文章编号：1007-791X (2009) 04-0341-06

基于样本不确定性和代表性相结合的可控主动学习算法研究

胡正平[1,*]，高文涛[1]，万春艳[2]

(1. 燕山大学 信息科学与工程学院，河北 秦皇岛 066004; 2. 齐齐哈尔市第三中学，黑龙江 齐齐哈尔 161000)

摘　要：通过选取最有信息量的样本提交专家进行标注，主动学习算法可以有效地减少无效标注样本的工作量。在充分考虑位于分类边界的不确定样本和基于先验分布的具有代表性样本的基础上，本文构造了不确定性与代表性相结合的可控主动学习算法。首先利用样本的kNN 分布状况建立不确定性置信度模型，该思路不需要知道样本分布的具体类型和参数计算；然后在样本聚集度模型的基础上进行聚类，在此基础上建立代表性置信度模型。最后将不确定性置信度模型与代表性置信度模型进行综合，构造可控的主动学习策略，使得每次主动学习选择的样本更具有“价值”。在 UCI 机器学习数据库上的仿真实验结果表明本文的思路是合理可行的，在实验所用数据集上，当达到相同的目标正确率时，本文的方法比随机采样算法所需的样本数量少得多。

关键词：可控主动学习；不确定性样本；样本先验分布；代表性样本

中图分类号：TP181　　**文献标识码**：A

0　引言

传统监督学习问题中，学习算法以外界给定的已标注样本集作为训练集进行训练得到分类器，而非监督学习算法则是使用未标注样本集进行聚类进而研究样本集中蕴含的规律。在很多现实应用中，常常遇到的是半监督学习问题[1]，即拥有较小量的已标注样本集和大量的未标注样本集。在半监督学习问题中，对样本进行标注的代价较大并且难度较高，而获取未被标注样本则相对容易。关于半监督学习问题关键有两点：一是通过合理利用未标注样本集的信息来提高监督学习算法的性能；二是利用一个好的样本选择策略对众多未标注样本进行选择标注，使之能够加入到训练集中进行训练。为解决半监督学习问题，主动学习得到众多的国内外学者的广泛关注。

主动学习算法可以从未标注样本集中选择最有价值的样本交由专家进行标注，从而在不损失训练精度的情况下减少标注样本的代价[2-4]。以两类分类问题为例：假设已有部分已标注样本集$D=\{x_1, x_2, \cdots, x_n\}\subset R^d$，其中$D_l$为已标注样本集，且$D_l$中每个样本的标签为$y\in\{1, -1\}$中的一个，未标注样本集为$D_u=D\setminus D_l$，系统以已标注样本集$D_l$为训练集训练出初始分类器，然后在未标注样本集$D_u$中根据某种原则选择部分样本进行标注并将其加入到训练集中，从而训练出新的分类器，整个过程循环多次，直到分类器的某种评价指标达到预设值或循环次数达到预设值为止。总体上主动学习过程包括两部分：学习引擎和选择引擎[5]。学习引擎的作用是使用一个基准监督学习算法在已标注样本集D_L上训练出分类器，而选择引擎则是负责在未标注样本集D_u中选择要标注的样本，提交专家标注后再将样本交给学习引擎进行学习。

采用什么原则选取最有“价值”的样本进行标注是主动学习的关键，因为“价值”的评判标准不同导致算法性能不同。例如，Seung 等人提出投票选择方法，从版本空间（version space）中随机选择若干假设构成一个委员会，然后选择委员会中的假设预测分歧最大的样本进行标注[6]。该方法所需的样本复杂度较低且能迅速缩减版本空间，但这种

