基于粗糙集理论的协同训练算法*

盛小春[1]，岳晓冬[2,3]

(1. 江苏理工学院 云计算与智能信息处理常州市重点实验室，江苏 常州 213001；2. 上海大学 计算机工程与科学学院，上海 200444；3. 同济大学 计算机科学与技术系，嵌入式系统与服务计算教育部重点实验室，上海 201804)

摘　要： 为了提升风险决策环境下协同训练的效果，提出了一种基于粗糙子空间的协同决策算法。首先利用粗糙集属性约简的概念，将部分标记数据属性空间分解为两差异性较大的粗糙子空间；在各子空间上训练分类器，并依据各分类器决策风险代价及隶属度将无标记数据划分为可信、噪声和待定样本。综合两分类器的分类结果，标注少量可信无标记样本后重复协同训练。从理论上分析了算法性能提升的区间界，并在 UCI 数据集上进行实验，验证了模型的有效性及效率。

关键词： 协同训练；属性约简；粗糙集；粗糙子空间；决策风险

中图分类号： TP182　　**文献标志码：** A　　**文章编号：** 1001-3695(2013)12-3546-05

doi:10.3969/j.issn.1001-3695.2013.12.007

Novel co-training algorithm based on rough sets

SHENG Xiao-chun[1], YUE Xiao-dong[2,3]

(1. *Key Laboratory of Cloud Computing & Intelligent Information Processing of Changzhou City, Jiangsu University of Technology, Changzhou Jiangsu* 213001, *China*; 2. *School of Computer Engineering & Science, Shanghai University, Shanghai* 200444, *China*; 3. *Key Laboratory of Embedded System & Service Computing of Ministry of Education, Dept. of Computer Science & Technology, Tongji University, Shanghai* 201804, *China*)

Abstract: In order to improve the performance of co-training in the context of decision with risk, this paper proposed a rough subspace-based co-training algorithm. Based on the concept of attribute reduction in rough sets, the algorithm first splitted all condition attributes of partially labeled data into two diverse rough subspaces. Then the two classifiers trained from derived rough subspaces and classified the unlabeled data into confident, noise and uncertain samples with consideration of the classification risk and membership of decision class. Finally, it labeled a few of confident samples for the two classifiers to learn from each other in iterative manner. It theoretically analyzed the performance of proposed algorithm, and empirical results on selected UCI data sets also show its effectiveness.

Key words: co-training; attribute reduction; rough sets; rough subspace; decision risk

0　引言

传统机器学习问题主要包括有标记数据的分类和无标记数据的聚类，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，以致有类别标记的数据较为稀少，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利用。如果仅在有标记数据上通过约简而产生相应的分类器，其分类预测效果可能不理想；而不利用稀有的有标记数据进行聚类，则亦造成重要信息的损失。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，因此在理论和实践中都受到越来越多的关注。

Blum 等人[1,2]提出的协同训练 co-training 是一种典型的双视图半监督学习算法。算法假设部分标记数据的条件属性能自然地分割成两个充分且独立的视图(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。然而在实际问题中，co-training 算法面临充分冗余视图的获取和无标记样本的选择两个问题。针对充分冗余视图的获取问题，Nigam 等人[3]指出协同训练条件满足时性能表现较优，而实际问题如不存在充分冗余视图，将充分大的属性集随机划分成两个视图也可能取得较好的效果。很明显，随机划分未利用任何数据信息，视图的充分性条件亦不能满足，因此随机划分策略并非总能奏效。Feger 等人[4]基于互信息概念，提出了基于图的视图独立最大化的划分算法。唐焕玲等人[5]利用互信息(MI)和卡方统计量(CHI)评估属性之间的相互独立性，提出了新的视图划分算法 PMID-MI 与 PMID-CHI，能有效地将一个属性集合划分成两个独立性较强的视图。该方法较随机划分效果好，但各视图的

收稿日期： 2013-04-03；**修回日期：** 2013-05-27　　**基金项目：** 国家自然科学基金资助项目(61103067)；常州市云计算与智能信息处理重点实验室资助项目(CM20123004)；江苏理工学院青年基金资助项目(KYY11093)

作者简介： 盛小春(1980-)，女，江苏常州人，讲师，硕士，主要研究方向为数据挖掘、机器学习和计算机应用(sxc@jsut.edu.cn)；岳晓冬(1980-)，男，山西太原人，博士(后)，主要研究方向为数据挖掘、图像处理等.

