计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2014.20131049
51(8):1811-1820,2014

邻域粗糙协同分类模型

张 维[1,2,3] 苗夺谦[1,3] 高 灿[4] 岳晓冬[5]
[1](同济大学电子与信息工程学院 上海 201804)
[2](上海电力学院计算机科学与技术学院 上海 200090)
[3](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)
[4](中联重科股份有限公司 长沙 410013)
[5](上海大学计算机工程与科学学院 上海 200444)
(zhangweismile@163.com)

A Neighborhood Rough Sets-Based Co-Training Model for Classification

Zhang Wei[1,2,3], Miao Duoqian[1,3], Gao Can[4], and Yue Xiaodong[5]
[1](*School of Electronics and Information, Tongji University, Shanghai* 201804)
[2](*School of Computer Science and Technology, Shanghai University of Electric Power, Shanghai* 200090)
[3](*Key Laboratory of Embedded System and Service Computing (Tongji University), Ministry of Education, Shanghai* 201804)
[4](*Zoomlion Heavy Industry Science and Technology Development Co., Ltd., Changsha* 410013)
[5](*School of Computer Engineering and Science, Shanghai University, Shanghai* 200444)

Abstract Pawlak's rough set theory, as a supervised learning model, is only applicable for discrete data. However it is often the case that practical data sets are continuous and involve both few labeled and abundant unlabeled data, which is outside the realm of Pawlak's rough set theory. In this paper, a neighborhood rough sets based co-training model for classification is proposed, which could deal with continuous data and utilize the unlabeled and labeled data to achieve better performance than the classifier learned only from few labeled data. Firstly, a heuristic algorithm based on neighborhood mutual information is put forward to compute the reduct of partially labeled continuous data. Then two diverse reducts are generated. The model employs the two reducts to train two base classifiers on the labeled data, and makes the two base classifiers teach each other on the unlabeled data to boot the their performance iteratively. The experimental results on selected UCI datasets show that the proposed model are more effective to deal with partially labeled continuous data than some representative ones in learning accuracy.

Key words neighborhood rough sets; neighborhood mutual information; semi-supervised reduction; co-training; continuous data

摘 要 Pawlak粗糙集理论是一种有监督学习模型,只适合处理离散型数据.但在一些现实问题中存在着大量的连续型数据,并且有标记数据很有限,更多的是无标记数据.结合邻域粗糙集和协同学习理论,提出了适合处理连续型数据并可有效利用无标记数据提升分类性能的邻域粗糙协同分类模型.该模型首先构建了邻域粗糙半监督约简算法,并利用该算法提取两个差异性较大的约简构造基分类器,然后迭代地在无标记数据上交互协同学习.UCI数据集实验对比分析表明,与其他同类模型相比,该模型有较好的性能.

关键词 邻域粗糙集;邻域互信息;半监督约简;协同学习;连续型数据

中图法分类号 TP181

收稿日期:2013-07-22;**修回日期**:2014-03-28
基金项目:国家自然科学基金项目(61075056,61273304,61202170,61103067);中央高校基本科研业务费专项资金项目

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.近年来,随着应用领域的扩展和数据环境的变化,研究人员针对实际应用问题从多个视角对经典 Pawlak 粗糙集理论进行了丰富与完善.

从数据关系角度,Pawlak 粗糙集的上、下近似概念是定义在等价关系上的,然而在许多实际应用中数据往往无法满足严格的等价关系.因此相关研究用非等价关系替代等价关系提出了多种粗糙集扩展模型,如邻域粗糙集、模糊粗糙集、相容粗糙集、覆盖粗糙集、基于优势关系的粗糙集等.此外,基于概率方法量化数据关系,相关研究又提出了概率型粗糙集模型,如决策粗糙集、0.5-概率粗糙集、可变精度粗糙集、参数化粗糙集、贝叶斯粗糙集等.文献[3]对多种粗糙集扩展模型做了比较全面的总结,这些模型极大丰富了经典粗糙集的理论与应用研究.

从数据类型角度,Pawlak 粗糙集只适于处理离散型数据,对具有连续属性值域的数据不能直接处理.然而,实际应用中普遍存在大量连续型数据,如股票价格、工业排水的化学成份、房屋的价格等.应用经典粗糙集理论处理该类数据时,通常采用离散化算法把连续型数据转化为离散型数据[4-7],但离散化策略不可避免地带来信息损失,并影响最终分类识别精度[8].为此,Pawlak 粗糙集被扩展为模糊粗糙集、相似关系粗糙集和邻域关系粗糙集等模型来处理连续型数据[9].

从数据质量角度,Pawlak 粗糙集在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的实验,代价很大.如果仅在有限的有标记数据上训练分类器,通常难以达到理想的分类性能.而大量无标记数据的获取相对容易[10],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就很有意义.文献[11-14]将粗糙集理论引入部分标记数据分类应用中,文献[15-16]结合半监督学习的思想,提出了可有效处理部分标记数据的粗糙协同半监督学习模型,并用实验验证该模型具有较高的学习性能,但是上述方法无法有效处理连续型数据.

针对上述问题,本文提出了邻域粗糙协同分类模型,可有效处理连续型数据,并能利用无标记数据提升分类学习性能.邻域粗糙协同分类模型将邻域粗糙集与半监督协同学习理论相结合,构建了邻域粗糙半监督约简算法,并利用该算法提取两个差异性较大的约简构造基分类器,然后迭代地在无标记数据上交互协同学习,最终形成高性能分类器.本文首先给出关于邻域粗糙集、邻域互信息与半监督协同学习的相关概念,接着提出邻域粗糙半监督约简算法,然后对不存在自然分割的属性集获取两个较大差异性的属性约简,给出邻域粗糙协同分类模型,最后用实验验证模型的有效性.