收稿日期：2009-03-07　　**基金项目**：河北省自然科学基金资助项目（F200800891），中国博士后科学基金资助项目（20080440124）

作者简介：*胡正平（1970- ），男，四川仪陇人，博士，副教授，主要研究方向为统计学习理论与模式识别，Email：tnpochw@263.net。

算法的性能对初始值敏感。文献 [7-8] 提出选取能够最小化未来分类的期望错误率的样本，这种方法直接针对分类器评价的最终标准，理论上具有更好的效果，但它的计算量较大，计算复杂度较高。Koller 等人选取当前分类器最不能确定其分类的样本进行标注，这种方法被称为不确定采样（Uncertainty Based Sampling，UBS）[9-10]，此思路在许多应用中都获得不错的效果，不足之处在于算法不够稳定。Xu 等人 [11] 先利用k均值算法对位于 SVM 分类间隔内的样本聚类，进而选择位于类中心具有代表性的样本进行标注，算法稳定性好，收敛速度比较慢。在上述方法中，采用不确定采样方法，但这一方法忽略了对样本先验分布信息的利用。样本的先验分布在一定程度上能够反映样本的代表性，因此需要具有代表性的样本来训练分类模型。另外，考虑到聚类以后位于同一类的不同样本实际中可能具有相同的标签，因此尽量选择具有代表性的样本，这可以避免在同一类中重复选择不必要的样本去标记，减小了标记代价。在充分考虑位于分类边界的不确定样本和基于先验分布的具有代表性样本的基础上，本文构造了不确定性与代表性相结合的可控主动学习算法，这一算法平衡了所选样本的不确定性和代表性，使得所选样本更具有“价值”。最后的实验结果表明：在不影响分类精度的情况下，主动学习选择标记的样本数量大大低于随机选择标记的样本数量，这降低了人工标记的工作量。

1 基于样本不确定性和代表性相结合的可控主动学习算法

1.1 系统组成

主动学习包括两个方面：学习引擎和选择引擎。学习引擎使用 SVM 作为基准学习算法得到分类器，对训练集进行训练，对测试集进行测试。选择引擎综合考虑未标注样本集中样本的不确定性和代表性两个因素，选择最有“价值”的样本进行专家标注，并将最终选择的样本加入到训练集中进行学习以更新分类模型，本文提出的可控主动学习方法系统组成如图 1 所示。

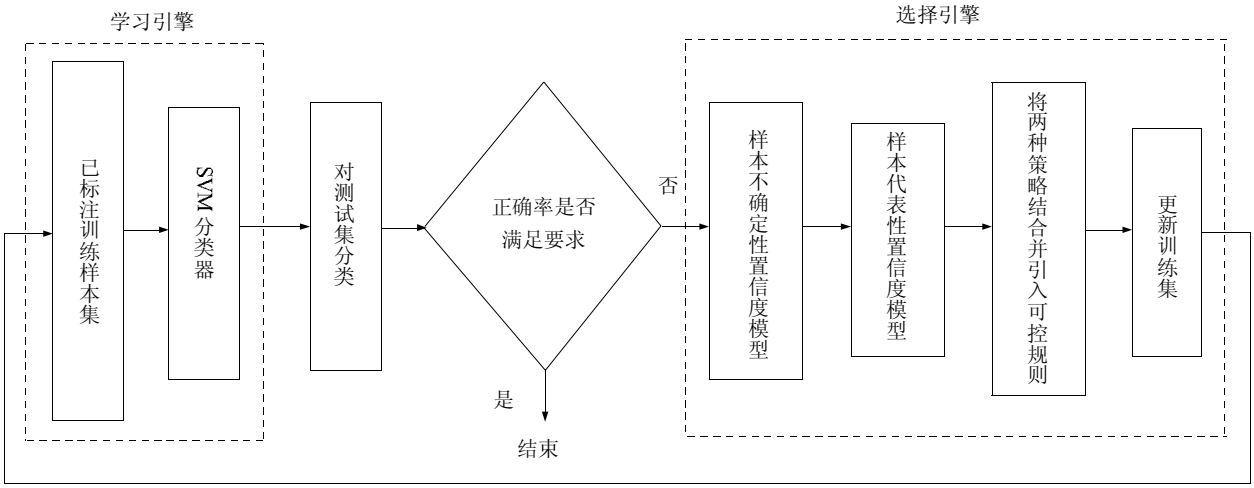


图 1 基于样本不确定性和代表性的可控主动学习系统组成原理框图

Fig. 1 Block scheme of the proposed controlling active learning algorithm

1.2 SVM 学习算法

SVM 学习算法基于结构风险最小化原则，将原始数据集合压缩到支持向量集合，然后用子集学习得到新知识，同时给出由支持向量决定的分类规则。针对两类学习问题，SVM 在原空间或投影后的高维空间中寻找最优分类超平面将两类示例完全分开。当样本集线性可分时，分类超平面为$w\cdot x+b=0$，其中w为权系数，b是分类阈值。最优分类超平面可通过解凸二次优化问题获得：$\min\phi(w)=\|w\|^2/2$，约束条件为$y_i(w\cdot x_i)+b-1\geq 0,\ i=1, 2, \cdots, n$，通过求解可得最优分类超平面的分类判别函数为

$$f(x)=\mathrm{sgn}\left[\sum_{i=1}^{n}\alpha_i y_i(x_i\cdot x)+b\right] \tag{1}$$

