

一种处理部分标记数据的粗糙集属性约简算法

张 维^{1,2,3} 苗夺谦^{1,3} 高 灿^{4,5} 李 峰^{1,3}

(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)¹

(上海电力学院计算机科学与技术学院 上海 200090)²

(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)³

(深圳大学计算机与软件学院 广东 518060)⁴ (香港理工大学应用科学与纺织学院 香港)⁵

摘 要 属性约简是粗糙集理论中重要的研究内容之一,是数据挖掘中知识获取的关键步骤。Pawlak 粗糙集约简的对象一般是有标记的决策表或者是无标记的信息表。而在很多现实问题中有标记数据很有限,更多的是无标记数据,即半监督数据。为此,结合半监督协同学习理论,提出了处理半监督数据的属性约简算法。该算法首先在有标记数据上构造两个差异性较大的约简来构造基分类器;然后在无标记数据上交互协同学习,扩大有标记数据集,获得质量更好的约简,构造性能更好的分类器,该过程迭代进行,从而实现利用无标记数据提高有标记数据的约简质量,最终获得质量较好的属性约简。UCI 数据集上的实验分析表明,该算法是有效且可行的。

关键词 粗糙集,增量式属性约简,协同学习,部分标记数据,半监督学习

中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.01.005

Rough Set Attribute Reduction Algorithm for Partially Labeled Data

ZHANG Wei^{1,2,3} MIAO Duo-qian^{1,3} GAO Can^{4,5} LI Feng^{1,3}

(Department of Computer Science and Technology, School of Electronics and Information Engineering,
Tongji University, Shanghai 201804, China)¹

(Department of Computer Science and Technology, Shanghai University of Electric Power, Shanghai 200090, China)²

(The Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai 201804, China)³

(School of Computer and Software, Shenzhen University, Guangdong 518060, China)⁴

(Institute of Textiles & Clothing, The Hong Kong Polytechnic University, Hong Kong, China)⁵

Abstract Attribute reduction, as an important preprocessing step for knowledge acquiring in data mining, is one of the key issues in rough set theory. Rough set theory is an effective supervised learning model for labeled data. However, attribute reduction for partially labeled data is outside the realm of traditional rough set theory. In this paper, a rough set attribute reduction algorithm for partially labeled data was proposed based on co-training which capitalizes on the unlabeled data to improve the quality of attribute reducts from few labeled data. It gets two diverse reducts of the labeled data, employs them to train its base classifiers, and then co-trains the two base classifiers iteratively. In every round, the base classifiers learn from each other on the unlabeled data and enlarge the labeled data, so better quality reducts could be computed from the enlarged labeled data and employed to construct base classifiers of higher performance. The theoretical analysis and experimental results with UCI data sets show that the proposed algorithm can select a few attributes but keep classification power.

Keywords Rough sets, Incremental attribute reduction, Co-training, Partially labeled data, Semi-supervised learning

Pawlak 粗糙集理论^[1,2]作为一种处理不精确、不一致和不完备数据的一种重要的智能信息处理技术,已经在模式识别、机器学习、人工智能、知识获取以及数据挖掘^[3]等方面得到广泛应用。属性约简尝试从高维数据中去除不相关和冗余的特征,同时保持学习器的性能,是机器学习、模式识别和数

据挖掘中知识获取的重要处理过程,在多年的研究工作中也取得了很大的进展^[4,5]。而在较多现实应用中,如网页分类、语音识别、自然语言解析、垃圾邮件过滤等,获取有标记数据可能需要大量的人力、特殊的设备以及多次实验,代价很高。如果仅在有限的有标记数据上获取约简,通常质量不高。而

到稿日期:2015-10-09 返修日期:2015-12-13 本文受国家自然科学基金项目(61273304),2013年度高等学校博士学科点专项科研基金(20130072130004)资助。

张 维(1978—),女,博士生,主要研究方向为粗糙集、机器学习等;苗夺谦(1964—),男,教授,博士生导师,主要研究方向为粗糙集、粒计算、模式识别、人工智能等;高 灿(1983—),男,博士,主要研究方向为粗糙集、机器学习等;李 峰(1989—),男,博士生,主要研究方向为粗糙集、机器学习等。

大量无标记数据的获取相对容易,因此研究这种有少量标记数据、大量无标记数据即半监督数据的属性约简受到越来越多的关注。近年来,研究学者提出了一些半监督属性约简模型和算法,大致可以分为3类:1)基于类别标号的方法,如半监督概率PCA^[6]、分类约束降维^[7]、半监督判别分析^[8]、半监督局部Fisher判别分析^[9]和基于流形学习的半监督局部Fisher判别分析^[10]等;2)基于成对约束的方法,如约束Fisher判别分析^[11]、基于约束的半监督降维框架^[12]、约束的局部保持投影^[13]和邻域保持半监督降维^[14,15]等;3)基于其他监督信息的方法,如扩充关系嵌入^[16]、语义子空间映射^[17]等。

属性约简是粗糙集理论重要的研究内容之一,Pawlak粗糙集约简的对象一般是有标记的决策表或者是无标记的信息表。对于部分标记数据,文献[18]将部分标记数据转换为有标记数据,为每个无标记数据赋予“伪类标记”。无标记数据的决策值可能与有标记数据和其他无标记数据不同,需要保留其分辨信息,因此对每一个无标记数据赋予“伪类标记”要使其能与其他所有数据可区分。但由此计算的约简会导致更多的冗余属性。对于该问题,现有的粗糙集尚无较好的方法。

针对上述问题,本文基于半监督协同学习理论,提出了一种处理部分标记数据的粗糙集属性约简算法。首先在有标记数据上构建了两个差异性较大的约简构造基分类器,然后在无标记数据上交互协同学习,扩大有标记数据集,获得质量更好的约简,构造性能更好的分类器,该过程迭代进行,从而实现利用无标记数据提高有标记数据的约简质量,最终获得质量较好的属性约简。

