文章编号: 1001-0920(2019)04-0718-09 DOI: 10.13195/j.kzyjc.2017.1342

基于三支决策的主动学习方法

胡 峰†, 张 苗, 于 洪

(1. 重庆邮电大学 计算机科学与技术学院，重庆 400065；
2. 重庆邮电大学 计算智能重庆市重点实验室，重庆 400065)

摘 要: 主动学习是机器学习领域研究的热点之一,旨在解决样本无标签问题. 将三支决策的思想应用到主动学习中,通过引入决策函数,并基于无标签样本的不确定性,将无标签样本划分为3个不同的域: 正域、负域、边界域. 针对不同区域的样本进行相应处理,提出一种基于三支决策理论的主动学习方法(TWD_Active方法). 通过主动学习方法选出最有用的样本交给专家进行标记,扩大训练集,创建更有效的模型. 与传统的被动学习相比,该方法可以选择信息量高、有代表性的样本进行打标,可避免样本的冗余添加. 通过反复迭代的训练学习达到预设的迭代次数或期望的性能指标. 实验结果表明,所提出的算法在F-value、AUC等评价指标上均可取得良好的效果,验证了该算法的有效性.

关键词: 主动学习；机器学习；三支决策；决策函数；无标签样本；不确定性

中图分类号: TP39 **文献标志码:** A

An active learning method based on three-way decision model

HU Feng†, ZHANG Miao, YU Hong

(1. School of Computer Science and Technology，Chongqing University of Posts and Telecommunications，Chongqing 400065，China；2. Key Laboratory of Computational Intelligence，Chongqing University of Posts and Telecommunications，Chongqing 400065，China)

Abstract: Active learning is one of the focuses in the field of machine learning, aiming to solve the unlabeled problem of samples. In this paper, a three-way decision model is applied to active learning. By introducing decision functions, the unlabeled samples are divided into three different parts: positive region, boundary region and negative region based on the uncertainty of unlabeled samples. Different solutions are adopted to process samples for each region. Then, an active learning method based on the three-way decision model, namely TWD_Active, is developed. The most useful samples are selected using the active learning method, and are labeled by experts, so more effective models can be trained by the expanded training set. Compared with traditional passive learning, this method can choose the informational and representative samples to label, avoiding the redundant addition of sample. The models are continuously trained until the expected number of iterations or performance indicators are achieved. Experimental results show that the proposed algorithm has a better performance in measures F-value, AUC and the effectiveness of the algorithm is verified.

Keywords: active learning；machine learning；three-way decision；decision function；unlabeled samples；uncertainty

0 引 言

传统的分类学习算法需要通过给定的有标签数据训练模型. 在具体的实践过程中发现,人们可以很方便地获取到数据,但是这些数据往往都是没有标签的. 特别是随着大数据时代的来临,数据更是影响着人们的方方面面,直接获取的数据往往是冗余的、繁琐的、无标签的. 对数据进行打标往往需要专业领域知识,如果对每个数据一一打标,显然需要消耗大量的资金和精力,这也是不现实的.

主动学习方法是机器学习方法中的一种,其特点是尽可能使用少量的训练样本实现较高的分类性能[1]. 通过主动学习算法可以选出最有用的数据交

收稿日期: 2017-10-11；**修回日期:** 2018-04-13.

基金项目: 国家自然科学基金项目(61533020, 61472056, 61309014, 61751312)；教育部人文社科规划基金项目(15XJA630003)；重点产业共性关键技术创新专项(cstc2017zdcy-zdyfX0001, cstc2017zdcy-zdzx0046)；重庆市基础与前沿项目(cstc2017jcyjAX0408).

责任编委: 阳春华.

作者简介: 胡峰(1978−), 男, 教授, 博士, 从事数据挖掘、Rough集和粒计算等研究；张苗(1993−), 女, 硕士生, 从事数据挖掘、Rough集的研究.

†通讯作者. E-mail: 2925808583@qq.com.

给专家进行标记,避免了数据的冗余添加和不必要添加,同时减少了大批量标记数据所需的人力和物力. 根据度量无标签样本信息价值的标准不同,主动学习中样本选择的评估方法主要分为以下3种:基于不确定性的信息度量方法、基于版本空间缩减的度量方法和基于误差缩减的度量方法.

1)基于不确定性的信息度量方法是适用性最广、研究最为充分的一种方法. 对于二分类,选择分类器对样本x预测为y^*的概率值$p(y^*|x)$最接近0.5的样本进行标记,可以有效提高分类器的精确度和泛化能力[2]. Lewis等[3]将其思想应用到决策树模型,构造了基于决策树的不确定性函数. 对于多分类问题,可以使用Margin策略[4]度量样本的不确定性. 在主动学习中,熵值是度量模型对样本的不确定性程度的重要方法,故基于样本后验概率输出的熵值方法也是计算样本不确定性的代表性方法. Settles[5]认为熵值更适合处理log损失函数的情况,而Margin策略更适合处理减少分类误差的情况. 当选择SVM为分类器时,通常将样本到分类超平面的距离作为不确定的评价标准[6-7]. 进一步的,学者在以SVM作为分类器时,将样本的分布信息考虑在内,提出了结合样本代表性的主动学习算法[8-12].

2)基于版本空间缩减的度量方法能够最大程度缩减版本空间的样例来进行打标. 委员会投票选择算法(QBC)[13]是最大程度缩减版本空间的代表性方法. Abe等[14]提出了Boosting-QBC和Bagging-QBC的委员会构建方法,这两种委员会的构建方法分别采用了两种著名的集成学习方法Boosting和Bagging.

3)基于误差缩减的度量方法建立在统计学基础之上. Cohn等[15]提出了主动学习的统计学模型,即基于模型方差最小化的缩减策略,但这种策略比较适用于方差简单可求的概率模型.