1 基本概念

一般地,信息系统可表示为 $IS=(U,A,V,f)$,其中,U 是对象集合 $\{x_1,x_2,\cdots,x_n\}$;A 是属性非空集合 $\{a_1,a_2,\cdots,a_m\}$;$V=\bigcup_{a\in A}V_a$,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 $A=C\cup D$,$C\cap D=\varnothing$,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.

1.1 邻域粗糙集

为了解决 Pawlak 粗糙集不能直接处理连续型数据的问题,文献[17-18]将邻域关系引入到粗糙集,将其扩展为邻域粗糙集.

对于任意对象 $x_i\in U$,$B\subseteq C$,定义 x_i 在属性空间 B 的邻域 $\delta_B(x_i)$ 为:

$$\delta_B(x_i)=\{x_j\mid x_j\in U,\Delta^B(x_i,x_j)\leqslant\delta\},$$

其中,Δ 是 U 上的距离函数,满足 $\Delta(x_i,x_j)\geqslant 0$.在实际应用中,欧氏距离是常用的度量:

$$\Delta(x_i,x_j)=\left(\sum_{k=1}^{n}|x_{ik}-x_{jk}|^2\right)^{1/2}.$$

论域中所有对象的邻域形成了论域的粒化,邻域粒子族构成了论域空间中的基本概念系统,通过这些基本概念定义上下近似,可以逼近空间中的任意概念.N 为论域空间 U 上邻域信息粒子族导出的邻域关系,概念 $X\subseteq U$ 的下、上近似以及边界为:

$$\underline{N}X=\{x_i\mid\delta(x_i)\subseteq X,x_i\in U\},$$

$$\overline{N}X=\{x_i\mid\delta(x_i)\cap X\neq\varnothing,x_i\in U\},$$

$$BNX=\overline{N}X-\underline{N}X.$$

1.2 知识的邻域熵与邻域互信息

在文献[19]里,香农提出的熵的概念被引入到粗糙集里作为知识的度量,定义了知识的熵与互信息.但这些概念只适合离散型数据的度量.为此,基于

邻域关系的定义,文献[20]将邻域互信息引入了信息理论,使熵的概念更加一般化,可以用于连续型数据的度量.下面给出邻域互信息的相关基本概念,公式中对数的底最常用的是以 2 为底,还可以采用 e 等其他的底,并可进行互换.相关理论的详细介绍请参阅文献[20].

设 $S\subseteq C$,x_i 在 S 上的邻域表示为 $\delta_S(x_i)$,其知识 S 的熵定义为:

$$NH_\delta(S)=-\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\delta_S(x_i)\|}{n}.$$

设 $R,S\subseteq C$,x_i 在属性空间 $R\cup S$ 的邻域表示为 $\delta_{R\cup S}(x_i)$,联合邻域熵定义为:

$$NH_\delta(R,S)=-\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\delta_{R\cup S}(x_i)\|}{n}.$$

特别地,如果 D 是类属性,定义 $\delta_{R\cup D}(x_i)=\delta_R(x_i)\cap D_{x_i}$,那么

$$NH_\delta(R,D)=-\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\delta_R(x_i)\cap D_{x_i}\|}{n}.$$

R 与 S 的邻域互信息定义为 $NMI_\delta(R;S)=-\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\delta_R(x_i)\|\cdot\|\delta_S(x_i)\|}{n\|\delta_{R\cup S}(x_i)\|}$.

同样,特别地,如果 D 是类属性,文献[20]已证明,$NMI(R;D)=-\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\delta_R(x_i)\|\cdot\|D_{x_i}\|}{n\|\delta_R(x_i)\cap D_{x_i}\|}$.

邻域熵与邻域互信息均可用于连续型数据的度量,分别度量了信源提供的平均信息量的大小以及一个信源从另一个信源获取的信息量的大小.

1.3　半监督协同学习

Blum 和 Mitchell[21]提出的协同学习(co-training)是一种处理部分标记数据的经典算法.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,用学得的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,以此提高学习性能.

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的视图属性集.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:一类是尽量满足该条件,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[22-27]的工作,但是这些方法不能保证视图的充分性或者是视图分割不稳定.另外一种策略是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[28-31]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[15-16]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.但是该模型只适合处理离散型数据.

2　邻域粗糙半监督约简

现实问题中,部分标记数据往往包括少量有标记数据和大量无标记数据.而 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,现有的粗糙集尚无较好的方法.文献[15-16]对此进行了探讨,根据差别矩阵的约简特性,定义了半监督差别矩阵,提出了基于 Markov 覆盖的启发式半监督属性约简算法,有效地扩展了有监督粗糙集模型.但是该算法的研究对象只针对离散型数据.下面我们将结合邻域粗糙集理论,提出连续型数据的半监督约简算法.

首先将部分标记数据转换为决策表.转换的策略是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.这样既充分考虑了有标记数据的决策信息,又考虑了无标记数据的区分信息,保证约简后决策表的分类能力不变.

为了方便描述,下面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 L 指有标记数据集合,N 是无标记数据集合,决策属性 D' 的值域 $V_{D'}$ 可取空值.

