文章编号:1671-4598(2014)06-1938-03　　中图分类号:TP18; TP391.4; TP301.6　　文献标识码:A

融合主动学习的改进贝叶斯半监督分类算法研究

刘建峰，吕　佳
(重庆师范大学 计算机与信息科学学院，重庆　401331)

摘要：半监督学习是人工智能领域一个重要的研究内容；在半监督学习中，如何有效利用未标记样本来提高分类器的泛化性能，是机器学习研究的热点和难点；主动学习可解决未标记样本有效利用的问题，将主动学习引入到半监督分类中，并改进贝叶斯算法，提出了一种基于改进贝叶斯算法的主动学习与半监督学习结合算法；实验结果表明，该方法取得了较好的分类效果。

关键词：半监督分类；主动学习策略；概率模型；贝叶斯分类；KL距离

Study on Improved Bayesian Algorithm Semi-supervised Classification Integration of Active Leaning

Liu Jianfeng, Lv Jia
(College of Computer and Information Science, Chongqing Normal University, Chongqing　401331, China)

Abstract: As for semi-supervised learning, it is an important field of artificial intelligence research, how to make full use of unlabeled data to improve the generalization performance of classifier is a hot and difficult issue in machine learning. Active learning can effectively utilize unlabeled data, and it is introduced to semi- supervised classification, and improved Bayesian algorithm. Combination of active learning and semi-supervised learning based on improved Bayesian algorithm was proposed in this paper. Experimental results show that the algorithm is effective and feasible.

Key words: semi-supervised classification; active learning strategy; probability model; Bayes classification; KL distance

0　引言

人工智能领域中，对于如何完成大量数据、文本、网页、邮件等自动分类、预测和控制，很多学者进行了重点研究，并取得了很多成果。半监督学习作为人工智能领域一个研究的热点，研究者对未标记数据分类中的作用表现出很大的兴趣[1]。半监督学习在学习过程中对于未标记样本的利用是被动地选择，结果往往造成学习性能下降。很多学者提出基于主动学习策略的半监督分类算法[2-5]。文献[2]提出了主动选择距离SVM分类超平面最近的点，即信息量最大的点是进行查询注释。文献[3]提出了一种基于图的局部和全局主动选择的半监督分类。文献[4]提出主动半监督支持向量机，该算法利用主动学习来选择类边界样本。文献[5]提出了S-SOINN算法，该算法是建立在自增量神经网络的基础上选择样本。

半监督分类模型中，常在概率分布上进行建模，比较有代表性是基于朴素贝叶斯假设的概率模型和EM算法[6]。贝叶斯分类器诸多算法中朴素贝叶斯分类模型是最基本的[7]，该模型是基于特征属性间彼此独立的假设，实际上样本特征属性对分类的贡献不一样[8]。这样加权贝叶斯分类器被提出来，Webb等[9]提出了APNBC方法，使加权参数作用于类别节点上，但是该算法二次加权后需要进行线性调整，降低了算法对复杂数据处理的性能。Mark Hall[10]利用决策树对属性加权，但是该方法要多次扫描数据集，从而导致算法低效。

1　半监督分类问题

半监督分类问题描述如下：

给定(1)式所示训练集，根据算法确定 $x_{l+1},x_{l+2}\cdots,x_n$ 对应的在 $\{-1,+1\}$(类标号)中取值的输出 $y_{l+1},y_{l+2},\cdots,y_n$ 的值。

$$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_l,y_l),x_{l+1},\cdots x_n\} \quad (1)$$

其中：$n=l+u,u>>l,x_i\in R^d(i=1,2,\cdots,n),y_i\in\{+1,-1\}(i=1,2,\cdots,l)$。

半监督分类的目的就是通过训练这些少量有标记样本和大量的未标记样本，得到一个精确度较高的分类模型，从而完成对未标记数据的标记和分类。从概率的角度来说是利用训练样本的输入边缘概率 $P(x)$ 和条件输出概率 $P(y\mid x)$，来确定样本 X 所属的类别，在一定假设的前提下，$P(x)$ 和 $P(y\mid x)$ 之间存在相关的联系，通过未标记样本获得关于 $P(x)$ 的知识来推测 $P(y\mid x)$ 的结果。

2　主动学习策略

结合 KL 距离能有效选择出信息量最大的样本这一的特点，提出了一种基于 KL 距离的主动选择策略。对于离散型的概率分布：样本 M 和 N 之间的KL距离定义如下[11]：

$$KL(M\|N)=\sum_{i=1}^{t}m_i\lg\frac{m_i}{n_i} \quad (2)$$

其中：$M=\{m_i\},N=\{n_i\},i=1,2,\cdots,t,m_i$ 和 n_i 分别代表样本 M 和 N 的每个属性。当 M 和 N 距离越接近，它们的 KL 距离越小，分布越相似。但是，由于 KL 距离不满足对称性，不具有三角形，因此定义了一种对称的距离