三支决策[16-17]是由Yao在概率粗糙集和决策粗糙集的基础上研究得出的,Yao在文献[18]中对三支决策问题给出了形式化定义. 近年来,众多学者对三支决策进行了不同方向的研究和应用,得到了一定的发展. 例如:Li等[19]提出了基于三支决策的代价敏感人脸识别方法;Liu等[20]基于回归分析和决策粗糙集提出了一种新的分类方法;Liu等[21]将三支决策理论应用在不完备信息系统中;Yu等[22-23]根据三支决策的区间集描述形式提出了基于三支决策聚类的思想,并实现了基于三支决策的重叠聚类;Ma等[24]基于决策粗糙集实现了多类属性的约简;Chen等[25]提出了一种基于邻域三支决策的属性约减方法.

三支决策通过引入两个参数α和β,将整个空间划分为3个域:正域、负域和边界域. 对这三个域分别采用不同的处理方法,为解决复杂、不确定性问题提供了一种有效的策略. 本文将三支决策理论运用到主动学习中,基于样本的不确定性对样本空间进行区域划分,并对于不同域中的样本分别选择不同的解决方案,这种从个体差异出发,有的放矢进行样本选择的方法可更好地选择出信息量高、代表性强的样本,从而更好地提升分类器各项性能. 该方法对主动学习进行了扩展,可有效用于处理目前大量分类属性缺失的情况.

1　相关概念

1.1　主动学习工作过程

主动学习的工作流程如图1所示.

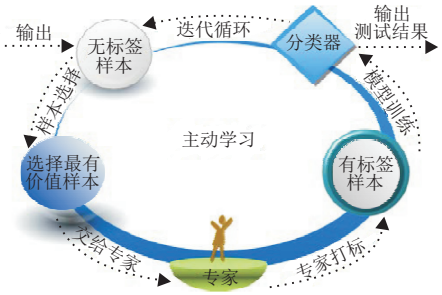

图1　主动学习流程

主动学习的工作一般可分为两个环节:1)样本选择:选择最具打标价值的样本交给人类专家进行打标,之后将打标后的样本添加至训练集;2)模型训练:根据提供的训练集进行有监督学习,并选择合适的评估指标,衡量分类器的分类性能. 在主动学习过程中,以上两个环节交替执行,迭代循环,当达到预设的迭代次数或预设的分类性能时,算法终止.

1.2　邻域粗糙集

Lin[26]提出了邻域模型,通过样本点的邻域将样本空间粒化,定义如下:

定义1 [27]　对于$\forall x_i \in U, B \subseteq C, x_i$在属性子集$B$上的邻域$\delta_B(x_i)$定义为

$$\delta_B(x_i) = \{x_j | x_j \in U,\ \mathrm{dis}(x_i, x_j) \leqslant \delta\}, \tag{1}$$

其中$\mathrm{dis}(x_i, x_j)$为距离的度量函数.

假设x_i, x_j为N维空间样本$A=\{a_1, a_2, \cdots, a_N\}$的两个样本,$f(x, a_i)$为样本$x$在属性$a_i$上的取值,则样本的Minkowsky距离可定义为

$$\mathrm{dis}(x_1, x_2) = \left(\sum_{i=1}^{N} |f(x_1, a_i) - f(x_2, a_i)|^p\right)^{1/p}. \tag{2}$$

当p取2时即为欧拉距离.

对于离散型属性,文献[28]提出了VDM(Value dierence metric). 假设样本x_1、x_2在分类型属性的两个值为v_1、v_2,他们之间的距离定义为

$$f(x_1,v_1)-f(x_2,v_2)=\sum_{i=1}^{N}\left|\frac{C_{1i}}{C_1}-\frac{C_{2i}}{C_2}\right|^t. \quad (3)$$

其中:C_1为所有样本中属性值为v_1的样本个数,C_{1i}为类别为i的样本个数;C_2为所有样本中该属性值为v_2的个数,C_{2i}为其中类别为i的样本个数;t为常数,通常是1.

关于邻域半径δ的计算,文献[27]给出了如下计算公式:

$$\delta=\min(\mathrm{dis}(x_i,s))+\omega\times\mathrm{range}(\mathrm{dis}(x_i,s)),$$
$$0\leqslant\omega\leqslant 1. \quad (4)$$

其中:$\min(\mathrm{dis}(x_i,s))$表示在训练集中样本$s$与距离其最近样本之间的距离,$\mathrm{range}(\mathrm{dis}(x_i,s))$表示在训练集中样本$s$与其他样本之间距离的取值范围.

1.3 三支决策理论

Yao在粗糙集和决策粗糙集理论的基础上提出了三支决策理论,该理论为粗糙集的3个域提供了合理的语义解释.

根据文献[29],为了实现三支决策,首先需要引入实体的评价函数$f(x)$,也称为决策函数,它的值称为决策状态值,其大小反映实体的好坏程度;然后,引入一对阈值α和β,根据决策状态值和阈值将论域中事件对象划分为正域、边界域和负域. 三支决策定义如下.

定义2[29] 给定实数空间上的非空有限样本集合$U=\{X_1,X_2,\cdots,X_n\},\forall x\in U$,给定目标函数$f(x)$,则三支决策定义如下.

1) 如果$f(x)\geqslant\alpha$,则$x\in\mathrm{POS}(X)$;

2) 如果$\beta<f(x)<\alpha$,则$x\in\mathrm{BND}(X)$;

3) 如果$f(x)\leqslant\beta$,则$x\in\mathrm{NEG}(X)$.

2 基于三支决策的主动学习

三支决策是一种符合人类认知能力的决策模式,将三支决策理论应用到主动学习中,可以为主动学习提供一套新的处理问题的思路. 这里,给出基于三支决策的主动学习的主要思路.

2.1 对冗余信息进行删减

计算样本间的相似度,样本间的相似度越高说明样本的特点越一致. 引入多样性准则,对相似度高的样本进行删减以减少冗余样本的添加. 若相似度大于某一设定的阈值,则视为样本信息冗余.

使用相似系数计算样本点之间的相似性,具体计算方式如下.