本文首先介绍了关于粗糙集与半监督协同学习的相关概念,接着提出属性增量式更新算法,在此基础上给出基于协同学习的部分标记数据属性约简算法,最后基于理论分析与实验仿真验证模型的有效性。

1 基本概念

1.1 粗糙集概念

一般地,信息系统可表示为 $IS=(U,A,V,f)$,其中, U 是对象集合 $\{x_1,x_2,\dots,x_n\}$; A 是属性非空集合 $\{a_1,a_2,\dots,a_m\}$; $V=\bigcup_{a\in A}V_a$, V_a 表示属性 a 的值域; $f:U\times A\rightarrow V$ 是信息函数,指定 U 中每一个对象 x 的属性值,即对 $a\in A,u\in U,f(u,a)\in V_a$ 。如果属性集 A 可分为条件属性集 C 和决策属性集 D ,即 $A=C\cup D,C\cap D=\emptyset$,则该信息系统称为决策信息系统或决策表。上述符号在后续行文中直接引用,其含义不再赘述。

定义1 对于 $\forall x_i,x_j\in U(x_i\neq x_j)$,如果 $C_i=C_j$ 则 $D_i=D_j$,那么称 x_i 和 x_j 是一致的;否则为不一致。条件属性 $R\subseteq C$,如果 $R_i=R_j$ 则 $D_i=D_j$,那么称 x_i 和 x_j 是 R -一致的;否则称 x_i 和 x_j 是 R -不一致。

定义2 设给定一个论域 U 和论域 U 上的一个等价关系 R ,以及论域 U 的一个划分 $F=\{X_1,X_2,\dots,X_n\}$, $RF=\{RX_1,RX_2,\dots,RX_n\}$ 。定义 R -近似分类质量为 $\gamma_R(F)=(\sum card(RX_i)/card U)$ 。

Skowron和Rauszer^[19]提出了差别矩阵,利用差别矩阵存放可辨识的对象对的属性集。Hu和Cercone^[20]提出了关系数据库中简化的差别矩阵。然而,如果决策表是不一致的,

基于简化的差别矩阵的属性约简算法可能得不到Pawlak约简。文献[21]提出了改进的差别矩阵。为了算法有更大的普适性,本文采用了该定义。下面给出改进的差别矩阵定义。

定义3 在决策表 $DT=(U,C\cup D,V,f)$ 上定义差别矩阵 $M=\{m_{ij}\}$ 。

$$m_{ij}=\begin{cases} \{a\in C:f(x_i,a)\neq f(x_j,a)\},f(x_i,d)\neq f(x_j,d),x_i,x_j\in U_1,d\in D \\ \{a\in C:f(x_i,a)\neq f(x_j,a)\},x_i\in U_1,x_j\in U_2' \end{cases}$$

$$U_1=\bigcup_{i=1}^k CX_i(U),U_2=U-U_1$$

$$U_2'=\{x_i|\forall x_i\in U_2,\text{not}\exists x_j\in U_2',a\in C,f(x_i,a)=f(x_j,a)\text{ and }f(x_i,d)\neq f(x_j,d),d\in D\}$$

其中, U_2' 其实就是不一致对象集合 U_2 去掉条件属性重复的对象的集合。

1.2 半监督协同学习

半监督学习是分析和处理部分标记数据的有效方法,近年来成为机器学习的研究热点之一。相比传统的学习方法,半监督学习可以同时利用无标记数据和有标记数据,因此在理论与应用中都受到越来越多的关注。Blum和Mitchell^[22]提出的协同学习(co-training)是一种经典的半监督学习算法。该算法的前提条件是数据集有两个充分冗余的视图,即每个属性集都足以训练一个分类器,且给定标记时,每个属性集都条件独立于另一个属性集。协同学习算法在这两个视图上利用有标记数据分别训练一个分类器,用学得分类器对未标记数据进行预测,然后从每个分类器的预测结果中挑选若干置信度较高的数据加入另一个分类器的训练集,以便对方用扩大的训练集来更新分类器。迭代该过程,以此提高学习能力。

然而在实际问题中,协同学习面临的问题之一是往往不存在自然分割的两个充分且冗余视图的属性集。为了使协同学习获得较好的性能,研究人员采用了不同的策略,大致可以归为以下两类:1)尽量满足该条件,比如采取随机划分方法、遗传最优化方法或者是基于互信息、卡方统计量等评估标准使两个视图独立最大化等,如文献[23-26]的工作,但是这些方法却不能保证视图的充分性或者是视图分割不稳定;2)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代替充分冗余视图条件假设,如文献[27,28]的工作,这些方法实际上只使用了一个视图,在一定程度上放松了协同学习的约束条件,从而影响了协同学习中两个视图优势互补的性质。文献[18]提出了基于粗糙集属性约简的差异性视图分割方法,构造了粗糙协同分类模型。该模型不仅使其构造的分类器之间具有较大的差异性,同时粗糙集约简的性质也使其满足充分性,因而能有效地处理部分标记数据。后续算法将沿用该思想构建两个差异较大的属性约简。

2 属性增量式更新算法

理论上讲,决策表的属性约简的最优结果是能够找到包含条件属性数目最少的约简,也称为最小约简,它能使决策规则的数目最少,而又不损失决策表的任何信息。但在实际应用中,还应考虑问题求解的成本以及算法的计算复杂度等因素,

故而一般采用启发式算法。现有的属性约简算法主要包括基于正区域的约简算法、基于差别矩阵的约简算法和基于信息熵的约简算法等。当新增一个数据时,就需要考虑约简的变化。根据以上 3 类算法的思想可知,基于差别矩阵的约简算法最能直观地表示出约简的变化,故本文基于差别矩阵构建属性的增量式更新算法。

2.1 算法的基本思路

现在的问题是根据已有的数据集构建差别矩阵 M , 获取了属性约简 R , 新增数据 $aSample$ 后, 差别矩阵如何更新, 以及更新后怎样求得新的属性约简 R' 。

根据文献[21], 新增数据 $aSample$ 后, 差别矩阵的更新情况如下:

1) $aSample$ 与 U_2' 不一致, 则 M 不变;