从 PS 转换后的数据表示为 $PS'=(U'',A=C\cup D'',V'',f'')$.决策属性 D'' 的值域 $V_{D''}$ 不包括空值,但和 PS 相比,增加了无标记数据的伪类标记.

将部分标记数据中所有无标记数据标注正确的决策表表示为 $DT=(U,A=C\cup D,V,f)$.后续的工作

基于 PS' 进行. 因此,需要首先分析 PS' 与 DT 的约简的关系以证明 PS' 约简的有效性. 文献[32]给出了决策表约简的信息表示,证明了其与 Pawlak 的代数表示的等价性. 可知互信息相等可以作为寻找决策表约简的条件. 在此基础上,给出以下命题.

命题 1. 假设部分标记数据表示为 $PS=(U'=L\cup N, A=C\cup D', V', f')$,如果 $RED_{PS'}$ 是从 PS 转换后得到的部分标记数据 PS' 的约简,则 $RED_{PS'}$ 至少包含一个 DT 的约简 RED_{DT}. 即对任意 $RED_{PS'}$ 至少存在一个 DT 的约简 RED_{DT},使 $RED_{DT}\subseteq RED_{PS'}$.

证明. RED_{DT} 表示 DT 的约简,其属性集表示为 R. 根据邻域互信息的知识,

$$NMI(R;D)=-\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\delta_R(x_i)\|\cdot\|D_{x_i}\|}{n\|\delta_R(x_i)\cap D_{x_i}\|},$$

$$NMI(C;D)=-\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\delta_C(x_i)\|\cdot\|D_{x_i}\|}{n\|\delta_C(x_i)\cap D_{x_i}\|},$$

因此可以得出:$NMI(R;D)=NMI(C;D)$.

设 $Diff=NMI(C;D)-NMI(R;D)$,可以表示为:

$$Diff=-\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\delta_C(x_i)\|\cdot\|\delta_R(x_i)\cap D_{x_i}\|}{\|\delta_C(x_i)\cap D_{x_i}\|\cdot\|\delta_R(x_i)\|},$$

那么 $\dfrac{\|\delta_C(x_i)\|\cdot\|\delta_R(x_i)\cap D_{x_i}\|}{\|\delta_C(x_i)\cap D_{x_i}\|\cdot\|\delta_R(x_i)\|}=1$.

对于 $\forall x\in U, R\subseteq C, \delta_R(x_i)\supseteq\delta_C(x_i)$ 成立,那么可以得出 $\dfrac{\|\delta_R(x_i)\cap D_{x_i}\|}{\|\delta_C(x_i)\cap D_{x_i}\|}\geqslant 1$.

$RED_{PS'}$ 表示 PS' 的约简,其属性集表示为 R'.

设 $Diff'=NMI(C;D)-NMI(R';D)$,可以表示为:

$$Diff'=-\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\delta_C(x_i)\|\cdot\|\delta_{R'}(x_i)\cap D'_{x_i}\|}{\|\delta_C(x_i)\cap D'_{x_i}\|\cdot\|\delta_{R'}(x_i)\|}.$$

从 PS 转换为 PS' 后,PS' 中有标记数据保持不变,每一个无标记数据被赋予可与其他数据可区分的伪类标记. 因此 $\|D'_{x_i}\|=\|D_{x_i}\|$ 或者 $\|D'_{x_i}\|=1$. 那么 $\dfrac{\|\delta_{R'}(x_i)\cap D'_{x_i}\|}{\|\delta_C(x_i)\cap D'_{x_i}\|}$ 和 $\dfrac{\|\delta_R(x_i)\cap D_{x_i}\|}{\|\delta_C(x_i)\cap D_{x_i}\|}$ 相比保持不变或者是减小了.

因此,要 PS' 约简后保证分类能力不变,即 $Diff'=0$,$\|\delta_{R'}(x_i)\|$ 应当不变或者是减小. 那么必定是 R 不变或者是有新的属性加入 R. 因此,$R'\supseteq R$, $RED_{PS'}\supseteq RED_{DT}$. 证毕.

根据以上命题的结论,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,说明了在 PS' 上求取的约简的有效性.

在正式给出约简算法以前,先给出一些相关概念及定义.

定义 1. 给定部分标记数据 $PS=(U'=L\cup N, A=C\cup D', V', f')$,对于其转换后的部分标记数据 PS' 中的任意属性 $a\in C$,其属性重要度可表示为:$Sig(a,R,D)=NMI_\delta(R\cup\{a\};D)-NMI_\delta(R;D)$,这里 $R\subseteq C, a\in C-R$.

定义 2. 给定部分标记数据 $PS=(U'=L\cup N, A=C\cup D', V', f')$,任意属性子集 $R\subseteq C$ 为转换后的部分标记数据 PS' 的约简当且仅当以下条件成立:

1) $NMI_\delta(R;D)=NMI_\delta(C;D)$;

2) $\forall a\in R, NMI_\delta(R-\{a\};D)\neq NMI_\delta(C;D)$.

基于上述概念,邻域粗糙半监督约简算法描述如下:

算法 1. 邻域粗糙半监督约简算法.

输入: 部分标记数据 $PS=(U'=L\cup N, A=C\cup D', V', f')$、控制邻域大小的域值 δ、计算邻域采用的范数 $norm$;

输出: 半监督约简 R.