假设样本x和样本x_k的属性空间分别为$x=\{x^1,x^2,\cdots,x^j,\cdots,x^m\},x_k=\{x_k^1,x_k^2,\cdots,x_k^j,\cdots,x_k^m\}$,则用余弦公式计算两者的相似度,即

$$\mathrm{sim}(x,x_k)=\cos(x,x_k)=\frac{\sum_{j=1}^{m}x^jx_k^j}{\sqrt{\sum_{j=1}^{m}x^jx^j}\sqrt{\sum_{j=1}^{m}x_k^jx_k^j}}. \quad (5)$$

2.2 基于不确定性对无标签样本进行区域划分

2.2.1 计算无标签样本的不确定性

在主动学习中,无标签样本的不确定性是指当前分类器确定其分类的程度. 不确定性越高,说明该样本在分类时越容易分错,对样本的整体分类效果影响越大,对其打标可以提高分类性能. 文中选择基于Margin策略的不确定性度量方法[4]计算无标签样本的不确定性,计算公式如下:

$$D\_\mathrm{value}(x)=p(y_{\mathrm{first}}|x,L)-p(y_{\mathrm{second}}|x,L). \quad (6)$$

其中:L为创建好的分类模型;$p(y_{\mathrm{first}}|x,L)$为最大的后验概率输出值,$p(y_{\mathrm{second}}|x,L)$为次最大的后验概率输出值,两者做差值运算,差值越小(即$D_\mathrm{value}(x)$越小),无标签样本的不确定性越高,其打标的价值越高.

如果分类模型选择非概率模型,则可以对概率输出值进行近似. 例如:如果选择分类器SVM,则可将样本到超平面的距离等价于概率差值;如果选择分类器KNN,则可将近邻样本的类别分布等价于后验概率输出值.

2.2.2 区域划分

由式(5)可知,$D_\mathrm{value}(x)$的取值范围为[0,1]. 下面以二分类为例讨论无标签样本的不确定性.

当$D_\mathrm{value}(x)=0.01$时,有如下方程组成立:

$$\begin{cases}p(y_{\mathrm{first}}|x)+p(y_{\mathrm{second}}|x)=1,\\ p(y_{\mathrm{first}}|x)+p(y_{\mathrm{second}}|x)=0.01.\end{cases}$$

解该方程组可得$p(y_{\mathrm{first}}|x)=0.505,p(y_{\mathrm{second}}|x)=0.495$,最大后验概率值与次最大后验概率值很接近,分类器对这类样本进行预测非常容易出错,如果对其打标则可以提高分类精度.

当$D_\mathrm{value}(x)=0.99$时,有如下方程组成立:

$$\begin{cases}p(y_{\mathrm{first}}|x)+p(y_{\mathrm{second}}|x)=1,\\ p(y_{\mathrm{first}}|x)+p(y_{\mathrm{second}}|x)=0.09.\end{cases}$$

解该方程组可得$p(y_{\mathrm{first}}|x)=0.995,p(y_{\mathrm{second}}|x)=$

0.005, 最大后验概率值远远大于次最大后验概率值, 分类器对这类样本进行预测是相对不容易出错的. 然而, 在主动学习的过程中, 选择已经可以预测出结果的样本打标是无意义的.

综上所述, 在样本选择的过程中, 由于无标签样本的不确定性不同, 样本所具有的打标价值也就不同. 为了更好地选择具有打标价值的样本, 利用三支决策的思想将无标签样本按照不确定性划分为低、中、高3等, 分别对应正域、边界域和负域. 其中: 正域是不确定性低的样本, 即当前分类器对其决策不易产生误分类, 故打标价值低; 负域是不确定性高的样本, 即当前分类器极易产生误分类, 故打标价值高; 边界域意味着推迟决策, 需重新确定是否打标.

将无标签样本 (u_Set) 进行区域划分的过程如下:

1) 对于$\forall x \in u$_Set, 根据式(5)计算x 的不确定性D_value(x).

2) 根据不确定性的大小对 u_Set 进行由大到小排序, 得到样本 x 的排列次序 getRank(x), 即 getRank(x) 的值越小, 样本 x 的不确定性越高, 属于负域的概率越大, getRank(x) 的值越大, 样本 x 的不确定性越低, 属于正域的概率越大.

3) 给定阈值 α 和 β, 可对样本 x 进行如下区域划分:

如果 getRank$(x) \geqslant \alpha$, 则 $x \in \mathrm{POS}(X)$;

如果 $\beta <$ getRank$(x) < \alpha$, 则 $x \in \mathrm{BND}(X)$;

如果 getRank$(x) \leqslant \beta$, 则 $x \in \mathrm{NEG}(X)$.

其中: $0 \leqslant \beta \leqslant$ selectNum, selectNum 为每次迭代期望打标的数量. 本文根据实验经验结果选取 $\beta =$ selectNum$/2, \alpha = |u_\mathrm{Set}| - \beta + 1$.

根据以上过程, 将无标签样本划分为正域、边界域和负域. 在此基础上, 对不同区域中的数据进行相应的处理.

2.3 对不同区域的样本分别进行处理

Step 1: 处理正域的样本. 对于正域的样本, 认为分类器可以确定该类样本的分类, 对该类样本打标意义不大, 不做处理.

Step 2: 处理负域的样本. 负域的样本都是不确定性高的样本, 将负域的样本交由人类专家进行打标, 以提高分类精度.

Step 3: 处理边界域的样本. 对于边界域的样本, 如果某样本周围有较多的无标签样本, 即位于无标签数据集的高密度区域, 则说明该样本具有代表性. 对其打标有利于该样本及周边样本的分类, 从而降低周围样本的不确定性, 提高分类器的性能. 边界域样本的处理包括以下两个环节:

1) 计算无标签样本的代表性.

结合样本邻域, 计算样本的代表性.

i) 计算边界域样本的邻域.

首先, 根据距离的计算公式(2)和(3), 计算 u_Set 中样本两两之间的距离 $\mathrm{dis}(x_1, x_2)$; 然后, 利用式(4)计算邻域半径 δ; 最后, 利用式(1)得到边界域上每个样本 x 在无标签数据集的邻域 $\delta(x)$.

ii) 基于邻域计算边界域样本的代表性.