其中α_i为拉氏乘子，拉氏乘子不为 0 的解向量为支持向量。对于非线性可分情况，SVM 引入核函数$k(x_i, x_j)=\varphi(x_i)\cdot\varphi(x_j)$，通过简单的内积计算实现非线

性向线性的转化，非线性可分情况的分类判别函数为

$$f(x)=\mathrm{sgn}\left[\sum_{i=1}^{n}\alpha_i y_i K(x_i\cdot x)+b\right] \tag{2}$$

不难看出，SVM 分类器仅与支持向量有关，与其它向量无关。

1.3　样本不确定性置信度模型

分类器对最不确定的样本进行分类时容易出错，分类结果的置信度（confidence）不高。因此，专家在选择对哪些样本进行标注时，不确定性是一个重要考虑因素。利用样本的kNN 分布状况建立样本不确定性置信度模型，采用距离计算的方法根据已标注样本集对未标注样本集中的样本进行选择标注，依据此模型确定样本的不确定性因素。这里采用直推信度机（Transductive Confidence Machines TCM）[12] 构造一种检测函数进行估算，该检测函数的值称为P值。

在计算待标注样本的P值之前，先给出奇异值（strangeness）的定义。它由结合kNN 方法计算样本特征向量在特征空间上的欧式距离获得。一般说来，同类别的样本由于具有相似性，它们的特征向量在特征空间上的分布具有聚集性，样本之间距离比较小；不同类别的样本由于具有相异性，它们的特征向量在特征空间上的分布具有分散性，样本之间距离比较大。因此，待标注样本x_i相对于类别y的奇异值α_{iy}定义为

$$\alpha_{iy}=\frac{\sum_{j=1}^{k}D_{ij}^{y}}{\sum_{j=1}^{k}D_{ij}^{-y}} \tag{3}$$

其中，D_{ij}^{y}表示待标注样本x_i与类别y中所有样本的距离序列中第j个最短的距离，而D_{ij}^{-y}代表待标注样本x_i与其它类别中（除类别y外）所有样本的距离序列中第j个最短的距离，参数k为要考虑的最近邻数目，这里取为 5。奇异值实际上是待标注样本x_i与待加入类中所有样本最小的k个距离之和，与其他类别中样本最小的k个距离之和的比率。若相对于类别y的奇异值越小，则待标注样本x_i属于类别y的可能性就越大。

有了奇异值，那么待标注样本x_i相对于类别y的P值如下定义

$$P(\alpha_i)=\frac{\#\{j:\ \alpha_j\geq\alpha_i\}}{n+1} \tag{4}$$

其中，#表示集合的“势”，这里为满足条件的元素个数；α_i为x_i的奇异值；n为y中样本个数；α_j表示y中任意样本的奇异值；j为y中奇异值大于x_i奇异值的样本个数。本质上P值就是待标注样本属于已存在的几类样本空间的概率，其相对于哪一类的概率值越大，则它属于该类的可能性越大。在未标注样本集中每个样本对应于每一类都有一个P值，因此一个未标注样本有一系列P值，将这些P值进行降序排列，若前两个最大的P值之差很小，则表明不确定该样本属于哪一类，即预测置信度较低，该样本“价值”较大。下式给出利用样本的kNN 分布状况建立的样本不确定性置信度模型：

$$C(x_i)=|P_j-P_k| \tag{5}$$

式中，P_j和P_k代表任意一个样本的前两个最大的P值，当$C(x_i)$值小于一个非常小的接近零的实数时，则表明对应的样本具有较大不确定性，在人工标注时需充分考虑。

1.4　样本代表性置信度模型

样本的代表性可以避免在同一类中重复选择不必要的样本去标记，使得所选样本更具有“价值”。这里利用样本先验分布建立样本代表性置信度模型，因为样本本身蕴藏着许多重要信息，而先验分布信息较容易获得并且本质上与样本的代表性相符。