1) 将部分标记数据 PS 转换为 PS';

2) 根据 δ 和 $norm$ 计算核属性 $Core$, $R=Core$;

3) 如果 $NMI_\delta(R;D)\neq NMI_\delta(C;D)$,重复以下过程:

① 根据 δ 和 $norm$ 计算 PS' 中属性 $a=\max\limits_{a}\ sig(a,R,D)(a\in C-R)$;

② $R=R\cup\{a\}$;

4) 返回约简 R,算法结束.

讨论该算法的时间复杂度. 假设有 m 个候选属性,我们需要计算 m 次 n 个样本的属性子集与类属性的互信息,从中选取使其互信息最大的那个属性,时间复杂度为 $O(mn)$. 假定在第 k 轮中,已选择了 k 个属性,那么还有 $m-k$ 个候选属性,我们需要计算 $m-k$ 次 n 个样本的属性子集与类属性的互信息,时间复杂度为 $O((m-k)n)$,最坏情形下总的时间复杂度为 $O(\sum\limits_{k=1}^{m}(m-k)n)=O(m^2n)$.

3 邻域粗糙协同分类模型

3.1 基本思想及框架

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,并且只适用于离散型数据. 对于部分标记数据,采用前面的转换策略将其变为决策表后,应用邻域粗糙半监督约简算法能够

保证约简后的属性子集不会损失分类信息.并且一般来讲,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.因此,我们可以由此分割原有的属性集,得到两个差异较大的约简结果,由此获取两个充分冗余的视图进行协同学习.即满足下述条件的两个约简:1)每个约简的属性集都足以描述该问题,也就是说,如果训练样例足够,在每个属性集上都足以学得一个强学习器;2)在给定标记时,每个约简的属性集都条件独立于另一个属性集.协同学习的过程在1.3节已有描述,这里不再赘述.最后对迭代学习后的两个分类器进行合并,形成最终的分类器.模型的框架图如图1所示:

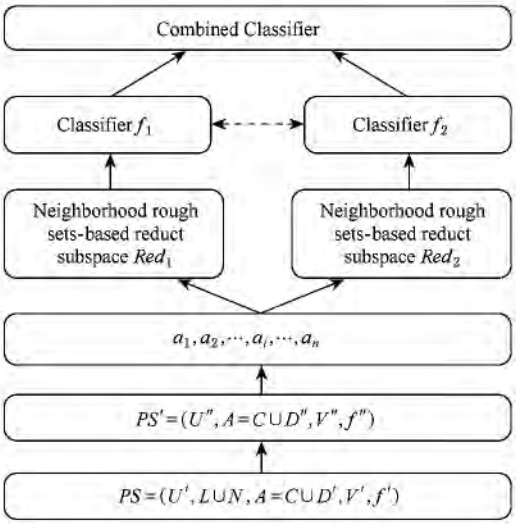

Fig. 1　Framework of the neighborhood rough sets-based co-training model for classification.

图1　邻域粗糙协同分类模型框架

3.2　邻域粗糙差异性约简子空间的协同学习算法

为了满足半监督协同训练对两个视图充分冗余的要求,我们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.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,以此设计差异性属性约简获取算法,具体描述如算法2.

算法2. 邻域粗糙集差异性半监督约简获取算法.

输入:部分标记数据 $PS=(U'=L\cup N, A=C\cup D', V', f')$、控制邻域大小的域值 δ、计算邻域采用的范数 $norm$;

输出:差异性约简 R_1 与 R_2.

1) 将部分标记数据 PS 转换为 PS';

2) 根据 δ 和 $norm$ 计算核属性 $Core$;

3) $R_1=Core$,调用算法1步骤3)获得约简 R_1;

4) $R_2=Core$,将 $C-R_1-Core$ 的属性优先考虑加入 R_2,调用算法1步骤3).如果无法满足 $NMI_\delta(R;D)=NMI_\delta(C;D)$,再将 R_1 的属性作为候选属性,调用算法1步骤3),得到约简 R_2;

5) 返回约简 R_1 与 R_2,算法结束.

该算法获取约简的方法同算法1.但在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1个约简与核属性以外的属性,计算与类属性邻域互信息,如果形成约简则算法结束;如果还不能形成约简,再将第1个约简的属性作为候选属性.算法的复杂度与算法1相当,这里不再赘述.

至此,我们从原始属性集上获取了两个差异性较大的约简,下面给出邻域粗糙协同分类的算法.

算法3. 邻域粗糙协同分类算法.

输入:部分标记数据 $PS=(U'=L\cup N, A=C\cup D', V', f')$、控制邻域大小的域值 δ、计算邻域采用的范数 $norm$;

输出:分类器 f.

1) 运用算法2将部分标记数据 PS 的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;

3) 将 f_1 与 f_2 的无标记样本分别加入集合 N_1 与 N_2;

4) 如果 $N_1\neq\varnothing$ 或 $N_2\neq\varnothing$,重复以下步骤:

① 遍历 N_1,分类器 f_2 挑选出若干标记置信度(即对无标记样本赋予正确标记的置信度)较高的样本进行标记,并把标记后的样本加入 f_1 的训练集 L_1,再训练分类器 f_1,同时更新 $N_1=U-L_1$;

② 遍历 N_2,分类器 f_1 挑选出若干标记置信度较高的样本进行标记,并把标记后的样本加入 f_2 的训练集 L_2,再训练分类器 f_2,同时更新 $N_2=U-L_2$;

5) 返回合成的分类器 f,算法结束.