根据式(5)可知 D_value(x) 的取值范围为 $[0, 1]$, D_value(x) 的值越小, 不确定性越大. 为了避免分母取零的情况, 这里对 D_value(x) 进行加1处理, 将样本的代表性重新定义如下:

$$\mathrm{representative}(x) = \frac{\frac{1}{|\delta(x)|}\sum_{k=1}^{|\delta(x)|}\mathrm{sim}(x, x_k)}{(D\_\mathrm{value}(x)+1)^a}. \quad (7)$$

其中: x_k 为 x 邻域中的样本; $|\delta(x)|$ 为 x 的邻域中样本的个数; $\frac{1}{|\delta(x)|}\sum_{k=1}^{|\delta(x)|}\mathrm{sim}(x, x_k)$ 代表样本 x 基于邻域的密度均值, $\mathrm{sim}(x, x_k)$ 选择通过余弦公式(5)来计算; a 为权重.

2) 选择无标签样本进行打标.

根据式(7)计算样本的代表性. 选取 top$K$ 个代表性强的样本并交由人类专家打标.

综上, 整个处理过程通过选择信息量高、代表性强的样本进行标记, 将这些打标好的样本添加到训练集, 创建新的分类器. 通过反复迭代训练, 不断对无标签样本打标, 直到达到迭代的预设次数或期望的评估标准.

2.4 算法描述

根据以上分析给出样本集划分算法和基于三支决策的主动学习方法, 如算法1和算法2所示. 算法1用于将样本集按照一定比例划分为训练集、无标签数据集和测试集; 算法2在算法1的基础上, 结合三支决策进行主动学习, 当满足算法终止条件时, 算法终止.

算法1 数据集划分算法.

输入: 数据集 $D = \{x_1, x_2, \cdots, x_i, \cdots, x_n\}$, 初始训练集所占的比例 p, 无标签数据集所占的比例 q;

输出: 划分结果: 训练集 tr_Set, 无标签数据集 u_Set, 测试集 te_Set.

Step 1: 初始化.

tr_Set $= \varnothing$, u_Set $= \varnothing$, te_Set $= \varnothing$.

Step 2: 调用随机函数将数据集随机化.

$D = \text{random}(D)$.

Step 3: 对数据集D进行划分.

for each in i index of D

if $i \leqslant np$ then

$\text{tr\_Set} = \text{tr\_Set} \bigcup \{x_i\}$;

else If $i > np$且$i \leqslant n(p+q)$

$u\_\text{Set} = u\_\text{Set} \bigcup \{x_i\}$;

else $\text{te\_Set} = \text{te\_Set} \bigcup \{x_i\}$;

end if

end for

Step 4: return $\text{tr\_Set}, u\_\text{Set}, \text{te\_Set}$.

算法2 基于三支决策的主动学习算法.

输入: 数据集$D = \{x_1, x_2, \cdots, x_i, \cdots, x_n\}$,初始训练集所占的比例$p$,无标签数据集所占比例$q$,冗余删减的阈值$T$,三支决策阈值$\alpha$和$\beta$,邻域半径参数$\omega$,计算代表性公式参数$a$,每次迭代期望标记的样本数selectNum;

输出: 测试样本的类别集合Result.

Step 1: 划分样本集. 调用算法1划分样本集,得到$\text{tr_Set}, u_\text{Set}, \text{te_Set}$.

Step 2: 计算样本的距离矩阵. 使用式(2)和(3)计算u_Set中样本间的两两距离,得到距离矩阵Matrix,使用式(5)计算u_Set中样本间的相似度,如果相似度大于阈值T,则进行删减.

Step 3: 初始化集合. 待标记集合$\text{select} = \varnothing$,正域$\text{POS}(X) = \varnothing$,边界域$\text{BND}(X) = \varnothing$,负域$\text{NEG}(X) = \varnothing$.

Step 4: 划分无标签样本. 使用训练集tr_Set构造分类器,文中选用朴素贝叶斯分类器[30].

for each x_i in u_Set do

$D\_\text{value}(x) = p(y_{\text{first}}|x, L) - p(y_{\text{second}}|x, L)$;

end for

按照无标签样本的不确定性由大到小排序;

for each x_i in u_Set do

if $\text{getRank}(x_i) \geqslant \alpha$ then

$\text{POS}(X) = \text{POS}(X) \bigcup \{x_i\}$;

else if $\beta < \text{getRank}(X) < \alpha$ then

$\text{BND}(X) = \text{BND}(X) \bigcup \{x_i\}$;

else $\text{NEG}(X) = \text{NEG}(X) \bigcup \{x_i\}$;

end if

end for

Step 5: 分区域对样本进行处理.

1) 对负域的样本进行处理.

for each x_i in $\text{NEG}(X)$ do

//将负域的样本添加到待标记集合中

$\text{select} = \text{select} \bigcup \{x_i\}$;

end for

2) 对边界域的样本进行处理.

for each x_i in $\text{BND}(X)$ do

根据Step 2得到的距离矩阵Matrix,利用式(5)计算边界域样本x_i在u_Set中的邻域半径δ;

利用式(1)计算x_i的邻域$\delta_{u_\text{Set}}(x_i)$;

利用式(7)计算样本的代表性$\text{representative}(x_i)$.

end for

按照样本的代表性对样本进行排序,表示为$\text{Rank}(\text{BND}(X))$;

取top(selectNum $-\ \beta$)个代表性强的样本添加至select中.

Step6: 为样本打标,添加至训练集,训练新的分类模型.

if $\text{select} = \varnothing$ then

算法终止; return NULL;

else

for each x_i in select do

对x_i进行打标;

$\text{tr\_Set} = \text{tr\_Set} \bigcup \{x_i\}$;

$u\_\text{Set} = u\_\text{Set} - \{x_i\}$;

end for

训练最终的分类模型,实现对测试集的预测,将预测结果保存到Result中,即$\text{Result} = \{y''\}$;

重复步骤Step 3 ~ Step 6进入下次迭代,直到算法终止.

end if

算法的复杂度分析: 令训练样本数目为n,样本的属性数目为m. 算法1中Step 1的时间复杂度为$O(1)$,Step 2的时间复杂度为$O(n)$,Step 3的时间复杂度为$O(n)$,Step 4的时间复杂度为$O(1)$,故算法1的时间复杂度为$O(n)$,空间复杂度为$O(m \times n)$. 在算法2中,Step 1的时间复杂度为$O(n)$,Step 2的时间复杂度为$O(m \times n^2)$,Step 3的时间复杂度为$O(1)$,Step 4的时间复杂度为$O(n^3)$,Step 5的时间复杂度为$O(n \log n)$,Step 6的时间复杂度为$O(n)$,故算法2(基于三支决策的主动学习算法)的时间复杂度为$\max\{O(n^3), O(m \times n^2)\}$,空间复杂度为$O(m \times n)$.