首先在样本聚集度模型的基础上聚类，聚类的目的是为了找到具有代表性的样本，而不是对样本进行最终分类。这里采用K-medoid 算法确立样本聚集度模型，利用它对已标注样本集中样本聚类，找到K个具有代表性的样本c_1, c_2, …, c_K，这些样本满足式 (6) 的要求，即已标注样本集中的所有样本与和它最近的代表性样本之间的距离之和最小。

$$\sum_{i=1}^{n}\min_{k=1,\ \cdots,\ K}d(x_i,\ c_k) \tag{6}$$

式中，n为已标注样本集中样本个数。其中c_1根据$\min\sum_{i=1}^{n}d(x_i,\ c_1)$确定。确定$c_1$之后根据式 (6) 依次确

定c_2, …, c_K，这里取K=10，即每次选取 10 个具有代表性的样本。

K个具有代表性的样本确定之后，在未标注样本集中反复执行式 (7) 和 (8) 得到类别先验概率α_k，其中σ^2是已标注样本集中样本方差，且对于不同类别都是一样的。

$$p(k|x_i)=\frac{\alpha_k\exp\left(-\frac{1}{2\sigma^2}\|x_i-c_k\|^2\right)}{\sum_{k'=1}^{K}\alpha_{k'}\exp\left(-\frac{1}{2\sigma^2}\|x_i-c_{k'}\|^2\right)} \tag{7}$$

$$\alpha_k=\frac{1}{n}\sum_{i=1}^{n}p(k|x_i) \tag{8}$$

根据类别先验概率α_k得到未标注样本的先验分布概率$p(x_i)$，那么样本代表性置信度模型最终确立，如下式所示

$$p(x_i)=\sum_{k=1}^{K}\alpha_k\exp\left(-\frac{1}{2\sigma^2}\|x_i-c_k\|^2\right) \tag{9}$$

$p(x_i)$的值越大，则说明未标注样本x_i越具有代表性，是应当充分考虑进行标注的样本。

1.5 基于样本不确定性和代表性相结合的可控主动学习算法

为了平衡所选样本的不确定性和代表性，样本选择策略应同时考虑两种类型的样本：靠近分类边界不确定的样本和具有代表性的样本。因此将样本不确定性置信度模型和样本代表性置信度模型综合考虑，构造了样本不确定性与代表性相结合的可控主动学习算法，由下式给出

$$s=\arg\max_{i\in L}(1-C(x_i))p(x_i) \tag{10}$$

为了快速提高分类器性能，初始时设定每次选择多个样本进行标注。由于最初分类器性能不高，可提高的空间较大，这样做可以较快改善其性能。随着学习次数增多，分类器性能提高的幅度趋缓，这时应考虑减少每次标注的样本数量。为此将每次要标注的样本个数t和当前分类器分类正确率η联系起来，如果当前分类正确率与前一次正确率的差值绝对值小于阈值δ

$$|\eta(i)-\eta(i-1)|<\delta \tag{11}$$

则每次要标注的样本个数t就在原基础上作相应的调整，具体规则由下式所示

$$t_{new}=\begin{cases}t_{old}-1 & t_{old}>1\\ t_{old} & t_{old}=1\end{cases} \tag{12}$$

选取阈值δ=0.05。这个简单的规则实现了分类正确率的变化$\Delta\eta$对每次要标注的样本个数t的控制，减小了标注代价。

主动学习目的是学习较少样本获得最多学习信息和最佳学习效果。每次通过对新样本的学习，分类器性能逐渐趋于稳定，便自动终止学习过程。

2 实验结果及分析

在本节，针对本文提出的可控主动学习算法的有效性和合理性进行验证。在 MATLAB 7.0 环境下进行了实验仿真，使用 LIBSVM 工具箱，实验数据采用 UCI 标准测试数据集。本文主要针对两类分类问题，采用的数据集共 6 个，其中 Iris*数据集只采用其中的 versicolor 和 virginica 两类数据。数据集的详细信息如表 1 所示。

表 1 实验数据集信息

Tab. 1 Information of datasets used in experiments

Dataset	Instances	Attribute	Class	Pos/Neg
Ionosphere	351	34	2	64.1%/35.9%
Statlog (Heart)	270	13	2	44.4%/55.6%
Haberman's Survival	306	3	2	73.5%/26.5%
Pima Indians Diabetes	768	8	2	65.1%/34.9%
German Credit Data	1000	20	2	70.0%/30.0%
Iris*	100	4	2	50.0%/50.0%