充分性没有得到保证。Salaheldin 等人[6]提出了遗传最优化的视图分割算法,自适应函数综合了视图的充分性及视图间的独立性和差异性信息,实验结果显示新方法较随机划分性能更优,但遗传划分方法存在随机不可重复性,即视图分割结果不稳定。王娇等人[7]基于随机子空间理论,提出了随机子空间协同训练算法 RASCO。随后,Yaslan 等人[8]对王娇的工作进行了改进,提出了基于相关随机子空间的协同训练算法,各子空间的质量较随机划分有较大的提升,但其充分性亦不能得到保证。此外,一些研究人员[9-11]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。

虽然异构分类器、随机子空间和重采样技术可用来训练多个具有差异性的分类器进行有效的协同训练,但在某种程度上放松了协同训练的约束条件,从而失去了一些较好的性质。如何将不存在自然分割的属性集划分成两个充分且冗余的视图,仍是一个没有完全解决的问题。

协同训练算法的质量不仅依赖于充分且冗余视图的获取,而且受所学习的无标记样本影响。已有协同训练算法的无标记样本选择大多数基于 0/1 风险函数,也即选择置信度较大的无标记样本进行标注学习。然而现实问题中不同分类决策往往带来不同的风险代价,因此无标记样本选择过程中应考虑样本的决策风险,以提升协同训练算法的质量。本文通过引入决策粗糙集理论,提出了充分且具差异性的视图获取方法;同时利用风险决策过程进行无标记样本选择,有效地提升了协同训练的效果。

1　粗糙集理论

定义 1[12]　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;$f:U\times A\to V$ 是一个映射,指定 U 中每一对象 x 的属性值,即对 $x\in U$,$a\in A$ 有 $a(x)\in V_a$。如果属性集合 A 可分为条件属性集 C 和决策属性集 D 且 $C\cap D=\varnothing$,则该信息系统称为决策表。

定义 2[12]　给定信息系统 $S=(U,A=C\cup D,V,f)$,对于任一属性子集 $B\subseteq A$,可定义不可分辨关系 $\mathrm{IND}(B)=\{(x,y)\in U\times U\mid \forall a\in B, a(x)=a(y)\}$,形成的等价类可表示为 $[x]_B=\{y\in U\mid \langle x,y\rangle\in \mathrm{IND}(B)\}$。

定义 3[12]　给定信息系统 $S=(U,A,V,f)$,设 $X\subseteq U$,对任一属性子集 $B\subseteq A$,X 关于 B 的上、下近似集分别表示为

$$\underline{B}(X)=\{x\in U\mid [x]_B\subseteq X\} \tag{1}$$

$$\overline{B}(X)=\{x\in U\mid [x]_B\cap X\neq\varnothing\} \tag{2}$$

定义 4[12]　给定信息系统 $S=(U,A,V,f)$,对任意的属性 $a\in A$,若 $\mathrm{IND}(A)=\mathrm{IND}(A-\{a\})$,则称属性 a 是不必要的,否则 a 是 A 中必要的。

定义 5[12]　给定信息系统 $S=(U,A,V,f)$,对象之间的分辨信息可表示为对称的矩阵 M,称为差别矩阵且元素项定义为 $m_{ij}=\{a\in A\mid a(x_i)\neq a(x_j)\}$。

定义 6[12]　给定信息系统 $S=(U,A,V,f)$ 及其差别矩阵 M,$P\subseteq A$,P 为信息系统 S 的属性约简 $\mathrm{RED}(A)$ 当且仅当以下条件成立:

a) $\forall r\in M, P\cap r\neq\varnothing$;

b) $\forall P'\subset P, \exists r\in M, P'\cap r=\varnothing$。

在决策粗糙集理论中,$\Omega=\{X,X^C\}$ 表示对象属于集合 X 或其补集的状态集合。$\zeta=\{a_P,a_B,a_N\}$ 分别表示当前对象属于正域、边界域或负域的可能决策。$P(X\mid[x])$ 和 $P(X^C\mid[x])$ 分别表示等价类 $[x]$ 中元素属于 X 和属于 X^C 的概率。不同状态下对象 x 采取不同决策方案的风险代价如表 1 所示。