3 实 验

3.1 实验数据集

为了验证算法的有效性,本次实验共选择11个数据集,包括二分类和多分类,数据集均来自公开的

UCI数据集[31]. 如表1所示.

表 1　实验数据集

数据集ID	数据集名称	特征数	样本数	类别数
1	austra	14	690	2
2	biodeg	41	1 055	2
3	diabetes	8	768	2
4	messidor-features	19	1 151	2
5	pima	8	768	2
6	sonar	59	208	2
7	cmc	9	1 473	3
8	mfeat-morphological	6	2 000	10
9	vehicle	18	846	4
10	yeast	7	1 484	10
11	dermatology	34	366	6

3.2　实验方法

为了对算法进行实验,需将数据集D分为训练集、无标签数据集和测试集,其中训练集占5%,无标签数据集占65%,测试集占30%. 2.1节中对冗余信息进行删减时设置的阈值为99.99%. 式(6)中选择朴素贝叶斯分类器计算无标签样本的后验概率值. 使用三支决策的划分方法对样本空间进行划分,selectNum为每次迭代期望标记的数量,每次选择初始无标签数据集的10%. 边界域上代表性的计算中,邻域半径参数$w = 0.01$,式(7)中参数a设置为常量1. 选择weka[32]平台下的随机森林作为分类器对测试集进行预测,参数都选择默认的参数. 为了测试分类器的性能,分别选择F_value和AUC作为评估指标进行验证. 实验中为了保证评估结果的稳定可靠,在数据划分之前调用随机函数将数据集进行随机化,重复进行10次实验评估,最后取平均值作为最终的评估结果.

3.3　实验结果

本次实验选择5种主动学习方法进行对比分析,除了本文提出的基于三支决策的主动学习方法(TWD_Active)之外,还包括了基于Margin策略的主动学习法(Margin)[4]、基于随机选择的主动学习法(Random)[11]以及基于代表性的主动学习法(Representive)[12]和基于委员会投票选择的主动学习法(QBC)[13]. 主动学习是一个迭代循环过程,每次迭代选择无标签数据进行标记,并将其添加至训练集,直到满足迭代的次数. 这里选择在添加无标签数据集的10%,20%,30%,40%,50%,80%时,对评估结果进行实验对比,如表2和表3所示.

为了直观展示不同主动学习算法在添加无标签样本过程中的性能指标的变化,图2和图3用柱状图的形式展示了在相同增量的情况下,性能指标在不同数据集上的平均值.

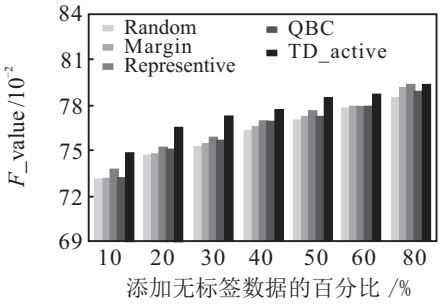

图 2　F_value值在不同数据集上的平均值

由表2、表3可知,本文提出的TWD_Active算法在大多数数据集上都取得了比较好的效果,在添加相同比例的样本时,TWD_Active算法多次优于其他算法. 观察图2和图3中性能指标在不同数据集上的平均值可知,与其他算法相比,TWD_Active算法的F-value和AUC的平均水平都得到较大的提升.

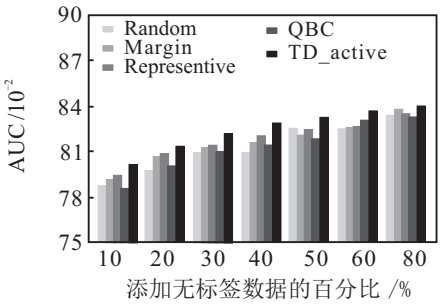

图 3　AUC值在不同数据集上的平均值

分析以上实验过程和实验数据可以发现,随着无标签样本的增加,分类模型的分类性能呈上升趋势,并趋向于某一固定值,但并不能严格保证添加的数据量越大,分类性能越好. 基于Margin的算法可以较好地选择出最不确定的样本,但是可能引入信息冗余的样本;基于Representive的算法将样本的邻域密度考虑在内,即考虑了样本的分布信息,但是在某些情况下会引入密度很高但价值量不高的样本;基于QBC的算法通过构建不同模型,选择分类结果最不一致的样本,但忽略了无标记样本的分布信息;基于Random算法对样本进行了随机选择,带有一定的盲目性,且每次选出的结果都不具有稳定性,但是有时候会取到意想不到的结果. 这些算法在分类性能上都没有取得最好的效果. 本文算法从不确定性角度出发,将问题分为3个部分进行分析讨论,旨在选择出不确定性高、代表性强的样本. 在F_value、AUC等指标上表现出了比较好的结果.