在使用这些数据集之前，首先需要对数据进行归一化处理，这是因为需要计算样本属性值之间的欧氏距离，而该计算容易出现由于样本属性值取值范围的差异造成一个数据影响另一个数据的情况。实验可以直接调用 LIBSVM 工具箱中的 Scale 函数将数据归一化，之后数据属性值的取值空间将被限定在−1~1 之间。测试时，将数据集随机打乱，然后取 3%作为已标注样本集，再将剩余样本分成 4∶1 的两份，大的一份作为未标注样本集，小的一份作为测试集。运行采样算法，每次选择t个样本标注，直至满足主动学习算法的终止策略为止。实验采用十折交叉验证（ten-fold cross validation）的方法，每种采样算法针对每个数据集的测试重复

10 次，取达到目标正确率时所需的平均采样数量作为评价标准。使用的基准学习算法为 SVM，它的参数设置是：类型为 C-SVC，核函数为 RBF (Radial Basis Function)，参数C与g通过交叉验证每次均取最佳参数。另外，采用随机采样方法作为对比算法，为了增加对比度，其余准则均与主动学习相同。

这里作为对比算法的随机采样方法可以认为是一种被动学习，被动地接受随机选定的训练样本，这种被动方式忽略了学习者对新样本学习的积极性。主动学习则不同，它从未标注样本集中选择最有价值的样本，可以更快改善分类器的性能。根据 PAC 理论，为获取期望错误率小于ε的分类器，被动学习算法的样本复杂度为$O\left[\frac{1}{\varepsilon}\ln\left[\frac{1}{\varepsilon}\right]\right]$，而主动学习算法能够将样本复杂度减少到$O\left[\ln\left[\frac{1}{\varepsilon}\right]\right]$。另外，这里对于可控主动学习算法进行了简单的时间复杂度估算。计算各样本的奇异值需要耗费$O(m^2)$的时间开销，且一旦有一个新样本加入到某一类已标注样本集中，便需为此类中的所有样本重新计算奇异值，并且代表性样本的选择也需重新计算，所以时间开销是比较大的。因此，样本集的数据规模和样本所对应的特征向量维数是影响本算法时间复杂度的主要因素。尽管主动学习选择策略的计算量和时间开销较大，但所需人工标注的样本数量比随机采样少很多，极大降低了人工标记的工作量。

实验特别关注了当分类正确率达到几乎同等较高水平时，采用主动学习方法所需人工标注的样本数量与随机采样方法所需人工标注的样本数量的对比结果，如表 2 所示。

表 2　数据集上达到目标正确率所需人工标注的样本数量

Tab. 2　Number of labeled samples for obtaining the target accuracy

Dataset	Random	TA/%	Active	TA/%
Ionosphere	67.4	91.33	37.5	92.06
Statlog (Heart)	46	82.04	26.9	82.85
Haberman's Survival	42.3	74.55	21.7	74.55
Pima Indians Diabetes	381.2	76.74	135.7	78.61
German Credit Data	354	76.30	190.6	76.33
Iris*	10.8	91.38	6.6	92.63

注：表 2 中 TA 表示目标正确率

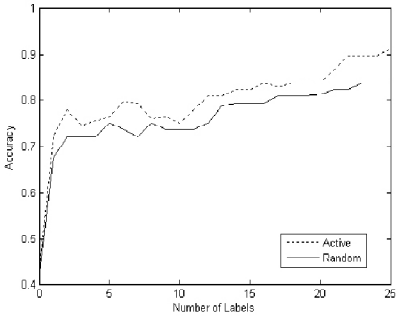
容易看出在实验数据集上主动学习算法的性能要比随机采样算法的性能好，要达到相同的分类性能，主动学习算法需要较少的人工标注样本数量。其中，在 Pima Indians Diabetes 数据集上，主动学习比随机采样算法节约了将近 2/3 的样本标记工作量，大大减小了标记代价。为了进一步证明主动学习算法的优势，表 3 列出了针对每一个数据集，主动学习所需标注的样本数与随机采样所需标注的样本数的比率，较容易看出主动学习比随机方法节约采样将近 50%。

表 3　主动学习与随机方法采样数目之比

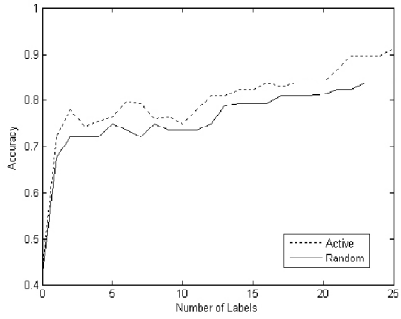
Tab. 3　Comparison of the number of sampling between active learning and random sampling