2) $aSample$ 与 U_1 不一致, 即 $\exists y \in U_1$ 使得 $aSample$ 和 y 是不一致的, 则删除数据 y 所在的行, 修改数据 y 所在的列, $U_2' = U_2' \cup \{y\}, U_1 = U_1 - \{y\}$;

3) $aSample$ 与 $U_1 \cup U_2'$ 一致, 则在 M 中增加数据 $aSample$ 对应的行和列, $U_1 = U_1 \cup aSample$ 。

新增数据 $aSample$ 后, 新增的差别矩阵 $M_{add} = \{m_{xi_add}\}$ 的计算方法为:

$$m_{xi_add} = \begin{cases} \{a \in C: f(aSample, a) \neq f(x_i, a)\}, f(aSample, d) \neq f(x_i, d), x_i \in U_1, d \in D \\ \{a \in C: f(aSample, a) \neq f(x_i, a)\}, x_i \in U_2' \end{cases}$$

求解的目的是获取属性集, 使得新的差别矩阵元素 m_{xi_add} 与其交集均不为 $\emptyset$ 。解决的思路是: 在 M_{add} 与 R 相交为 $\emptyset$ 的元素中, 首先计算出新增的核元素 $core_add$, 再优先选择出现频率较高的属性加入 R_{add} , 即 $R' = R \cup R_{add}$, 直至最终 R' 与 m_{xi_add} 的交集均不为 $\emptyset$ 。据此给出算法 1。

算法 1 新增约简属性求解算法

输入: (1) 增加的差别矩阵 M_{add}

(2) $R \in R(U)$ // $R(U)$ 是 U 上所有的约简集合

输出: 增加的约简属性 R_{add}

1. 对 $\forall c_{ij} \in M_{add}$, 如果 $c_{ij} \cap R \neq \emptyset$, 则 $c_{ij} = \emptyset$;
2. 对 $\forall c_{ij}$, 如果都有 $c_{ij} = \emptyset$, 则 $R_{add} = \emptyset$, 转到第 7 步, 否则转到第 3 步;
3. 计算此时 M_{add} 的核 $core_add$, $R_{add} = core_add$;
4. 对 $\forall c_{ij} \in M_{add}$, 如果 $c_{ij} \cap R_{add} \neq \emptyset$, 则 $c_{ij} = \emptyset$;
5. 对 $\forall c_{ij}$, 如果都有 $c_{ij} = \emptyset$, 则转到第 7 步, 否则转到第 6 步;
6. 计算当前 M_{add} 中每个属性出现的次数, 选出现次数最多的元素 a_m , 令 $R_{add} = R_{add} \cup a_m$, 转到第 4 步;
7. 返回 R_{add} 。

根据前述差别矩阵的更新情况以及算法 1, 给出属性约简增量式更新算法如下。

算法 2 属性约简增量式更新算法

输入: (1) $U_1 = \bigcup_{i=1}^k CX_i(U), U_2 = U - U_1, U_2' = \text{delrep}(U_2)$

(2) $R \in R(U)$ // $R(U)$ 是 U 上所有的约简集合

(3) 新增数据 $aSample$

输出: 更新的属性约简 R'

1. $R_{add} = \emptyset$;
- ① 如果 $aSample$ 与 $y(y \in U_2')$ 不一致;

$R_{add} = \emptyset$, 转到第 2 步;

② 如果与 $y(y \in U_1)$ 不一致;

计算新增 $aSample$ 后新增的差别矩阵 M_{add} , 调用算法 1, 得到 R_{add} , 转到第 2 步;

③ 如果 $aSample$ 与 $y(y \in U_1)$ 是 R -一致的;

$R_{add} = \emptyset$, 转到第 2 步;

否则, 计算新增 $aSample$ 后新增的差别矩阵 M_{add} , 调用算法 1, 得到 R_{add} ; 转到第 2 步;

2. 返回 $R' = R \cup R_{add}$ 。

2.2 算法复杂度分析

对新增数据 $aSample$, 最坏的情况就是需要根据差别矩阵的变化计算新增的约简属性。该情况下的运行时间包括:

1) 判别 $aSample$ 与其他数据是否为 R -一致, 其时间复杂度为 $O(|U_1| + |U_2'|)$ 。

2) 求解新增的差别矩阵元素的时间, 第 2 种情况为 $O(|C||U_1|)$, 第 3 种情况为 $O(|C|(|U_1| + |U_2'|))$ 。

3) 求解新增差别矩阵的核的时间, 时间复杂度为 $O(1 * (|U_1| + |U_2'|))$ 。

4) 将新增差别矩阵中包含核元素的矩阵元素置为 $\emptyset$, 时间复杂度为 $O(|U_1| + |U_2'|)$ 。

5) 挑选新增差别矩阵中出现频率最高的属性, 并将新增差别矩阵中包含该属性的元素置为 $\emptyset$, 该过程反复执行, 直至每个矩阵元素为 $\emptyset$ 。最坏时间复杂度为

$$O(|C-R|(|U_1| + |U_2'|)) + O(|C-R|-1)(|U_1| + |U_2'|-1) + \dots = O(|C-R|(|U_1| + |U_2'|))$$

因此, 在最坏情况下, 算法 2 的时间复杂度为 $O(|C-R|^2(|U_1| + |U_2'|))$ 。

3 基于协同学习的部分标记数据属性约简

3.1 基于协同学习的部分标记数据属性约简算法

Pawlak 粗糙集仅能处理有标记数据的决策表或者是无标记数据的信息表。对于部分标记数据, 仅在有限的有标记数据集 L 上难以获取高质量的属性约简。结合半监督协同学习理论, 可以利用大量无标记数据来提高属性约简的质量。

协同学习假设数据集有两个充分冗余的视图, 即每个属性集都足以训练一个分类器, 且给定标记时, 每个属性集都条件独立于另一个属性集。然而在实际问题中, 往往不存在自然分割的两个充分且冗余视图的属性集。