表 1　不同决策方案在不同状态下的风险代价

集合	决策		
	a_P	a_B	a_N
X	λ_{PP}	λ_{BP}	λ_{NP}
X^C	λ_{PN}	λ_{BN}	λ_{NN}

其中:λ_{PP}、λ_{BP} 和 λ_{NP} 表示在对象属于 X 的状态下分别采取决策 a_P、a_B 和 a_N 的风险代价值。λ_{PN}、λ_{BN} 和 λ_{NN} 表示在对象属于 X^C 的状态下分别采取决策方案 a_P、a_B 和 a_N 的风险代价值。一般地,正确决策所产生的风险代价要小于错误决策的风险代价,故有 $\lambda_{PP}\leqslant\lambda_{BP}\leqslant\lambda_{NP}$ 且 $\lambda_{PN}\leqslant\lambda_{BN}\leqslant\lambda_{NN}$ 依据具体的风险代价值,可计算出三种决策的期望风险分别为

$$R(a_P\mid[x])=\lambda_{PP}P(X\mid[x])+\lambda_{PN}P(X^C\mid[x]) \tag{3}$$

$$R(a_N\mid[x])=\lambda_{NP}P(X\mid[x])+\lambda_{NN}P(X^C\mid[x]) \tag{4}$$

$$R(a_B\mid[x])=\lambda_{BP}P(X\mid[x])+\lambda_{BN}P(X^C\mid[x]) \tag{5}$$

根据贝叶斯最小风险决策原则,可得如下决策规则:

(P)　若 $R(a_P\mid[x])\leqslant\min\{R(a_N\mid[x]),R(a_B\mid[x])\}$,则 $[x]\subseteq\mathrm{POS}(X)$;

(N)　若 $R(a_N\mid[x])\leqslant\min\{R(a_P\mid[x]),R(a_B\mid[x])\}$,则 $[x]\subseteq\mathrm{NEG}(X)$;

(B)　若 $R(a_B\mid[x])\leqslant\min\{R(a_P\mid[x]),R(a_N\mid[x])\}$,则 $[x]\subseteq\mathrm{BND}(X)$。

当决策代价出现相等情况时,需引入其他信息或仲裁机制,以保证每个对象仅属于某一决策区域。

2　基于粗糙子空间的协同决策算法

2.1　基本思想及框架

协同训练算法需要两个充分且独立的视图,以训练两个分类器交互学习,然而现实数据很难满足该条件,因此限制了协同训练算法的应用。一般来说,针对同一问题往往存在不同的解决思路和方法,而且人的思维也会从不同层次和视角来分析和解决复杂问题。相应地,现实数据往往也存在多个子空间,它们都能较好地描述数据的结构,但方式却有较大的不同。实际上,这种具有差异性描述的子空间可被综合利用来解决同一问题。属性约简是粗糙集理论重要研究内容之一,能有效地将高维数据降至低维属性子空间而不造成分类信息的损失。一般来说,数据集的属性约简(称为粗糙子空间)不是唯一的,即对同一个数据集可能存在多个粗糙子空间,而每个粗糙子空间都是充分的,都能训练较好的分类器。所以可运用属性约简思想对属性空间进行分割,寻找两个最具差异性的粗糙子空间进行有效的协同训练。

另外,协同训练算法的两分类器主要利用无标记数据提升学习模型的性能。然而在现实应用问题中,无标记数据中很可能存在噪声或奇异点,而这些数据往往对协同训练带来不利的影响。一般地,相对于协同训练,无标记数据可分为可信样本、不确定样本和噪声样本。决策粗糙集理论皆在不确定环境下

综合考虑各类决策所带来的风险,进行有效的决策推理。因此,在协同训练过程中,可结合两分类器的决策及其代价,运用决策粗糙集模型指导无标记数据的选择,以保证协同训练的质量。