Dataset	Active/Random (%)
Ionosphere	55.6
Statlog (Heart)	58.5
Haberman's Survival	51.3
Pima Indians Diabetes	35.6
German Credit Data	53.8
Iris*	61.1

图 2 给出了在 Ionosphere 和 Heart 数据集上主动学习算法与随机方法性能改进曲线。其中，纵轴给出分类器的分类正确率，横轴给出所需人工标注的样本数目。从学习曲线图上可看出，尽管曲线有小幅度的波动但总体上是上升的，且主动学习算法的上升速度更快，其性能明显优于随机采样方法。



(a) Ionosphere 数据集学习曲线



(b) Heart 数据集学习曲线

图 2　Ionosphere 和 Heart 数据集测试结果

Fig. 2　Test results on ionosphere and heart

3 结束语

本文针对主动学习算法中的采样问题，分析了基于样本 *k*NN 分布状况的不确定性置信度模型和基于样本先验分布的代表性置信度模型，将两种思路结合，提出了一种新的主动学习算法，实现了每次采样数目的可控。这种算法使得每次选择的样本更具有“价值”，最后的实验结果表明该方法可以有效减少主动学习算法的采样次数和所需样本数量，减小标注代价。如何将不确定性置信度模型与代表性置信度模型更好的融合起来，如何减少算法的计算量等都是值得进一步研究的问题。

参考文献

[1] Deng Chao, Guo Maozu. Tri-training and data editing based semi-supervised clustering algorithm [J]. Journal of Software, 2008,19 (3): 663-673.

[2] Steven C H Hoi, Rong Jin, Jianke Zhu, et al.. Batch mode active learning and its application to medical image classification [C] // Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, 2006: 417-424.

[3] Gong Xiujun, Sun Jianping, Shi Zhongzhi. An active Bayesian network classifier [J]. Journal of Computer Research and Development, 2002,39 (5): 574-579.

[4] Long Jun, Yin Jianping, Zhu En, et al.. An active learning algorithm by selecting the most possibly wrong-predicted instances [J]. Journal of Computer Research and Development, 2008,45(3):472-478.

[5] Nguyen H T, Smeulders A. Active learning using pre-clustering [C] // Proceedings of the twenty-first international conference on Machine learning, Banff, Alberta, Canada, 2004: 79-86.

[6] Constantinos Constantinopoulos,Aristidids Likas. Semi-supervised and active learning with the probabilistic RBF classifier [J]. Neurocomputing, 2008,71 (13): 2489-2498.

[7] Cohn D A, Ghahramani Z, Jordan M I. Active learning with statistical models [J]. Journal of Artificial Intelligence Research, 1996,4: 129-145.

[8] Roy N, McCallum A. Toward optimal active learning through sampling estimation of error [C] // Proceeding of 18th International Conference on Machine Learning, San Francisco, CA, 2001: 441-448.

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

[10] Mingkun Li, Ishwar K. Sethi. Confidence-Based active learning [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006,28 (8): 1251-1261.

[11] Xu Z, Yu K, Tresp V, et al.. Representative sampling for text classification using support vector machines [C] //25th European Conference on Information Retrieval Research, Pisa, Italy, 2003: 393-407.

[12] Li Fayin. Open set face recognition using transduction [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27 (11): 1686-1697.

Controlling active learning algorithm based on uncertainty and representative of data selection

HU Zheng-ping[1], GAO Wen-tao[1], WAN Chun-yan[2]

(1. College of Information Science and engineering, Yanshan University, Qinhuangdao, Hebei 066004, China; 2. Qiqihaer third middle school, Qiqihaer, Heilongjiang 161000)

Abstract: Active learning algorithm can alleviate effectively the efforts of ineffective labeling instances by selecting the most informative examples for experts to label. Fully considered the uncertain samples close to the classification boundary and representative samples near the center of the prior data distribution, the controlling active learning method based on uncertainty and representative of data selection is presented. Firstly, uncertainty confidence level model is constructed using *k*NN distribution of samples, and the idea needn't consider specific type and parameters calculation of samples distribution. Then representative confidence level model is introduced based on samples aggregation model. Finally, the two different confidence level models are combined together to form controlling active learning method that could select the most valuable samples in each training step. The simulation experimental results show that this method is valid and efficient, and it selects fewer instances than random sampling on used UCI datasets when obtaining the same target accuracy.

Key words: controlling active learning; uncertainty samples; prior data distribution; representative samples