通过邻域粗糙集差异性半监督约简算法得到了两个约简后,在此基础上利用有标记数据集分别训练分类器.然后在协同训练过程中,每个分类器从未标记示例中挑选出若干标记置信度较高的示例进行标记,并把标记后的示例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,再在更新后的训练集上重新训练分类器,迭代此过程协同学习.邻域粗糙协同分类模型正是以这种方式利用无标记数据提升

分类器的性能.

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4)的迭代学习过程中,两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$O(nt)$.

3.3 模型分析

前已述及,Pawlak粗糙集只适合处理离散数据,对于数值型数据不能直接处理,通常采用离散化的方法转换数据类型,这种方法会带来信息损失.文献[9]比较了采用离散化方法后Pawlak粗糙集属性约简与基于邻域粗糙集属性约简各自得到的特征数量与分类精度.实验分析表明,邻域粗糙集模型可以选择少量的特征并且保持甚至显著提高约简数据的分类精度,证明了该方法提高了约简的质量.和离散化方法相比较,模型中的约简方法提高了约简结果作为视图的充分性.

Blum和Mitchell[21]证明,当数据集的两个视图满足充分冗余的条件时,协同学习算法可以有效地通过利用未标记数据提升分类器的性能.Wang和Zhou[33]进一步证明了只要两个分类器有较大的差异,就可以通过利用无标记数据进行协同学习来提高分类性能.粗糙集约简的性质保证了视图的充分性条件,同时基于邻域粗糙集差异性半监督约简获取算法可以选择两个具有较少共同属性的约简作为两个协同学习的视图,使其具有较大的差异性.因此该模型应能有效地处理部分标记的连续型数据并获得较好的性能.

在有标记数据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_{h1}与n_{h2}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_{h2}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$l+n_{h2}$.类似地,f_2训练集的样本数增至$l+n_{h1}$.再次训练分类器,可能原有无法以较高置信度标记的样本出现能被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任意样本.

4 实验仿真

4.1 实验设置

实验选用了5个UCI标准数据集.为了验证邻域粗糙协同分类模型对连续型数据处理的有效性,所选数据集均为连续型数据,详细信息见表1所示:

Table 1 UCI Data Sets

表1 实验数据集

Data Set	Attribute Type	#Attributes	#Instances
Wine	Continuous	13	178
WDBC	Continuous	30	569
WPBC	Continuous	33	194
Vowel	Continuous	13	989
Waveform	Continuous	40	5 000

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 邻域粗糙协同分类实验结果与分析

为了验证邻域粗糙协同学习算法的有效性,实验选用自训练、随机协同学习两种传统半监督学习算法进行对比分析.自训练是在有标记数据上训练单分类器,然后通过自我标记的方式利用无标记数据提升分类器的性能.随机协同学习是对数据集的属性集进行随机等分形成两个视图.对于随机协同学习与邻域粗糙协同分类器的合成,为了简便,选用两分类器的平均值作为最终性能.

在给定的标记率与邻域δ值情况下,各数据集进行10重交叉验证,最终性能取平均值.表2显示了各算法采用决策树C4.5作为分类器、部分标记数据集标记率α为10%、数据归一化到[0,1]区间、邻域δ为0.1时的分类性能.表2中"NRS Co-training"表示领域粗糙协同学习的性能;NoUnlabeled代表算法在初始标记数据上的性能,随机协同与粗糙协同的计算方法是将两个分类器在初始标记数据上的性能取平均值;WithUnlabeled代表通过无标记数据学习后的性能;"平均值"表示各算法在所选数据集上的平均性能.

从表2中可以看出,各算法在绝大部分数据集上利用无标记数据有效地提升了学习性能.其中,自训练算法的性能提升比较小;除了数据集Vowel,随机协同算法提升的程度更大;邻域粗糙协同在初始

Table 2　Average Accuracy of the Compared Algorithms under the Labeled Rate $\alpha=10\%$ and $\delta=0.1$

表 2　在标记率 $\alpha=10\%$,邻域大小 $\delta=0.1$ 下算法性能对比

Data Set	Self-training		Random Co-training		NRS Co-training	
	NoUnlabeled	WithUnlabeled	NoUnlabeled	WithUnlabeled	NoUnlabeled	WithUnlabeled
Wine	0.760 0	0.769 2	0.751 1	0.763 7	0.767 5	0.846 5
WPBC	0.717 8	0.719 3	0.691 3	0.730 0	0.702 9	0.734 8
WDBC	0.893 6	0.901 9	0.891 7	0.910 9	0.901 0	0.924 9
Vowel	0.570 6	0.572 7	0.492 9	0.447 3	0.563 9	0.692 8
Waveform	0.622 9	0.625 9	0.619 5	0.639 0	0.635 2	0.662 0
Avg.	0.713 0	0.717 8	0.689 3	0.698 2	0.714 1	0.772 2

标记数据上的性能与随机协同相比稍好,与自训练相比差别不大,但是通过无标记数据学习后,数据集的训练性能均比随机协同与自训练得到了更大的提升,取得了更好的效果.