利用以上两思想,可构建改进的协同训练学习算法,其框架如图 1 所示。

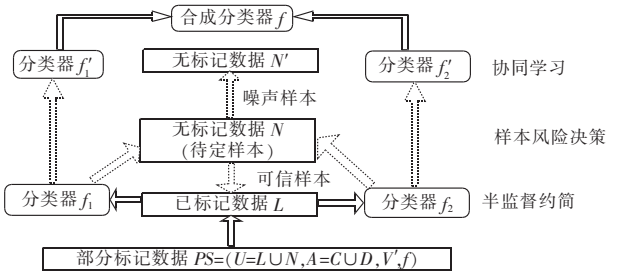

图 1 协同决策算法框架

2.2 协同训练视图分割

实现协同决策算法,首先需对属性空间进行分割,以获取两个差异性的粗糙子空间进行协同训练。粗糙集属性约简算法一般针对有标记数据或无标记数据,因此不能直接应用于部分标记数据。互信息[13]是一种度量变量相关性的有效测度,可以较好地度量属性的重要性及冗余性。实际上在有标记数据上,属性的重要性可依据互信息进行度量;而在无标记数据上,由于缺乏相关的类别信息,无法度量属性在无标记数据的重要性。但无标记数据上存在数据与属性的内在相关信息,如属性的相关性,因此在属性约简过程中可将无标记数据属性相关信息加入目标函数,以改善降维的效果。为此可将无标记数据看成具有特别标记的数据,该标记与已有的数据标记都不同,且特殊标记对象也不代表属于同类。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。而通过特殊标记化后的数据可表示为 $TS=(U',A=C\cup D,V'',f)$,其决策属性 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。

定义 7 给定部分标记数据 $PS=(U=L\cup N,A=C\cup D,V',f)$,其特殊标记化后的数据表示为 $TS=(U',A=C\cup D,V'',f)$,则任意属性 $a\in C$ 的属性重要性可表示为 $\mathrm{sig}(a,P,D)=(\mathrm{MI}(P;D)-\mathrm{MI}(P-\{a\};D))/(\mathrm{MI}(P-\{a\};\{a\})+1)$。

定义 8 给定部分标记数据 $PS=(U=L\cup N,A=C\cup D,V',f)$,其特殊标记化后的数据表示为 $TS=(U',A=C\cup D,V'',f)$,则任意属性 $a\in C$ 为部分标记数据核属性的充分必要条件为 $\mathrm{sig}(a,C,D)>0$。

定义 9 给定部分标记数据 $PS=(U=L\cup N,A=C\cup D,V',f)$,其特殊标记化后的数据表示为 $TS=(U',A=C\cup D,V'',f)$,则任意属性 $a\in C$ 为决策表的相对冗余属性的充分必要条件为 $\mathrm{sig}(a,C,D)=0$。

定义 10 给定部分标记数据 $PS=(U=L\cup N,A=C\cup D,V',f)$,其特殊标记化后的数据表示为 $TS=(U',A=C\cup D,V'',f)$,任意属性子集 $P\subseteq C$ 为部分标记数据 PS 的粗糙子空间当且仅当以下条件成立:

a) $\mathrm{MI}(P,D)=\mathrm{MI}(C,D)$;

b) $\forall a\in P,\mathrm{MI}(P-\{a\},D)\neq\mathrm{MI}(C,D)$。

基于上述定义,可构造算法 1 求得部分标记数据最优一半监督粗糙子空间。对于另一半监督粗糙子空间,理论最优方法是计算部分标记数据的所有粗糙子空间,再选择与最优半监督粗糙子空间差异度最大的一个。但部分标记数据属性过多时,计算所有粗糙子空间代价非常大。实际上可利用启发式思想求得另一半监督粗糙子空间,具体实现过程见 2.4 节。

算法 1 基于互信息的半监督粗糙子空间

输入:部分标记数据 $PS=(U=L\cup N,A=C\cup D,V',f)$。

输出:半监督粗糙子空间 Red。

a)将部分标记数据进行转换;

b)计算核属性 Core 并设置优先候选属性集为 C-Core,Red = Core;

c)如果 $\mathrm{MI}(\mathrm{Red},D)\neq\mathrm{MI}(C,D)$:

(a)计算 C-Red 中各属性的重要性;

(b)选择重要性最大的属性 a,$\mathrm{Red}=\mathrm{Red}\cup\{a\}$。

d)返回粗糙子空间 Red,结束。

算法 1 在核属性基础上,依次加入属性重要性最大的属性,直至当前属性集合的互信息与原互信息相等。假设对象 $|U|=n$,条件属性 $|C|=m$,算法 1 循环 $|C|=m$ 次必定结束,所以总体时间复杂度为 $O(m^2n)$,空间复杂度为 $O(n)$。

2.3 无标记样本风险决策

在决策粗糙集理论中,决策不仅由所处理的数据本身决定,也较大程度受其决策风险代价影响,因此不同的风险代价值可能带来完全不同的决策结果。协同决策算法从多个视角分析数据的结构信息,实质上是多模型(分类器)协同决策过程,其决策行为不由单个模型决定,而需综合考虑多个模型的决策代价,最终形成一致最小风险代价决策。给定一个无标记样本 x,各分类器根据风险代价可判定为正域、负域或边界域样本。而在不同子空间下的分类器将出现多种不同的决策结果,其具体情况如表 2 所示。

表 2 多模型协同决策

决策 1	决策 2		
	a_{P}^2	a_{B}^2	a_{N}^2
a_{P}^1	P	P	N
a_{B}^1	P	B	P
a_{N}^1	N	P	P

其中:$a_i^k(k\in\{1,2\},i\in\{\mathrm{P},\mathrm{B},\mathrm{N}\})$ 表示分类器 k 对样本 x 采取 i 决策,矩阵中元素 P、N 和 B 分别表示在两分类器决策 a_i^1、$a_j^2(i,j\in\{\mathrm{P},\mathrm{B},\mathrm{N}\})$ 下所作的最终接受、拒绝和待定决策。当分类器对样本 x 的决策一个为正域另一个为负域时,分类器的决策出现冲突,此时如加入该样本至分类器的训练集,必定会造成分类器的性能下降,因此在多模型协同决策过程中要将此类样本认定为拒绝样本(噪声样本)。而当两分类器都将两样本判定为边界域时,表明两分类器在当前状态下都不能作出明确的接受和拒绝决策,因此需将该样本加入待定区域,等待进一步决策。在其他的状态下,分类器要么是相同的接受或拒绝决策,要么是一个分类器作出确定的决策,另一分类器的决策为待定。此时可以确定决策信息传播至另一个待定的分类器,以提升各分类器的确定的决策信息。依据表 2 分器类的决策情况,可计算三类不同决策的分险代价为

$$R(b_P|x)=\min_{(i,j)\in P}\{R(b_P|a_i^1,a_j^2)\}=\min_{(i,j)\in P}\{R(a_i^1|x)+R(a_i^2|x)\}$$
$$R(b_B|x)=\min_{(i,j)\in B}\{R(b_B|a_i^1,a_j^2)\}=\min_{(i,j)\in B}\{R(a_i^1|x)+R(a_i^2|x)\}$$
$$R(b_N|x)=\min_{(i,j)\in N}\{R(b_N|a_i^1,a_j^2)\}=\min_{(i,j)\in N}\{R(a_i^1|x)+R(a_i^2|x)\}$$

其中:$R(b_P|x)$、$R(b_B|x)$和$R(b_N|x)$分别表示对样本 x 作接受、待定和拒绝决策代价;$(i,j)\in h(h\in\{P,B,N\})$表示两分类器取表 2 中的接受、待定和拒绝状态。根据贝叶斯最小风险决策原则,可选择三类决策中风险最小者作为样本的最终的决策。

2.4　算法描述

根据协同决策算法的框架,首先需对属性空间进行分割。算法 1 可求得最优粗糙子空间,另一粗糙子空间可通过对算法 1 进行适当调整而获得。根据属性空间差异性度量标准,两个粗糙子空间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粗糙子空间中的属性。换句话说,算法在求取另一粗糙子空间过程中应优先考虑未出现在最优粗糙子空间中的属性,则生成的粗糙子空间与最优粗糙子空间具有较少的共同属性,两者的差异性则较大。在两个具有较少共同属性的粗糙子空间上可构造两个差异性较大的分类器,然后利用决策粗糙集理论选择可信的无标记样本进行协同学习,以有效地提升各分类器的性能。实现协同决策的算法描述如下:

算法 2　基于粗糙子空间的协同决策算法

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。

a)计算核属性 Core,$\mathrm{Red}_1=\mathrm{Red}_2=\mathrm{Core}$;

b)粗糙子空间 Red_1 的优先候选集 At 置为 C-Core,调用算法 1 步骤 c)得粗糙子空间 Red_1;

c)粗糙子空间 Red_2 的优先候选集 At 置为 C-Red_1,调用算法 1 步骤 c)得粗糙子空间 Red_2;

d)两分类器的训练集置为 $L_1=L_2=L$,未标记数据集置为 $N_1=N_2=N$,在粗糙子空间 Red_1 和 Red_2 构造分类器 f_1 和 f_2;

e)如 N_1,N_2 不为空且 $N_1\neq N_2$,重复以下过程:

(a)利用决策粗糙集理论将无标记样本划分为可信、噪声和待定样本,并从各无标记样本集 N_1、N_2 中删除噪声样本;

(b)将分类器 f_1 的可信样本标记决策值加入 f_2 的训练集 L_2,更新分类器 f_2 及无标记样本集 $N_2=U-L_2$;

(c)将分类器 f_2 的可信样本标记决策值加入 f_1 的训练集 L_1,更新分类器 f_1 及无标记样本集 $N_1=U-L_1$;

f)输出合成的分类器 f。

假设 $|U|=n$,$|L|=l$,$|N|=t$,$|C|=m$,则初始分类器训练时间为 $O(ml)$。算法 2 对属性集进行分割的过程实际上是两次调用启发式算法 1,因此时间复杂度为 $O(m^2n)$。而决策粗糙集对无标记样本划分过程只涉及少量公式计算,时间代价可忽略。在最坏情况下,协同学习过程循环次数为 t。所以算法 2 的总体时间复杂度为 $O(ml+m^2n+m(l+t)t)\approx O(mn^2)$(一般地,$l<<n$ 且 $t<<n$)。

2.5　理论分析

算法首先选择两个具有较少共同属性的粗糙子空间构造其初始分类器,不仅满足了协同训练的充分性假设条件,也保证了分类器的差异性。而风险决策过程保证了无标记样本的质量,提升了分类器协同训练的效果。因此,新算法应能有效地处理部分标记数据。

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 $|L|=l$,$|N|=t$。在有标记数据 L 上,可分别以粗糙子空间训练分类器 f_1 和 f_2。两分类器对无标记数据 N 的预测结果可分三种情况:较大信度预测正确(correct)、较大信度预测错误(incorrect)和不确定决策(uncertain),则两分类器在 t 个无标记数据上的差异性可表示为表 3。

表 3　分类器差异矩阵

预测 1	预测 2		
	f_2 correct(c)	f_2 incorrect(i)	f_2 uncertain(u)
f_1 correct(c)	t_{cc}	t_{ci}	t_{cu}
f_1 incorrect(i)	t_{ic}	t_{ii}	t_{iu}
f_1 uncertain(u)	t_{uc}	t_{ui}	t_{uu}

其中:t_{cc}、t_{ii}和 t_{uu}分别表示两分类器都以较大信度预测正确、错误以及不确定决策的样本数目;t_{ci}和 t_{ic}表示一分类器较大信度预测正确而另一分类器较大信度预测错误的对象数目;t_{cu}和 t_{uc}表示一分类器较大信度预测正确另一分类器为不确定决策的样本数目;一分类器较大信度预测错误而另一分类器为不确定决策的样本分别以 t_{iu}和 t_{ui}表示。在协同训练之前,两分类器的正确率分别为$(t_{cc}+t_{ci}+t_{cu})/t$ 和$(t_{cc}+t_{ic}+t_{uc})/t$。在第一次协同训练过程中,分类器 f_2 将 t_{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 预测正确的样本增加 t_{uc}。与 f_1 类似,f_2 预测正确的样本增加 t_{cu}。更新各分类器后,原有 t_{uu}个不确定决策的样本可能出现两分类器中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。通过将分类器的训练集从有标记数据扩充至无标记数据,各分类器的性能得到较大的提升。

