, 2001)

[5] Zhang Wenxiu, Wu Weizhi, Liang Jiye, *et al*. Rough Sets Theory and Methods. Beijing, China: Science Press, 2003 (in Chinese)
(张文修,吴伟志,梁吉业,等. 粗糙集理论与方法. 北京: 科学出版社, 2003)

[6] Liang Jiye, Li Deyu. Uncertainty and Knowledge Acquisition in Information System. Beijing, China: Science Press, 2005 (in Chinese)
(梁吉业,李德玉. 信息系统中的不确定性与知识获取. 北京: 科学出版社, 2005)

[7] Miao Duoqian, Li Daoguo. Rough Sets Theory, Algorithms and Applications. Beijing, China: Tsinghua University Press, 2008 (in Chinese)
(苗夺谦,李道国. 粗糙集理论,算法与应用. 北京: 清华大学出版社, 2008)

[8] Duan Qiguo, Miao Duoqian. Jin Kaimin. A Rough Set Approach to Classifying Web Page without Negative Examples // Proc of the 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Nanjing, China, 2007: 481-488

[9] Lingras P, Chen Min, Miao Duoqian. Semi-Supervised Rough Cost/Benefit Decisions. Fundamenta Informaticae, 2009, 94(2): 1-12

[10] Gu X P, Tso S K. Applying Rough-Set Concept to Neural-Network-Based Transient-Stability Classification of Power Systems // Proc of the 5th International Conference on Advances in Power System Control, Operation and Management. Hong Kong, China, 2000: 400-404

[11] Wang Sheng, Wang Xue, Bi Daowei, *et al*. Collaborative Statistical Learning with Rough Feature Reduction for Visual Target Classification // Proc of the 5th International Joint Conference on Neural Networks. Hong Kong, China, 2008: 1151-1156

[12] Settles B. Active Learning Literature Survey. Computer Sciences Technical Report, 1648. Madison, USA: University of Wisconsin-Madison, 2009

[13] Long Jun, Yin Jianping, Zhu En, *et al*. A Survey of Active Learning. Journal of Computer Research and Development, 2008, 45(Z1): 300-304 (in Chinese)
(龙 军,殷建平,祝 恩,等. 主动学习研究综述. 计算机研究与发展, 2008, 45(Z1): 300-304)

[14] Blum A, Mitchell T M. Combining Labeled and Unlabeled Data with Co-Training // Proc of the 11th Annual Conference on Computational Learning Theory. Madison, USA, 1998: 92-100

[15] Zhou Zhihua, Wang Jue. Machine Learning and Its Application. Beijing, China: Tsinghua University Press, 2007 (in Chinese)
(周志华,王 珏. 机器学习及其应用. 北京: 清华大学出版社, 2007)

[16] Chapelle O, Schölkopf B, Zien A. Semi-Supervised Learning. Cambridge, USA: MIT Press, 2006

[17] Zhu Xiaojin. Semi-Supervised Learning Literature Survey (Revised Edition). Technical Report, 1530. Madison, USA: University of Wisconsin-Madison, 2008

[18] Liang Jiye, Gao Jiawei, Chang Yu. The Research and Advances on Semi-Supervised Learning. Journal of Shanxi University: Nature Science Edition, 2009, 32(4): 528-534 (in Chinese)
(梁吉业,高嘉伟,常 瑜. 半监督学习研究进展. 山西大学学报: 自然科学版, 2009, 32(4): 528-534)

[19] Nigam K, Ghani R. Analyzing the Effectiveness and Applicability of Co-Training // Proc of the 9th ACM International Conference on Information and Knowledge Management. McLean, USA, 2000: 86-93

[20] Goldman S, Zhou Yan. Enhancing Supervised Learning with Unlabeled Data // Proc of the 17th International Conference on Machine Learning. San Francisco, USA, 2000: 327-334

[21] Zhou Zhihua, Li Ming. Tri-Training: Exploiting Unlabeled Data Using Three Classifiers. IEEE Trans on Knowledge and Data Engineering, 2005, 17(11): 1529-1541

[22] Li Ming, Zhou Zhihua. Improve Computer-Aided Diagnosis with Machine Learning Techniques Using Undiagnosed Samples. IEEE Trans on Systems, Man and Cybernetics, 2007, 37(6): 1088-1098

[23] Balcan M F, Blum A, Yang K. Co-Training and Expansion: Towards Bridging Theory and Practice // Proc of the 19th Annual Conference on Neural Information Processing Systems. Whistler, Canada, 2005: 89-96

[24] Wang Wei, Zhou Zhihua. Analyzing Co-Training Style Algorithms // Proc of the 18th European Conference on Machine Learning. Warsaw, Poland, 2007: 454-465

[25] Muslea I, Minton S, Knoblock C. Selective Sampling with Redundant Views // Proc of the 17th National Conference on Artificial Intelligence. Austin, USA, 2000: 621-626