表2 F_value值

数据集ID	算法	10%	20%	30%	40%	50%	60%	80%
1	Random	0.8458	0.8574	0.8509	0.8532	0.8616	0.8663	0.8693
	Margin	0.8581	0.8548	0.8616	0.8665	0.8652	0.8664	0.8653
	Representive	0.8571	0.8546	**0.8687**	0.8686	0.8672	0.8694	0.8726
	QBC	0.8494	**0.8673**	0.8593	0.8685	**0.8740**	0.8634	0.8724
	TWD_Active	**0.8630**	0.8633	0.8607	0.8699	0.8670	**0.8788**	**0.8780**
2	Random	0.6984	**0.7344**	0.7252	0.7420	0.7507	0.7678	0.7760
	Margin	0.6881	0.6763	0.6955	0.7530	0.7564	0.7745	**0.7788**
	Representive	0.7004	0.6732	0.7134	0.7520	0.7605	**0.7763**	0.7643
	QBC	0.6681	0.6946	0.7019	0.7317	0.7366	0.7399	0.7754
	TWD_Active	**0.7099**	0.7259	**0.7299**	**0.7624**	**0.7643**	0.7701	0.7785
3	Random	0.8058	0.8078	0.8093	0.8113	0.8097	0.8143	0.8179
	Margin	0.8031	0.8095	0.8118	0.8154	0.8153	0.8189	0.8171
	Representive	0.8090	0.8064	0.8072	0.8137	**0.8178**	0.8159	0.8161
	QBC	0.8083	0.8078	0.8105	**0.8221**	0.8108	0.8184	0.8229
	TWD_Active	**0.8088**	**0.8192**	**0.8203**	0.8207	0.8107	**0.8235**	**0.8260**
4	Random	**0.6064**	**0.6117**	**0.6197**	0.6296	0.6298	0.6319	0.6403
	Margin	0.5887	0.6045	0.5883	0.6050	0.6165	0.6409	0.6519
	Representive	0.5978	0.5993	0.6086	0.6352	0.6417	0.6462	0.6593
	QBC	0.6081	0.5951	0.6084	0.5898	0.6135	0.6342	0.6540
	TWD_Active	0.6010	0.6029	0.6158	**0.6439**	**0.6520**	**0.6527**	**0.6629**
5	Random	0.7965	0.7893	0.7902	**0.8096**	0.8064	0.8032	0.8028
	Margin	**0.8021**	0.8010	0.7976	0.8015	0.8048	0.7954	0.8025
	Representive	0.7956	0.8034	0.8005	0.7982	0.7943	0.8034	0.8048
	QBC	0.7985	0.8048	0.8041	0.8076	0.8036	0.8007	0.8061
	TWD_Active	0.7983	**0.8052**	**0.8105**	0.7876	**0.8150**	**0.8044**	**0.8103**
6	Random	0.6313	0.6134	0.6452	0.6728	0.6829	**0.7229**	0.723
	Margin	0.6151	0.6652	0.6682	0.6943	0.6993	0.7194	0.7531
	Representive	0.6601	0.6715	0.6968	0.7114	0.7052	0.7195	**0.7592**
	QBC	0.6405	0.6651	0.6862	**0.7213**	0.7088	0.7155	0.7231
	TWD_Active	**0.6865**	**0.6938**	**0.7152**	0.6949	**0.7218**	0.7215	0.7501
7	Random	0.5522	0.5748	0.5775	**0.5882**	0.5971	0.5888	0.5955
	Margin	0.5561	0.5720	0.5774	0.5830	0.5827	0.5787	0.5896
	Representive	0.5509	**0.5885**	0.5776	0.5828	0.5790	0.5798	0.5887
	QBC	**0.5615**	0.5738	0.5817	0.5799	0.5852	0.5809	**0.5985**
	TWD_Active	0.5601	0.5860	0.5870	0.5877	0.5971	0.5980	0.5930
8	Random	0.9899	0.9883	**0.9921**	0.9921	0.9907	0.9898	0.9898
	Margin	0.9889	0.9899	0.9913	0.9904	0.9905	0.9906	**0.9907**
	Representive	0.9898	0.9868	0.9913	0.9923	0.9913	0.9916	0.9913
	QBC	0.9877	0.9887	**0.9921**	0.9921	**0.9921**	**0.9921**	0.9899
	TWD_Active	**0.9919**	**0.9920**	0.9918	**0.9926**	0.9911	0.9921	0.9903
9	Random	0.9751	0.9851	0.9863	0.9882	0.9899	**0.9961**	0.9912
	Margin	0.9921	0.9910	0.9940	0.9938	0.9897	0.9949	0.9950
	Representive	0.9847	**0.9934**	0.9904	0.9918	0.9916	0.9944	0.9921
	QBC	0.9831	0.9825	0.9875	0.9883	0.9874	0.9951	0.9966
	TWD_Active	**0.9926**	0.9907	**0.9953**	**0.9962**	**0.9922**	0.9936	**0.9967**
10	Random	0.5128	0.5239	0.5281	0.5297	0.5413	0.5496	0.5683
	Margin	0.5184	0.5235	0.5355	0.5181	0.5497	0.5403	0.5775
	Representive	0.5338	0.5368	0.5253	0.5223	0.5440	0.5237	**0.5852**
	QBC	0.5215	0.5373	0.5567	0.5524	0.5541	0.5591	0.5633
	TWD_Active	**0.5397**	**0.5572**	**0.5697**	**0.5732**	**0.5733**	**0.5698**	0.5560
11	Random	0.6220	0.7362	0.7428	0.7813	0.8122	0.8333	0.8615
	Margin	0.6380	0.7450	0.7783	0.8018	0.8360	0.8595	0.8852
	Representive	0.6316	0.7599	0.7688	0.8001	0.8504	0.8509	**0.8907**
	QBC	0.6314	0.7405	0.7382	0.8194	0.8357	**0.8655**	0.8787
	TD_Active	**0.6759**	**0.7839**	**0.8067**	**0.8226**	**0.8525**	0.8578	0.8899