为了满足半监督协同训练对两个视图充分冗余的要求, 首先需要对属性集进行分割, 获取两个差异较大的约简作为两个视图。一般来讲, 一个决策表的知识约简不是唯一的, 即对同一个决策表可能存在多个约简。理论上, 两约简相同属性越少, 则差异性越大。在获得一个约简后, 优先考虑条件属性中排除该约简后余下的属性, 以此获得两个差异性属性约简。基于差别矩阵算法获取一个约简后再计算另一约简时, 采用了如下的策略: 首先包括核属性, 然后优先考虑第一个约简与核属性以外的属性, 按照频率的高低依次加入约简, 使得约简与所有的差别矩阵元素交集均不为空; 如果还无法满足该条件, 再将第一个约简的属性作为候选属性。

至此, 从原始属性集上获取了两个差异性较大的约简, 由此获取两个充分冗余的视图构建基分类器 f_1 与 f_2 。然后迭代协同学习。在每一轮的学习过程中, 每个分类器对无标记

数据进行标记,并把标记后的数据加入另一个分类器的有标记训练集中,另一分类器也采用同样的方式,以扩大两个分类器的有标记样本数量,获取质量更好的属性约简,再在更新后的训练集与属性约简上重新训练分类器,得到性能更好的分类器。如此迭代协同学习,利用无标记数据提升最终的属性约简质量。算法的示意图如图 1 所示。

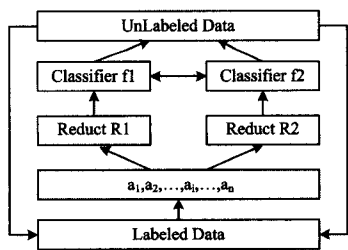


图 1 基于协同学习的部分标记数据属性约简算法框架

下面给出基于协同学习的部分标记数据属性约简算法。

算法 3 基于协同学习的部分标记数据属性约简算法

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

输出: 约简 R_1', R_2'

1. 在有标记数据 L 上将条件属性子集分割为充分且差异较大的属性子集 R_1 与 R_2 ;
2. 设两分类器的训练集 $L_1=L_2=L$, 并分别在 R_1 与 R_2 上训练分类器 f_1 与 f_2 , 并设 $N_1=N_2=N$;
3. 如果 $N_1=\emptyset$ 或 $N_2=\emptyset$, 重复以下步骤:
 - ① 遍历 N_1 , 分类器 f_2 对无标记样本进行标记, 并把标记后的样本加入 f_1 的训练集 L_1 , 在更新的训练集 L_1 上应用算法 2 获取新增的约简属性 R_1' , 同时更新 $N_1=U-L_1$;
 - ② 遍历 N_2 , 分类器 f_1 对无标记样本进行标记, 并把标记后的样本加入 f_2 的训练集 L_2 , 在更新的训练集 L_2 上应用算法 2 获取新的属性约简 R_2' , 再训练分类器 f_2 , 同时更新 $N_2=U-L_2$;
 - ③ 如果 $R_1=R_1'$, 则在更新的训练集 L_1 上再训练分类器 f_1 , 否则在更新的训练集 L_1 以及更新的属性 R_1' 上再训练分类器; 如果 $R_2=R_2'$, 则在更新的训练集 L_2 上再训练分类器 f_2 , 否则在更新的训练集 L_2 以及更新的属性 R_2' 上再训练分类器 f_2 ;
4. 返回 R_1', R_2' 。

最初仅在标记数据集 L 上获取的属性约简质量可能不高,但是在后面迭代学习的过程中,每个分类器对无标记数据进行标记,并传播至另一有标记数据集以扩大有标记数据集,从而重新学习的分类器的分类性能会更高,那么在下一轮的学习过程中,对无标记数据的分类准确率也会更高,因此扩大的有标记数据集的质量也就更高,从而最终得到质量较高的属性约简。需要说明的是,在每次迭代时加入的无标记数据保持了样本类别比例的分布。

设 $|L|=l, |N|=n, |C|=m$, 设单分类器在 m 个属性 l 个样本上训练的时间为 t 。在算法 3 步骤 3 的迭代学习过程中,两个分类器只有能以较高置信度标记的样本可以利用。假定在最坏的情况下,每次迭代时这样的样本只有一个,那么最多需要迭代 n 次,即两个分类器在迭代过程中在 m 个属性 $(l+1)$ 个样本上训练 n 次。因此,算法 3 的最坏时间复杂度为 $O(nt)$ 。

3.2 理论分析

协同学习最初在有标记数据集 L 上,分别训练两个分类器 f_1 与 f_2 。 f_1 与 f_2 能以较高置信度标记的无标记样本数量

用 n_{n1} 与 n_{n2} 表示。在协同学习没有开始以前,两个分类器 f_1 与 f_2 训练集的样本数为 l 。第 1 次协同学习过程中,分类器 f_2 将 n_{n2} 个标记置信度较高的样本进行标记,并传播至 f_1 的训练集中, f_1 训练集的样本数增至 $l+n_{n2}$ 。类似地, f_2 训练集的样本数增至 $l+n_{n1}$ 。再次训练分类器,可能出现原有的无法以较高置信度标记的样本能以较高置信度标记的情形,于是进行第 2 次协同训练。如此迭代训练,在理想情况下, f_1 与 f_2 均能以较高置信度标记其属性空间的任意样本。

那么在迭代学习过程中,在扩大的有标记数据集上,按照算法 3 得到的属性约简是否比原来的属性约简质量更高呢?下面给出命题 1。

命题 1 给定部分标记数据 $PS=(U=L \cup N, A=C \cup D, V, f)$, R 是有标记数据集 L 上的属性约简。新增有标记数据 $aSample$ 加入 L 后,扩大的有标记数据集表示为 L' ($L'=L+aSample$), R' 是按照算法 3 获取的有标记数据集 L' 上的属性约简。则 $\gamma_{R'}(U') \geq \gamma_R(U')$ ($U'=U+aSample$)。