为了进一步比较算法的效率,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

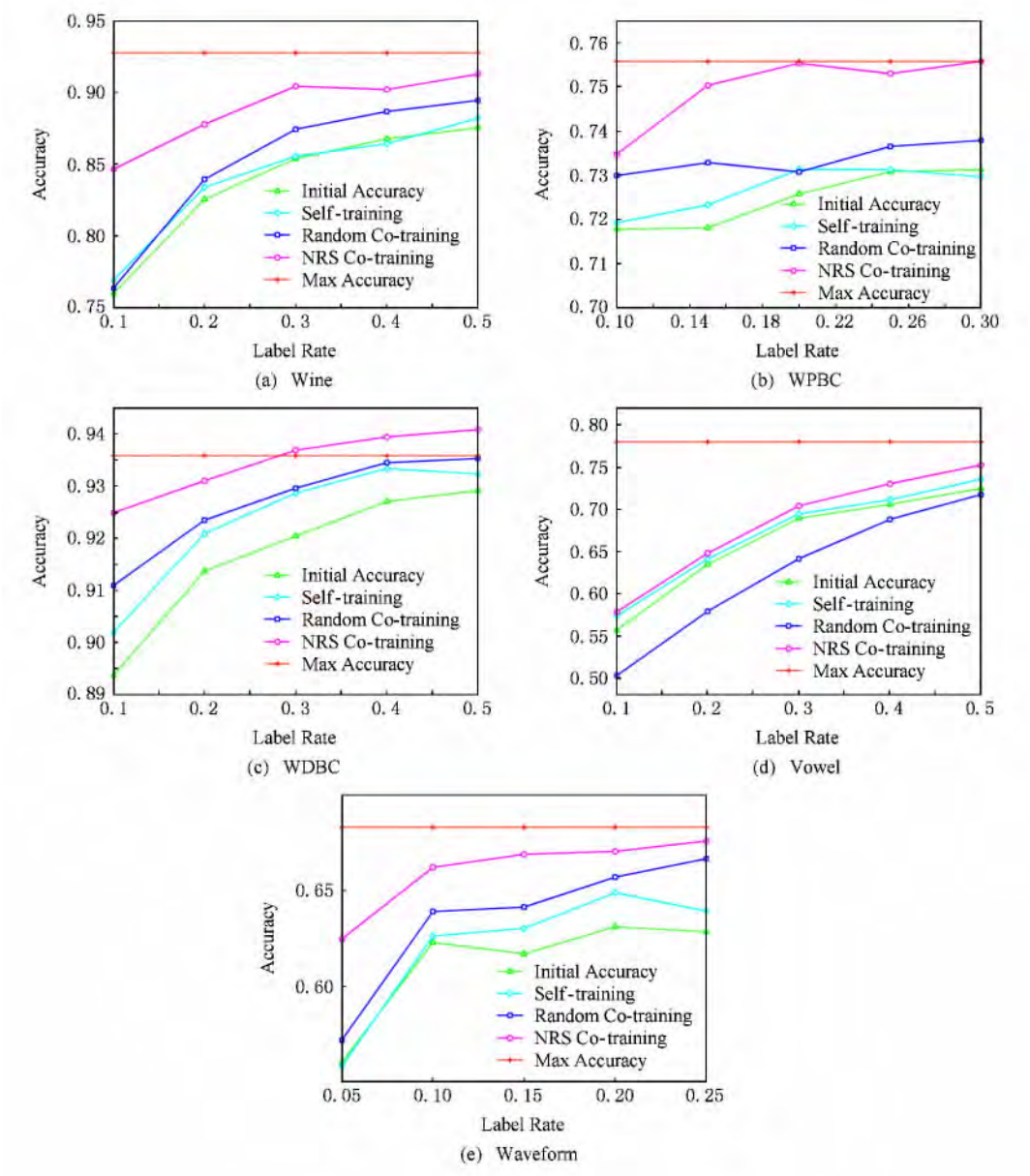

Fig. 2　Average accuracy of the compared algorithms under the different labeled rates.

图 2　不同标记率下算法性能对比

的“Initial Accuracy”代表不利用无标记数据的传统粗糙集方法性能;“NRS Co-training”表示领域粗糙协同学习的性能;“Max Accuracy”表示将所有训练集样本标注正确后传统粗糙集方法的性能.

从图 2 可以看出,邻域粗糙协同分类算法的分类精度在大多数数据集上取得了较大的提升,随机协同与自训练算法的分类精度也有提高,但是程度较小,最终性能也不及邻域粗糙协同分类算法. 自训练采用自我标记的方式,在初始分类效果不好的情况下,误标记的概率就会比较大,这些样本加入训练集后迭代训练,尽管也加入了正确标记的数据扩大了训练集,但可能会导致最终的分类性能提升不大,甚至更差,如图 2 中 Waveform 数据集标记率为 5% 所示. 随机协同与邻域粗糙协同划分了两个属性空间并在此基础上构建了两个分类器. 前者采用随机划分属性生成两个视图的方式,无法满足协同学习对视图充分性的要求,因此初始分类器的性能可能比后者差,如图 2 中的各数据集所示,以至最终学习性能较低,甚至弱于自训练算法在初始标记数据上的分类性能,如数据集 Vowel 的实验结果. 后者生成两个差异性较大的约简作为协同学习的两个视图,保证了视图的充分性与差异性,在协同训练过程,通过其他分类器置信的数据,获取了比自训练更多的信息,扩大自己的训练集,从而改善各个视图上的分类器性能,因此如图 2 中各数据集所示,在初始标记数据上以及通过无标记数据学习均比自训练与随机协同的性能更好.