表 3 AUC 值

数据集ID	算法	10%	20%	30%	40%	50%	60%	80%
1	Random	0.894 0	0.895 7	0.907 9	0.914 0	0.915 5	0.916 2	0.923 3
	Margin	0.896 0	0.901 4	0.904 6	0.908 8	0.915 7	0.921 1	**0.925 1**
	Representive	0.898 2	0.902 2	0.901 4	0.908 8	0.910 8	0.916 7	0.921 0
	QBC	0.896 7	0.908 4	0.908 1	0.917 3	0.908 8	0.928 6	**0.925 1**
	TWD_Active	**0.906 0**	**0.909 6**	**0.913 2**	**0.922 3**	**0.923 0**	**0.928 8**	0.927 0
2	Random	0.862 8	0.867 3	0.875 6	0.885 6	0.891 9	0.893 2	0.902 9
	Margin	0.866 8	0.872 4	0.879 8	0.895 6	0.892 7	0.898 2	0.910 2
	Representive	**0.868 7**	0.869 0	0.881 6	0.894 8	0.900 8	0.901 1	**0.914 1**
	QBC	0.864 4	0.878 8	0.869 9	0.888 4	0.899 8	0.901 7	0.905 5
	TWD_Active	0.8664	**0.879 3**	**0.888 9**	**0.906 2**	**0.903 4**	**0.907 2**	0.909 7
3	Random	0.766 3	0.775 9	0.772 0	0.780 3	0.781 7	0.781 1	0.789 5
	Margin	0.763 8	0.783 8	0.779	0.780 9	0.784 0	0.783 6	0.798 5
	Representive	0.761 4	0.783 2	0.768 5	0.780 3	0.782 1	0.795 2	**0.799 9**
	QBC	0.779 8	0.777 6	0.765 2	0.783 2	0.786 1	**0.799 6**	0.794 7
	TWD_Active	**0.776 8**	**0.784 0**	**0.780 5**	**0.784 4**	**0.788 3**	0.786 1	0.788 5
4	Random	0.661 1	0.667 6	0.675 7	0.702 0	0.703 5	0.695 1	0.716 4
	Margin	**0.679 2**	0.665 7	0.684 8	0.682 5	0.693 0	0.708 7	0.709 3
	Representive	0.673 9	0.678 4	0.680 6	0.699 9	0.702 8	0.702 4	0.716 3
	QBC	0.664 0	0.669 3	**0.687 2**	0.684 3	0.689 9	0.692 9	0.702 5
	TWD_Active	0.669 2	**0.692 4**	0.6783	**0.703 5**	**0.716 2**	**0.716 3**	**0.719 8**
5	Random	0.750 2	0.756 5	0.755 6	0.768 3	0.763 3	0.775 6	0.782 3
	Margin	0.749 5	0.759 2	0.768 1	0.768 4	0.755 3	0.770 2	0.786 0
	Representive	0.757 0	**0.766 9**	0.770 1	0.762 5	0.773 1	0.769 5	0.779 8
	QBC	0.754 1	0.759 1	0.7637	**0.770 6**	0.773 2	0.773 4	0.773 5
	TWD_Active	**0.757 1**	0.7583	**0.780 7**	0.7679	**0.785 0**	**0.783 2**	**0.790 8**
6	Random	0.722 0	0.754 1	0.788 6	0.810 2	**0.828 7**	0.828 8	0.850 2
	Margin	0.741 5	0.787 1	0.775 7	0.773 5	0.807 3	0.818 4	0.856 8
	Representive	0.756 7	0.789 6	0.797 4	0.806 2	0.816 7	0.806 4	0.823 9
	QBC	0.710 9	0.752 9	0.800 5	0.799 1	0.800 0	0.846 8	0.845 9
	TWD_Active	**0.770 6**	**0.790 7**	**0.819 9**	**0.826 2**	0.811 4	**0.850 2**	**0.864 0**
7	Random	0.633 8	0.656	**0.675 4**	0.668 2	0.670 4	0.672 3	0.675 6
	Margin	**0.667 3**	**0.665 9**	0.672 9	0.673 3	0.668 4	0.662 7	0.666 1
	Representive	0.648 8	0.661 6	0.667 8	0.681 3	0.666 3	0.664 3	0.669 8
	QBC	0.638 9	0.639 4	0.657 8	0.651 2	0.662 0	0.653 4	0.669 7
	TWD_Active	0.661 3	0.663 3	0.673 5	**0.683 9**	**0.678 7**	**0.678**	**0.675 9**
8	Random	0.976 3	0.977 1	0.975 4	0.976	0.976 3	0.975 2	0.974 2
	Margin	0.980 0	0.980 2	0.977 9	0.973 8	0.977 7	0.974 8	**0.976 1**
	Representive	0.980 7	0.980 6	0.977 7	0.973 7	0.977 7	0.974 6	0.975 3
	QBC	**0.981 1**	0.981 1	0.979 9	0.978 7	0.979 0	0.975 7	0.975 7
	TWD_Active	0.978 9	**0.985**	**0.980 8**	**0.979 7**	**0.983 6**	**0.982 9**	0.974 5
9	Random	0.766 7	0.773 8	0.790 2	0.801	0.812 3	0.814 2	0.815 7
	Margin	0.753 8	0.772 2	0.787 1	0.793 3	0.807 2	0.807 1	0.822 4
	Representive	0.751 0	0.779 1	**0.800 3**	0.791 4	0.801 0	0.810 6	**0.830 0**
	QBC	0.757 4	0.788 0	0.791 6	0.790 7	0.796 2	0.812 4	0.811 9
	TWD_Active	**0.768 8**	**0.788 2**	0.7974	**0.805 4**	**0.817 9**	**0.825 2**	0.825 7
10	Random	0.691 8	0.706	0.738 4	0.745 3	**0.760 6**	0.760 1	0.767 7
	Margin	0.714 6	0.734 4	0.743 0	0.741 0	0.745 2	0.751 7	0.767 6
	Representive	0.715 5	0.7378	**0.753 7**	0.753 6	0.755 0	0.763 5	0.769 8
	QBC	0.697 1	0.706 4	0.722 4	0.730 1	0.730 3	**0.768 2**	0.774 3
	TWD_Active	**0.730 8**	**0.747 7**	0.753 3	**0.758 6**	0.759 2	0.760 3	0.774 2
11	Random	**0.930 3**	0.937 5	0.940 9	0.959 4	0.965 6	0.960 1	0.971 8
	Margin	0.891 9	0.946 6	0.958 3	**0.978 8**	0.983 2	0.978 9	**0.990 9**
	Representive	0.922 3	0.945 0	0.955 1	0.969 1	0.974 0	0.982 5	0.977 9
	QBC	0.891 8	0.934 1	0.957 3	0.957 7	0.968 0	0.979 0	0.977 1
	TWD_Active	0.923 9	**0.947 4**	**0.970 5**	0.974 3	**0.983 4**	**0.984 9**	0.987 6

4 结 论

本文提出了一种基于三支决策的主动学习方法,用于解决样本无标签问题. 首先,基于Margin策略计算无标签样本的不确定性;其次,根据样本不确定性将无标签样本空间划分成3个不同的区域;最后,对不同区域上的样本进行相应处理,选择信息量大、价值高的样本进行标记. 不仅降低了对正域样本进行处理的复杂性,而且兼顾了无标签样本的不确定性和代表性. 在UCI数据集上的实验结果表明,相对于本文提到的其他主动学习算法,本文提出的算法在处理无标签样本的情况时,分类器的评估指标都得到了提升. 但是,主动学习本身就是一个迭代循环的过程,尤其当数据的维度很高或数据量很大时,会增加计算的复杂性,因此寻找有效的基于属性约简的主动学习方法或从算法本身出发实现算法的并行化将是下一步的研究重点.