3　实验仿真分析

实验选用 6 个 UCI 数据集,其详细信息如表 4 所示。其中数据集 ionosphere 的少量数值属性运用等频方法(三等分)进行离散化,而 lymphography 仅选取样本较多的两类。

表 4　UCI 数据集

数据集名称	属性数	实例数	类别数
tic-tac-toe(TTT)	9	958	2
lymphography(Lymp)	18	148	2
mushroom(MR)	22	8 124	2
breast cancer(Cancer)	30	569	2
ionosphere (Iono)	34	351	2
chess2(Chess)	36	3 196	2

实验采用 10 重交叉验证方法划分训练集和测试集,然后每重交叉验证按标记率将训练集随机划分为有标记和无标记样本集。由于 10 重交叉验证方法受数据集样本次序的影响,实验打乱样本进行了 10 次随机 10 重交叉验证以保证实验结果的有效性。

算法 1 能有效地去除部分标记数据的冗余属性,实现半监督属性约简。在标记率为 10% 时,所选数据集约简前后的属性数目与真实的约简子空间对比如图 2 所示。

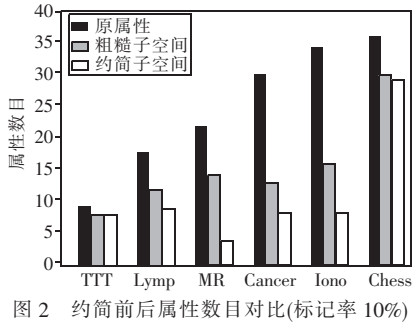

图 2　约简前后属性数目对比(标记率 10%)

从图 2 可见,算法 1 在所选数据集上都去除了冗余属性。在数据集 TTT 上,标记率为 10% 的粗糙子空间和真实约简子空间(标记率为 100%)完全一致,说明了算法 1 的有效性。

为了验证新算法的有效性,实验选用了自训练和随机协同[3]两种部分标记学习算法进行对比分析。实验过程中,各算法将采用 J48 决策树作为分类器。自训练和随机协同算法将选用置信度大于 0.75 的无标记对象作为可信样本。而新算法中正例、边界和负例样本风险代价统一设置为(1,1.5,3)。在标记率为 10% 时,各算法 10 次随机 10 重交叉验证的平均结果如表 5 所示。

表 5 各算法错误率对比(标记率为 10%)

数据集	自训练		随机协同		新算法	
	初始	最终	初始	最终	初始	最终
TTT	0.316 1	0.316 2	0.344 3	0.376 9	0.314 2	0.300 1
Lymp	0.352 4	0.329 7	0.321 8	0.266 0	0.253 1	0.251 4
MR	0.019	0.019	0.042 6	0.024 4	0.012 6	0.004 3
Cancer	0.102 2	0.097 8	0.107 3	0.110 1	0.095 1	0.082 6
Iono	0.209 0	0.209 0	0.251 0	0.266 7	0.213 5	0.201 3
Chess	0.276	0.227 4	0.396 6	0.289 8	0.297 2	0.179 5
平均	0.212 5	0.199 9	0.243 9	0.222 3	0.197 6	0.169 9

表 5 中,“初始”表示算法仅在有标记数据上的学习性能,“最终”表示算法利用无标记数据的学习性能。从表 5 可以看出,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不变和降低的情况,如数据集 TTT、MR 和 Iono。随机协同训练方法是双分类器模型,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,这就违反了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偿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,如 TTT、MR、Cancer、Iono 和 Chess。新算法运用粗糙集理论生成两差异性粗糙子空间,保证了两分类器充分且具有差异性,而决策粗糙集理论能有效地选择可利用的无标记样本,因此两分类器能通过协同训练获得较好的学习性能。

为了进一步比较算法的性能,实验测试了新算法与其他算法性能的统计显著性水平(表 6)。给定算法 A 和 B,各算法在相同参数下独立重复 10 次实验,然后运用双尾成对 t-检验评价两算法性能的差异性程度。具体来说,双尾成对 t-检验的显著性水平 p 值用于度量两算法性能的差异性。p 值越小,说明算法性能差异程度越大。一般来说,如果显著性水平 p 值小于 0.05 时(即 5.0e -2),统计差异性是显著的。