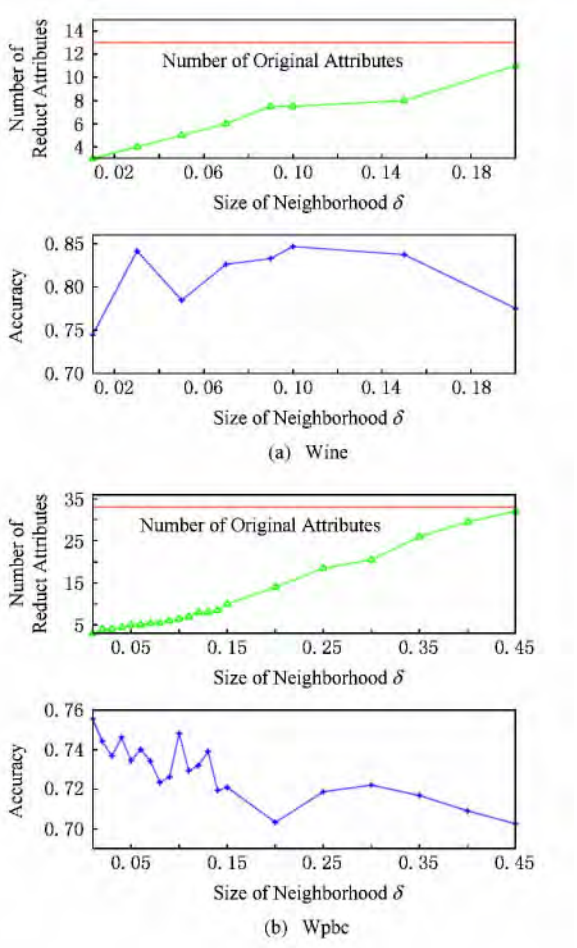

Fig. 3 Number and accuracy of selected features varying with the size of neighborhood δ.

图 3 约简中属性数量与约简分类能力随邻域大小 δ 的变化

从图 2 还可以看出,邻域粗糙协同分类算法在某些标记率下的性能能达到甚至优于传统粗糙集方法的最优值. 例如在标记率 20% 下,邻域粗糙协同分类算法在数据集 WPBC 上达到了传统粗糙集方法的最优性能;在标记率 30% 下,数据集 WDBC 取得了优于传统粗糙集方法的最优值. 这可归结于无标记数据的利用以及两个分类器协同学习.

图 3 展示了约简中属性数量、分类精度随邻域大小 δ 的变化情况. 约简属性个数取两个约简属性个数的平均值. 当 δ 较小时,发现的特征较少,这时分类精度可能较低;当 δ 较大时,发现的特征较多,这时两个约简的重复属性可能就较多,协同学习失去了良好的性质,分类精度也可能较低. 参数 δ 在 0.1 附近取值较为理想.

5 结束语

现实问题存在着大量的连续型数据,并且有标记数据由于标注代价过大则相对较少,而大量的无标记数据却容易获得. 本文把邻域粗糙集理论与半监督协同学习相结合,提出可有效利用无标记数据提升分类性能并适用于连续型数据的邻域粗糙协同分类模型,解决部分标记连续型数据的约简与分类学习问题. 实验仿真结果表明,该模型的应用可提高分类学习性能.

邻域粗糙集模型中 δ 的大小反映了对象之间的相似性. $\delta=0$ 就表示邻域里的对象在所有属性上取值相同,这时邻域粗糙集模型就退化为 Pawlak 粗糙集. 下一步研究该模型用于离散型数据与连续型数据共存的数据集的学习,以使其更适合现实问题的解决.

另外,邻域粗糙半监督约简算法的算法复杂度较高,降低算法复杂度也是后续的重要工作之一.

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 Guoyin, Yao Yiyu, Yu Hong. A survey on rough set theory and applications [J]. Chinese Journal of Computers, 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] Ching J Y, Wong A K C, Chan K C C. Class-dependent discretization for inductive learning from continuous and mixed-mode data [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 1995, 17(7): 641-651

[5] Miao Duoqian. A new method of discretization of continuous attributes in rough sets [J]. Acta Automatica Sinica, 2001, 27(3): 296-302 (in Chinese)
(苗夺谦. Rough Set 理论中连续属性的离散化方法[J]. 自动化学报, 2001, 27(3): 296-302)

[6] Yu Daren, Hu Qinghua, Bao Wen. Combining rough set methodology and fuzzy clustering for knowledge discovery from quantitative data [J]. Proceedings of the CSEE, 2004, 24(6): 205-210 (in Chinese)
(于达仁, 胡清华, 鲍文. 融合粗糙集和模糊聚类的连续数据知识发现[J]. 中国电机工程学报, 2004, 24(6): 205-210)

[7] Xie Hong, Cheng Haozhong, Niu Dongxiao. Discretization of continuous attributes in rough set theory based on information entropy [J]. Chinese Journal of Computers, 2005, 28(9): 1570-1574 (in Chinese)
(谢宏, 程浩忠, 牛东晓. 基于信息熵的粗糙集连续属性离散化算法[J]. 计算机学报, 2005, 28(9): 1570-1574)

[8] Jensen R, Shen Q. Semantics-Preserving dimensionality reduction: rough and fuzzy-rough-based approaches [J]. IEEE Trans on Knowledge and Data Engineering, 2004, 16(12): 1457-1471

[9] Hu Qinghua, Yu Daren, Xie Zongxia. Numerical attribute reduction based on neighborhood granulation and rough approximation [J]. Journal of Software, 2008, 19(3): 640-649 (in Chinese)
(胡清华, 于达仁, 谢宗霞. 基于邻域粒化和粗糙逼近的数值属性约简[J]. 软件学报, 2008, 19(3): 640-649)