参考文献(References)

[1] Gong X J, Sun J P, Shi Z Z. Active bayesian network classifier[J]. Computer Research and Development, 2002, 39(5): 574-579.

[2] Culotta A, McCallum A. Reducing labeling effort for structured prediction tasks[C]. Proc of AAAI 2005. Menlo Park: AAAI Press, 2005: 746-751.

[3] Lewis D, Catlett J. Heterogeneous uncertainty sampling for supervised learning[C]. Proc of ICML 1994. San Francisco: Morgan Kaufmann, 1994: 148-156.

[4] Scheffer T, Decomain C, Wrobel S. Active hidden Markov models for information extraction[C]. Int Conf on Advances in Intelligent Data Analysis. Berlin: Springer-Verlag, 2001: 309-318.

[5] Settles B. Active learning literature survey[J]. University of Wisconsinmadison, 2009, 39(2): 127-131.

[6] Tong S, Koller D. Support vector machine active learning with applications to text classification[J]. J of Machine Learning Research, 2001, 2(1): 45-66.

[7] Zhang C, Chen T. An active learni ng framework for content based information retrieval[J]. IEEE Trans on Multimedia, 2002, 4(2): 260-268.

[8] Xu Z, Yu K, Tresp V, et al. Representative sampling for text classification using support vector machines[C]. European Conf on IR Research. Berlin: Springer-Verlag, 2003: 393-407.

[9] Donmez P, Carbonell J G, Bennett P N. Dual strategy active learning[C]. Proc of the 18th European Conf on Machine Learning. Berlin: Springer-Verlag, 2008: 208-215.

[10] Hoi S C H, Jin R, Zhu J, et al. Semi-supervised svm batch mode active learning for image retrieval[C]. Proc of the IEEE Computer Society Conference on Computer Vision and Pattern Recongnition. Anchorage: IEEE Computer Society Press, 2008: 1-7.

[11] Huang S, Jin R, Zhou Z. Active learning by querying information and representative examples[C]. Proc of NIPS 2010. Cambridge: MIT Press, 2010: 892-900.

[12] Settles B, Craven M. An analysis of active learning strategies for sequence labeling tasks[C]. Proc EMNLP. Stroudsburg: ACL Press, 2008: 1069-1078.

[13] Seung H S, Opper M, Sompolinsky H. Query by committee[C]. Proc of the 5th ACM Workshop on Computational Learning Theory. New York: ACM Press, 1992: 287-294.

[14] Abe N, Mamitsuka H. Query learning strategies using boosting and bagging[C]. Proc of ICML 1998. San Francisco: Morgan Kaufmann, 1998: 1-9.

[15] Cohn D, Ghahramani Z, Jordan M I. Active learning with statistical models[J]. Artificial Intelligence Research, 1996, 4(1): 129-145.

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

[17] Yao Y Y. The superiority of three-way decision in probabilistic rough set models[J]. Information Sciences, 2011, 181(6): 1080-1096.

[18] Yao Y Y. An outline of a theory of three-way decisions[C]. Int Conf on Rough Sets and Current Trends in Computing. Heidelberg: Springer, 2012: 1-17.

[19] Li H, Zhang L, Huang B, et al. Sequential three-way decision and granulation for cost-sensitive face recognition[J]. Knowledge-Based Systems, 2016, 91(1): 241-251.

[20] Liu D, Li T R, Liang D C. Incorporating logistic regression to decision-theoretic rough sets for classifications[J]. Int J of Approximate Reasoning, 2014, 55(1): 197-210.

[21] Liu D, Liang D, Wang C, et al. A novel three-way decision model based on incomplete information system[J]. Knowledge-Based Systems, 2016, 91(1): 32-45.

[22] Yu H, Wang Y, Jiao P. Detecting and refining overlapping regions in complex networks with three-way decisions[J]. Information Sciences, 2016, 373: 21-41.

[23] Yu H, Zhang C, Wang G Y. A tree-based incremental overlapping clustering method using the three-way decision theory[J]. Knowledge-Based Systems, 2016, 91(1): 189-203.

[24] Ma X A, Wang G Y, Yu H, et al. Decision region distribution preservation reduction in decision-theoretic rough set model[J]. Information Sciences, 2014(278): 614-640.

[25] Chen Y, Zeng Z, Zhu Q, et al. Three-way decision reduction in neighborhood systems[J]. Applied Soft Computing, 2016, 38(1): 942-954.

[26] Lin T Y. Neighborhood systems and approximation in relational databases and knowledge bases[C]. Proc of the 4th Int Symposium on Metho-dologies of Intelligent Systems. Charlotte: Oak Ridge National Laboratory, 1989: 75-86.

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

[28] Stanfill C, Waltz D. Toward memory-based reasoning[J]. Communications of the ACM, 1986, 29(12): 1213-1228.

[29] Liu D, Li T R, Miao D Q, et al. Three-way decision and granular computing[M]. Beijing: Science Press, 2013: 12-30.

[30] Zhou Z H. Machine learning[M]. Beijing: Tsinghua University Press, 2016: 150-154.

[31] Asuncion A, Newman D J. UCI machine learning repository[CP/OL]. http://archive.ics.uci.edu/ml.

[32] Wikipedia Weka(machine learning)[CP/OL]. http://en.wikipedia.org/wiki/Weka.

(责任编辑:闫 妍)