表 6 新算法相对自训练、协同训练算法统计差异性对比

数据集	新算法 vs 自训练		新算法 vs 随机协同	
	win/tie/loss	p 均值	win/tie/loss	p 均值
TTT	9/1/0	1.7e -3	10/0/0	3.4e -5
Lymp	10/0/0	5.3e -7	10/0/0	3.2e -3
MR	5/3/2	4.1e -1	8/2/0	1.7e -2
Cancer	10/0/0	2.6e -3	10/0/0	8.3e -3
Iono	8/2/0	7.4e -2	10/0/0	3.3e -5
Chess	10/0/0	5.2e -3	10/0/0	4.8e -6

表 6 中,“win/tie/loss”显示了新算法与其他比较算法的性能统计差异性比较。“win”表示新算法在 10 次统计显著性测试中明显优于其他算法的次数;“loss”表示新算法在 10 次统计显著性测试中明显劣于其他算法的次数;“tie”则表示新算法与其他算法差异性不明显的次数。“p 均值”表示 10 次统计测试显著性水平的平均值。从表 6 中可明显看出,新算法要优于自训练和随机协同算法。在所有数据集上,新算法在 86.7% 和 96.7% 的比较实验中统计分别优于自训练和随机协同算法,仅有 3.3% 的比较实验弱于自训练算法。

4 结束语

协同训练是一种有效的半监督学习算法,然而现实数据较难满足其充分且独立视图的条件;此外,协同训练算法仅利用置信度信息来选择无标记数据,不考虑无标记样本决策的风险代价。本文通过将属性约简和风险决策概念引入协同训练,提出了可有效利用无标记样本提升分类性能的协同决策算法,解决了协同训练视图分割和无标记样本选择问题。理论分析和实验仿真结果表明,新算法较已有算法更优。下一步将考虑采用异构分类器和多分类器,进一步增强协同训练的效果,并将新算法应用于实际领域。

参考文献:

[1] BLUM A, MITCHELL T. Combining labeled and unlabeled data with co-training[C]//Proc of the 11th Annual Conference on Computational Learning Theory. New York: ACM Press, 1998: 92-100.

[2] 刘君,熊忠阳,王银辉. 蛋白质二级结构的协同训练预测方法[J]. 计算机应用研究,2011,28(5):1688-1691.

[3] NIGAM K, GHANI R. Analyzing the effectiveness and applicability of co-training[C]//Proc of the 9th ACM International Conference on Information and Knowledge Management. New York: ACM Press, 2000: 86-93.

[4] FEGER F, KOPRINSKA I. Co-training using RBF nets and different feature splits[C]//Proc of International Joint Conference on Neural Networks. 2006: 1878-1885.

[5] 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.

[6] SALAHELDIN A, El GAYAR N. New feature splitting criteria for co-training using genetic algorithm optimization[C]//Proc of the 9th International Conference on Multiple Classifier Systems. Berlin: Springer-Verlag, 2010: 22-32.

[7] 王娇,罗四维,曾宪华. 基于随机子空间的半监督协同训练算法[J]. 电子学报,2008,36(12A):60-65.

[8] YASLAN Y, CATALTEPE Z. Co-training with relevant random subspaces[J]. Neurocomputing, 2010, 73(10-12): 1652-1661.

[9] 于重重,商利利,谭励,等. 半监督学习在不平衡样本集分类中的应用研究[J]. 计算机应用研究,2013,30(4):1085-1089.

[10] 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.

[11] LI Ming, ZHOU Zhi-hua. Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples[J]. IEEE Trans on Systems, Man and Cybernetics, Part A: System and Humans, 2007, 37(6): 1088-1098.

[12] 李华雄,周献中,李天瑞,等. 决策粗糙集理论及其研究进展[M]. 北京:科学出版社,2011.

[13] 范会联,仲元昌. 结合互信息的多目标属性约简[J]. 计算机应用研究,2012,29(2):490-492.