证明:在算法 3 中,属性约简是基于改进的差别矩阵(定义 3)获取的。对于新增数据 $aSample$,差别矩阵的更新分为 3 种情况。

1) $aSample$ 与 $x(x \in U_2')$ 不一致,属性约简没有发生变化,则 $\gamma_{R'}(U') = \gamma_R(U')$ 成立。

2) $aSample$ 与 $x(x \in U_1)$ 不一致,这里分为两种情况。

① $|[x(U)]_R| = 1$ 。

$aSample$ 增加至 L' , $|[x(U')]_R| = 1+1$; U_1 中的所有数据都是一致的。 R' 是 L' 上获取的属性约简,与更新后的差别矩阵中每一个元素都有交集,那么 $|[x(U')]_R| = |[x(U')]_C| = 1+1$ 。 $|[x(U')]_R| = |[x(U')]_C|$ 成立。 $aSample$ 与 $y(y \in U_1-x)$ 一致,因此, $\sum |R'(X_i)| = \sum |R(X_i)|$ 成立。i. e. $\gamma_{R'}(U') = \gamma_R(U')$ 。

② $|[x(U)]_R| = n(n > 1)$ 。

(反证法)假设在这种情况下, $\gamma_{R'}(U') < \gamma_R(U')$, 根据定义 2, 那么 $\sum |R'(X_i)|/|U'| < \sum |R(X_i)|/|U'|$, i. e. $\sum |R'(X_i)| < \sum |R(X_i)|$ (formula 1)。 $aSample$ 增加至 L' , $|[x(U')]_R| = n+1$; U_1 中的所有数据都是一致的, R' 是 L' 上获取的属性约简,与更新后的差别矩阵中每一个元素都有交集,那么 $|[x(U')]_R| = |[x(U')]_C| = 1+1$ 。 $n > 1$, $|[x(U')]_R| > |[x(U')]_C|$ 成立。 $aSample$ 与 $y(y \in U_1-x)$ 一致。因此 $\sum |R'(X_i)| > \sum |R(X_i)|$ 成立。这与 formula 1 矛盾。假设不成立。

3) $aSample$ 与 $x(x \in U_1 \cup U_2')$ 一致,这里分为两种情况。

① $aSample$ 与 x 是 R -一致的,属性约简没有发生变化,则 $\gamma_{R'}(U') = \gamma_R(U')$ 成立。

② $aSample$ 与 x 是 R -不一致的。

(反证法)假设在这种情况下, $\gamma_{R'}(U') < \gamma_R(U')$ 。根据定义 2, 那么 $\sum |R'(X_i)|/|U'| < \sum |R(X_i)|/|U'|$ 成立, i. e. $\sum |R'(X_i)| < \sum |R(X_i)|$ 。 R' 是 L' 上获取的属性约简,与更新后的差别矩阵中每一个元素都有交集,那么 $[x(U')]_R \subseteq X_i$ 成立。由于 $\sum |R'(X_i)| < \sum |R(X_i)|$, 可得 $[x(U')]_R \subseteq X_i$ 。这和 $aSample$ 与 x 是 R -不一致相矛盾。所以假设不成立。

以上 3 种情况下 $\gamma_{R'}(U') \geq \gamma_R(U')$ ($U'=U+aSample$) 成立,命题得证。

因此,算法 3 获取部分标记数据的属性约简是合理且有效的。

4 实验仿真

4.1 实验设置

实验选用了 5 个 UCI^[29] 标准数据集。详细信息如表 1 所列。数据集 Iono, Wine 和 WPBC 的连续属性运用三等频方法^[30] 进行离散化预处理。

表 1 实验数据集			
数据集	属性数目	样本数目	类别数
Wine	13	178	3
Ionosphere (Iono)	34	351	2
Vote	16	434	2
Wisconsin Diagnostic Breast Cancer (WDBC)	30	569	2
Chess-kr-vs-kp(Chess)	36	3196	2

实验采用 10 重交叉验证方法划分训练集与测试集。每一重按标记率将训练集随机划分为有标记集 L 与无标记集 N 。考虑到样本次序的影响,实验将样本打乱进行了 10 次随机 10 重交叉验证划分。

4.2 部分标记数据的属性约简结果分析

为了验证部分标记数据的属性约简算法的效率,实验收集了标记率 $\alpha=10\%$ 下各数据集的约简信息,详细情况如表 2 所列。

表 2 标记率 $\alpha=10\%$ 下数据集的约简信息						
数据集	属性数目	真实约简	最小值	最大值	平均值	近似率 (%)
Wine	13	5	7	9	8	0.7142
Iono	34	8	9	14	12	0.8889
Vote	16	12	9	13	12	0.7500
WDBC	30	8	6	9	8	0.7500
Chess	36	29	19	29	25	0.6897
Avg.	26	12	10	15	13	0.7586

表中“最小值”、“最大值”和“平均值”表示在标记率 $\alpha=10\%$ 下 10 重交叉验证过程中部分标记数据属性约简数目的最小值、最大值和平均值;“真实约简”表示在标记率 $\alpha=100\%$ 下数据集的约简数目;表中“近似率”是部分标记数据属性约简数目的最小值与真实约简的比值,表示部分标记数据属性约简与真实约简的相似程度,近似率越高,约简算法越有效。从表中可看出,约简算法在数据集 Iono 上获得的约简与真实约简很接近;在其他数据集上,算法也取得了较好的效果。

4.3 在属性约简空间上构建的分类器性能分析

为了验证属性约简的质量,实验收集了下面 3 种情况下求得的属性约简个数以及在此基础上构建的分类器性能。

标记率 $\alpha=100\%$ 下数据集的属性约简,详细情况如表 3 所列。

标记率 $\alpha=10\%$ 下有标记数据的属性约简,详细情况如表 4 所列。

标记率 $\alpha=10\%$ 下文献[18]提出的算法(简称为 RBT-PD)获得的部分标记数据的属性约简及分类性能,详细情况如表 5 所列。

标记率 $\alpha=10\%$ 下部分标记数据的属性约简,即本文算法计算的属性约简,详细情况如表 6 所列。

实验显示了采用 CART 和 Bayes 两种分类算法在上述 3 种不同的属性约简空间的分类性能,各数据集进行 10 重交叉验证,最终性能取平均值,分别显示在表 3—表 6 中。其中 N_1, N_2 分别表示原始数据所有属性的个数和计算的属性约简的个数。分类性能 1 和分类性能 2 分别表示在原始所有属性与计算的属性约简空间构造的分类器的分类性能。