[10] Zhu Xiaojin. Semi-Supervised learning survey, TR1530 [R]. Madison: Department of Computer Sciences, University of Wisconsin, 2008

[11] Gu Xueping, Tso S K. Applying rough-set concept to neural-network-based transient-stability classification of power systems [C] //Proc of the 5th Int Conf on Advances in Power System Control, Operation and Management. London: Institution of Engineering and Technology, 2000: 400-404

[12] Duan Qiguo, Miao Duoqian, Jin Kaimin, A rough set approach to classifying web page without negative examples [C] //Proc of the 11th Pacific-Asia Conf on Advances in Knowledge Discovery and Data Mining. Berlin: Springer, 2007: 481-488

[13] Wang Sheng, Wang Xue, Bi Daowei, et al. Collaborative statistical learning with rough feature reduction for visual target classification [C] //Proc of the 5th Int Joint Conf on Neural Networks. Piscataway, NJ: IEEE, 2008: 1151-1156

[14] Lingras P, Chen Min, Miao Duoqian. Semi-supervised rough cost/benefit decisions [J]. Fundamenta Informatica, 2009, 94(2): 233-244

[15] Miao Duoqian, Gao Can, Zhang Nan, et al. Diverse reduct subspaces based co-training for partially labeled data [J]. International Journal of Approximate Reasoning, 2011, 52(8): 1103-1117

[16] Gao Can, Miao Duoqian, Zhang Zhifei, et al. A Semi-Supervised rough set model for classification based on active learning and co-training [J]. Pattern Recognition and Artificial Intelligence, 2012, 25(5): 745-754 (in Chinese)
(高灿, 苗夺谦, 张志飞, 等. 主动协同半监督粗糙集分类模型[J]. 模式识别与人工智能, 2012, 25(5): 745-754)

[17] Lin T Y. Neighborhood systems and relational database [C] //Proc of the 16th ACM Annual Computer Science Conf. New York: ACM, 1988: 725-728

[18] Lin T Y. Neighborhood systems: A qualitative theory for fuzzy and rough sets [J]. Advances in Machine Intelligence and Soft Computing, 1997, 4: 132-155

[19] Miao Duoqian, Wang Jue. On the relationships between information entropy and roughness of knowledge in rough set theory [J]. 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[J]. 模式识别与人工智能, 1998, 11(1): 34-40)

[20] Hu Qinghua, Zhang Lei, Zhang David, et al. Measuring relevance between discrete and continuous features based on neighborhood mutual information [J]. Expert Systems with Applications, 2011, 38(9): 10737-10750

[21] 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

[22] Nigam K, Ghani R. Analyzing the effectiveness and applicability of co-training [C] //Proc of the 9th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2000: 86-93

[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] Wang Jiao, Luo Siwei, Zeng Xianhua. A random subspace method for co training [J]. Acta Electronica Sinica, 2008, 36(12A): 60-65 (in Chinese)

(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)

[25] Tang Huanling, Lin Zhengkui, Lu Mingyu, 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 co-training using genetic algorithm optimization [C] //Proc of the 9th Int Workshop on Multiple classifier systems. Berlin: Springer, 2010: 22-32

[27] Yaslan Y, Cataltepe Z. Co-training with relevant random subspaces [J]. Neurocomputing, 2010, 73 (10/11/12): 1652-1661

[28] Goldman S, Zhou Yan. Enhancing supervised learning with unlabeled data [C] //Proc of the 17th Int Conf on Machine Learning. San Francisco: Margan Kaufmann, 2000: 327-334

[29] Zhou Yan, Goldman S. Democratic co-learning [C] //Proc of the 16th IEEE Int Conf on Tools with Artificial Intelligence. Piscataway, NJ: IEEE, 2004: 594-602

[30] Zhou Zhihua, Li Ming. Tri-training: Exploiting unlabeled data using three classifiers [J]. IEEE Trans on Knowledge and Data Engineering, 2005, 17(11): 1529-1541

[31] Li Ming, Zhou Zhihua. Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples [J]. IEEE Trans on Systems Man and Cybernetics—Part A: Systems and Humans, 2007, 37(6): 1088-1098

[32] Miao Duoqian, Wang Jue. An information representation of concepts and operations in rough set theory [J]. Journal of Software, 1999, 10(2): 113-116 (in Chinese)
(苗夺谦, 王珏. 粗糙集理论中概念与运算的信息表示[J]. 软件学报, 1999, 10(2): 113-116)

[33] Wang Wei, Zhou Zhihua. Analyzing co-training style algorithms [C] //Proc of the 18th Europe Conf on Machine Learning. Berlin: Springer, 2007: 454-465

Zhang Wei, born in 1978. PhD candidate in the Department of Computer Science and Technology at Tongji University. Student member of China Computer Federation. Her current research interests include rough set theory and machine learning.

Miao Duoqian, born in 1964. Professor and PhD supervisor in the Department of Computer Science and Technology at Tongji University. Senior member of China Computer Federation. His current research interests include rough set theory, granular computing, principal curve, artificial intelligence, etc (miaoduoqian@163.com)

Gao Can, born in 1983. Received his PhD degree from Tongji University in 2013. His current research interests include rough set theory and machine learning (2005gaocan@163.com).

Yue Xiaodong, born in 1980. Received his PhD degree from Tongji University in 2010. Member of China Computer Federation. His current research interests include soft computing, data mining and multimedia (yswantfly@gmail.com).