表 3 标记率 $\alpha=100\%$ 数据集的属性约简及分类性能						
数据集	属性个数		CART		Bayes	
	N_1	N_2	分类性能 1	分类性能 2	分类性能 1	分类性能 2
Wine	13	5	0.9017	0.9241	0.9533	0.9423
Iono	34	8	0.8559	0.8676	0.8459	0.7434
Vote	16	12	0.9664	0.9636	0.9052	0.9220
WDBC	30	8	0.9467	0.9463	0.9392	0.9494
Chess	36	29	0.9942	0.9905	0.8780	0.8898
Avg.	26	12	0.9330	0.9384	0.9043	0.8894

表 4 标记率 $\alpha=10\%$ 有标记数据的属性约简及分类性能						
数据集	属性个数		CART		Bayes	
	N_1	N_2	分类性能 1	分类性能 2	分类性能 1	分类性能 2
Wine	13	3	0.9017	0.7016	0.9533	0.7705
Iono	34	3	0.8559	0.7143	0.8459	0.6981
Vote	30	3	0.9467	0.8905	0.9392	0.8888
WDBC	16	3	0.9664	0.8770	0.9052	0.8780
Chess	36	12	0.9942	0.9433	0.8780	0.8399
Avg.	26	5	0.9330	0.8253	0.9043	0.8151

表 5 标记率 $\alpha=10\%$ 下算法 RBTPD 获得的部分标记数据的属性约简及分类性能

数据集	属性个数		CART		Bayes	
	N_1	N_2	分类性能 1	分类性能 2	分类性能 1	分类性能 2
Wine	13	8	0.9017	0.6713	0.9533	0.8437
Iono	34	13	0.8559	0.7095	0.8459	0.7194
Vote	30	12	0.9467	0.8879	0.9392	0.8996
WDBC	16	15	0.9664	0.9040	0.9052	0.8888
Chess	36	31	0.9942	0.9484	0.8780	0.8317
Avg.	26	16	0.9330	0.8242	0.9043	0.8366

表 6 标记率 $\alpha=10\%$ 下部分标记数据的属性约简及分类性能

数据集	属性个数		CART		Bayes	
	N_1	N_2	分类性能 1	分类性能 2	分类性能 1	分类性能 2
Wine	13	8	0.9017	0.74062	0.9533	0.8973
Iono	34	12	0.8559	0.74335	0.8459	0.7263
WDBC	30	8	0.9467	0.9032	0.9392	0.9196
Vote	16	12	0.9664	0.9420	0.9052	0.9026
Chess	36	25	0.9942	0.9508	0.8780	0.6622
Avg.	26	13	0.9330	0.8560	0.9043	0.8216

从表 3 可以看出,如果有足够多的有标记数据,求得的属性约简可以减少属性个数并且可以保持甚至提高分类性能。然而,在有标记数据个数有限的情况下,如表 4 所列,标记率仅为 10% ,获得的属性约简个数较少,并且以此构造的分类器分类性能也下降了很多。比较表 4、表 5 和表 6,可以发现无标记数据的利用大大提高了属性约简的质量,与表 1 中标注率为 100% 的情况很接近。但对比表 5 和表 6,可以发现大多数情况下本文算法计算得来的约简个数较少并且分类器性能比算法 RBTPD 要好。实验结果说明,文中提出的算法可以有效求得部分标记数据的属性约简。

为了进一步比较约简的质量,各算法在其他标记率下也进行了实验,结果如图 2 所示。其中,各图中的“初始性能”代表不利用无标记数据的传统粗糙集方法性能;“最优性能”表示将所有训练集数据标注正确标记后传统粗糙集方法的性能(即标记率为 100% 下传统粗糙集方法的性能)。

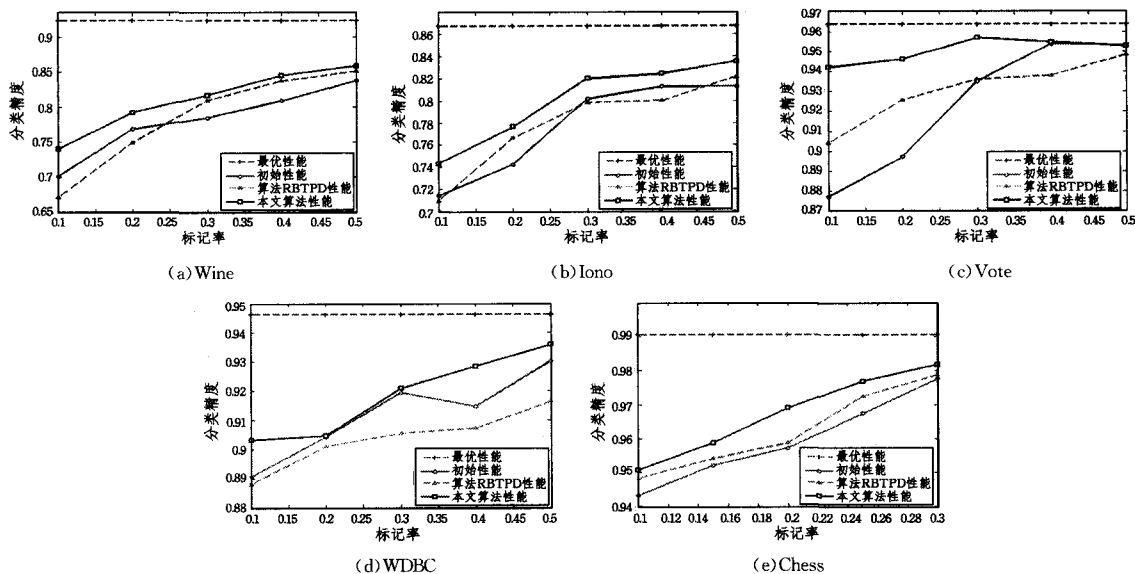


图2 不同标记率下算法的性能对比

从图2可以看出,随着标记率的增大以及有标记数据的增多,各算法的性能都有不同程度的提升;仅在基于有标记数据上获得的约简空间训练的分类器和基于算法RBTPD获得的约简空间训练的分类器性能也出现了增大的情况,但最终性能还是差于基于本文算法获得的约简空间训练的分类器性能,这也反映出本文算法求得的约简的质量较高。

仅在有关标记数据上计算约简,由于有标记数据的数量有限,因此求得的约简很难保持真实数据的区分能力;而算法RBTPD为每个无标记数据赋予的“伪类标记”由于需要保留其分辨信息,因此对每一个无标记数据赋予“伪类标记”要使其能与其他所有数据可区分,虽然保证了约简后决策表的分类能力不变,但这样也导致了约简可能会有更多的冗余属性,反而性能不如初始性能,如图中的数据集Wine, Iono, Vote和WDBC。本文算法利用了两个分类器协同学习,每个分类器可通过另一分类器标注有有用的无标记样本来提升分类性能,分类器原不可预测的数据在协同学习后将变得可预测,因此扩大了有标记数据集,进而求得了质量更好的约简,从而也就能训练出性能更好的分类器。该过程迭代进行,最终能获得质量较好的约简,从而训练的分类器有较好的分类性能。

结束语 现实问题存在着大量部分标记数据,即少量的有标记数据和大量的无标记数据。传统的粗糙集理论尚无较好的方法获取其约简。本文基于半监督协同学习理论,提出了处理部分标记数据的属性约简算法,理论分析与实验结果表明该算法是合理且有效的。但在实验的过程中也发现无标记数据的利用并非总是能提高学习性能,反而可能会使学习性能下降或者提高的性能很有限,在后续工作中,将会关注无标记数据的安全利用问题。

参考文献

- [1] PAWLAK Z. Rough sets[J]. International Journal of Computer and Information Science, 1982, 11(5): 341-356.
- [2] PAWLAK Z. Rough sets; Theoretical Aspects of Reasoning about Data[M]. Dordrecht, Netherlands: Kluwer Academic Publishers, 1991.
- [3] WANG Guo-yin, YAO Yi-yu, YU Hong. A survey on rough set theory and applications [J]. Chinese Journal of Computers, 2009, 32(7): 1229-1246. (in Chinese)

2009, 32(7): 1229-1246. (in Chinese)

王国胤, 姚一豫, 于洪. 粗糙集理论与应用研究综述[J]. 计算机学报, 2009, 32(7): 1229-1246.

- [4] THANGAVEL K, PETHALAKSHMI A. Dimensionality reduction based on rough set theory: A review[J]. Applied Soft Computing, 2009, 9(1): 1-12.
- [5] MIAO D Q, ZHAO Y, YAO Y Y, et al. Relative reducts in consistent and inconsistent decision tables of the Pawlak rough set model[J]. Information Sciences, 2009, 179(24): 4140-4150.
- [6] TIPPING E, BISHOP C M. Probabilistic principal component analysis[J]. Journal of the Royal Statistical Society, 1999, 61(3): 611-622.
- [7] COSTA J A, Hero A O. Classification constrained dimensionality reduction[J]. Statistics, 2008, 5: 1077-1080.
- [8] CAI D, HE X, HAN J. Semi-Supervised discriminant analysis [C]// International Conference Computer Vision. 2007: 1-7.
- [9] SUGIYAMA M, IDE T, NAKAJIMA S, et al. Semi-Supervised local fisher discriminant analysis for dimensionality reduction [J]. Machine Learning, 2008, 78(1/2): 35-61.
- [10] CHATPATANASIRI R, KIJSIKUL B. A unified semi-supervised dimensionality reduction framework for manifold learning [J]. Neurocomputing, 2010, 73(10-12): 1631-1640.
- [11] BAR-HILLEL A, HERTZ T, SHENTAL N, et al. Learning a Mahalanobis Metric from Equivalence Constraints[J]. Journal of Machine Learning Research, 2005, 6(2): 937-965.
- [12] ZHANG D, ZHOU Z, CHEN S. Semi-Supervised dimensionality reduction[C]// SIAM Conference on Data Mining. Minneapolis, 2007: 629-634.
- [13] CEVIKALP H, VERBEEK J J, et al. Semi-Supervised Dimensionality Reduction Using Pairwise Equivalence Constraints[C]// The Third International Conference on Computer Vision Theory and Applications. 2008: 489-496.
- [14] WEI J, PENG H. Neighbourhood preserving based semi-supervised dimensionality reduction[J]. Electronics Letters, 2008, 44(20): 1190-1191.
- [15] BAGHSHAH M S, SHOURAKI S B. Semi-Supervised Metric Learning Using Pairwise Constraints[C]// The International Joint Conference on Artificial Intelligence. 2009: 1217-1222.

- [16] LIN Y, LIU T, CHEN H. Semantic manifold learning for image retrieval[C]// ACM International Conference on Multimedia. 2005; 249-258.
- [17] YU J, TIAN Q. Learning image manifolds by semantic subspace projection[C]// ACM International Conference on Multimedia. 2006; 297-306.
- [18] MIAO D, GAO C, ZHANG N, et al. Diverse reduct subspaces based co-training for partially labeled data[J]. International Journal of Approximate Reasoning, 2011, 52(8): 1103-1117.
- [19] SKOWRON A, RAUSZER C. The discernibility matrices and functions in information systems[M]// Slowinski R, ed. Intelligent Decision Support, Handbook of Applications and Advances of the Rough Sets Theory. Dordrecht, Kluwer, 1992.
- [20] HU X H, CERCONI N. Learning in relational databases: A rough set approach[J]. International Journal of Computational Intelligence, 1995, 11(2): 323-338.
- [21] YANG Ming. An incremental updating algorithm for attribute reduction based on improved discernibility matrix[J]. Chinese Journal of Computers, 2007, 30(5): 815-822. (in Chinese)
杨明. 一种基于改进差别矩阵的属性约简增量式更新算法[J]. 计算机学报, 2007, 30(5): 815-822.
- [22] BLUM A, MITCHELL T. Combining labeled and unlabeled data with co-training[C]// Proc of the 11th Annual Conf on Computational Learning Theory. New York: ACM, 1998; 92-100.
- [23] FEGER F, KOPRINSKA I. Co-training using RBF nets and different feature splits[C]// Proc of the 2006 Int Joint Conf on Neural Networks. Piscataway, NJ: IEEE, 2006; 1878-1885.
- [24] YASLAN Y, CATALTEPE Z. Co-training with relevant random subspaces[J]. Neurocomputing, 2010, 73(10-12): 1652-1661.
- [25] TANG Huan-ling, LIN Zheng-kui, LU Ming-yu, et al. An advanced co-training algorithm based on mutual independence and diversity measures[J]. Journal of Computer Research and Development, 2008, 45(11): 1874-1881. (in Chinese)
唐焕玲, 林正奎, 鲁明羽, 等. 一种结合独立性模型与差异评估的 Co-Training 改进方案[J]. 计算机研究与发展, 2008, 45(11): 1874-1881.
- [26] SALAHELDIN A, GAYAR N El. New feature splitting criteria for cotraining using genetic algorithm optimization[C]// Proc of the 9th Int Workshop on Multiple classifier systems. Berlin: Springer, 2010; 22-32.
- [27] GOLDMAN S, ZHOU Yan. Enhancing supervised learning with unlabeled data[C]// Proc of the 17th Int Conf on Machine Learning. San Francisco: Morgan Kaufmann, 2000; 327-334.
- [28] ZHOU Zhi-hua, LI Ming. Tri-training: Exploiting unlabeled data using three classifiers[J]. IEEE Trans on Knowledge and Data Engineering, 2005, 17(11): 1529-1541.
- [29] BLAKE C, MERZ C J. UCI Repository of machine learning databases[OL]. <http://archive.ics.uci.edu/ml/datasets.html>.
- [30] ØHRN A, KOMOROWSKI J. ROSETTA: A rough set toolkit for analysis of data[C]// Proceedings of the 3rd International Joint Conference on Information Sciences. Durham, NC, USA: 1997; 403-407.
- (上接第6页)
- [40] JUSTESEN P. On independent circuits in nite graphs and a conjecture of Erdős and Pósa[J]. Annals of Discrete Mathematics, 1989, 41: 299-306.
- [41] WANG Hong. Independent cycles with limited size in a graph[J]. Graphs and Combinatorial, 1994, 10: 271-281.
- [42] GELEN J, GERARDS B, REED B, et al. On the odd-minor variant of Hadwiger's conjecture[J]. Journal of Combinatorial Theory, Series B, 2009, 99: 20-29.
- [43] METZLAR A, MURTY U S R. Disjoint Circuits in Planar Digraphs[J]. Journal of Combinatorial, Theory Series B, 1993, 57(2): 228-238.
- [44] KANN V. Maximum bounded 3-dimensional matching is MAX SNP-complete[J]. Information Processing Letters, 1991, 37: 27-35.
- [45] AUSIELLO G, CRESCENZI P, Gambosi G, et al. Combinatorial Optimization Problems and Their Approximability Properties [M]. Berlin: Springer, 1999.
- [46] HURKENS C A J, SCHRIJVER A. On the size of systems of sets every t, of which have an SDR, with an application to the worstcase ratio of heuristics for packing problems[J]. SIAM Journal on Discrete Mathematics, 1989, 2(1): 68-72.
- [47] EIMABROUK N. Genome Rearrangement by Reversals and Insertions/Deletions of Contiguous Segments[C]// Combinatorial Pattern Matching, 11th Annual Symposium. Montreal, Canada, 2000; 222-234.
- [48] BAFNA V, PEVZNER P A. Genome Rearrangements and Sorting by Reversals[J]. SIAM Journal on Computing, 1996, 25(2): 272-289.
- [49] SHAO Ming-fu, LIN Yu. Approximating the edit distance for genomes with duplicate genes under DCJ, insertion and deletion [J]. BMC Bioinformatics, 2012, 13(S-19): S13.
- [50] SHAO Ming-fu, LIN Yu, BERNARD M. An Exact Algorithm to Compute the DCJ Distance for Genomes with Duplicate Genes [C]// Research in Computational Molecular Biology 19th Annual International Conference. Pittsburgh, PA, USA, 2014; 280-292.
- [51] CHEN Xin, ZHENG Jie, FU Zheng, et al. Assignment of orthologous genes via genome rearrangement[J]. IEEE/ACM Transactions on Computational Biology Bioinformatics, 2005, 2(4): 302-315.
- [52] LI Shao-hua, WANG Jian-xin, FENG Qi-long, et al. Kernelization techniques and its application to parameterized computation [J]. Journal of Software, 2009, 20(9): 2307-2319. (in Chinese)
李绍华, 王建新, 冯启龙, 等. 参数计算中核心化计算及其应用的研究进展[J]. 软件学报, 2009, 20(9): 2307-2319.
- [53] ZHOU Xing, PENG Wei. Several techniques used in parameterized computation[J]. Computer Science, 2014, 41(6A): 18-23. (in Chinese)
周星, 彭伟. 参数计算中使用的若干技术[J]. 计算机科学, 2014, 41(6A): 18-23.
- [54] LU Sheng-guo, LI Shuang, ZHU Jian-guo, et al. Survey of genome rearrangements[J]. China Biotechnology, 2013, 30(7): 108-111. (in Chinese)
卢胜国, 李霜, 朱建国, 等. 基因组重排技术应用及进展[J]. 中国生物工程杂志, 2013, 30(7): 108-111.