收稿日期:2014-01-19；　**修回日期**:2014-02-26。
基金项目:国家自然科学基金数学天元基金项目(11326189)。
作者简介:刘建峰(1984-)，男，河南濮阳人，硕士研究生，主要从事机器学习和数据挖掘方向的研究。
吕　佳(1978-)，女，四川达州人，教授，博士，硕导，主要从事机器学习、数据挖掘及最优化技术方向的研究。

$$KL(M\|N)=\frac{1}{2}KL(M||N)+\frac{1}{2}KL(N\|M) \tag{3}$$

这个公式经常被用于分类中的特征选择。本文令 P 和 Q 分别代表样本的类别后验概率，即：$P=p(c_i\mid x)$，$Q=q(c_j\mid x)$，$i\neq j$ 且 $i,j\leqslant s$，s 为类别数，根据对称的 KL 距离计算公式，P、Q 之间的 KL 距离可改写为

$$KL\{p(c_i\mid x)\|p(c_j\mid x)=+\frac{1}{2}\sum_{i=1}^{s}p(c_j\mid x)\cdot\lg\frac{p(c_j\mid x)}{p(c_i\mid x)} \tag{4}$$

式（4）为样本 x 属于第 i 类和属于第 j 类的概率的差值，差值越大样本所属类别就越容易确定；反之，样本就处于比较模糊的边界。其中，$p(c_i\mid x)$ 是样本的类属后验概率，样本 x 属于第 i 类的概率越大，就确定为该样本属于该类。因此关键的是确定 $\max p(c_i\mid x)$ 的值，计算公式如下

$$\max p(c_i\mid x)=\frac{\max p(x\mid c_i)p(c_i)}{p(x)} \tag{5}$$

由于 $p(x)$ 对于所有类别均为常数，公式（4）变形为

$$\max p(c_i\mid x)=\max p(c_i)\prod_{k=1}^{t}p(x_k\mid c_i)^{w_{A_{k,v,i}}} \tag{6}$$

其中：$w_{A_{k,v,i}}$ 的具体定义见公式（15），此处的 $w_{A_{k,v,i}}$ 即是（15）中的 $w_{A_{k,v,i}}$。根据公式（3）和公式（5），令 $L_{i,j}$ 表示未标记样本 x 到各类之间的 KL 距离，计算公式为

$$\begin{aligned}L_{i,j}=&\frac{1}{2}p(c_i)\prod_{k=1}^{t}p(x_k\mid c_i)^{W_{Ak,v,i}}\left[\lg\frac{p(c_i)}{p(c_j)}+\right.\\&\left.\sum_{k=1}^{t}\lg\frac{p(x_k\mid c_i)^{W_{Ak,v,i}}}{p(x_k\mid c_j)^{W_{Ak,v,j}}}\right]+\\&\frac{1}{2}p(c_j)\prod_{k=1}^{t}p(x_k\mid c_j)^{W_{Ak,v,j}}\left[\lg\frac{p(c_j)}{p(c_i)}+\right.\\&\left.\sum_{k=1}^{t}\lg\frac{p(x_k\mid c_j)^{W_{Ak,v,j}}}{p(x_k\mid c_i)^{W_{Ak,v,i}}}\right]\end{aligned} \tag{7}$$

若一个未标记样本的 $L_{i,j}$ 处在预先设定的阈值 δ 范围内时，认定该样本是处在不同类之间的一个比较模糊的边界上，即该样本为不确定性最强的那个样本也就是信息量最大的样本，此时，交由专家（训练的 h 个初始分类器组成）标记，选择专家标记最多的那个类作为该样本的类标记。

3　基于改进贝叶斯半监督分类模型

利用贝叶斯分类的优点，提出改进加权方法以克服以上算法的缺陷。半监督分类模型如下所示，其中，$p(c_i)$ 是类的先验概率，$p(x_k\mid c_i)$ 是样本类属先验概率计算如下

$$c_i(x)=\underset{c_i\in C}{\operatorname{argmax}}\left\{p(c_i)\prod_{k=1}^{n}p(x_k\mid c_i)^{w_{Ak,t,i}}\right\} \tag{8}$$

$$p(c_i)=\frac{s_i}{s},p(x_k\mid c_i)=\frac{s_{ik}}{s_i}$$

其中：s_i 是类 c_i 中的训练样本数，s 是训练样本总数，s_{ik} 是特征属性中具有 x_k 的类 c_i 的训练样本数。公式（9）只适用于处理离散属性的样本，若样本 x 的属性是连续性的，一般假定该属性是遵循高斯分布的，此时计算公式如下

$$p(x_k\mid c_i)=\frac{1}{\sqrt{2\pi}\sigma_{c_i}}e^{\frac{(x_k-\mu_{c_i})}{2\sigma_{c_i}^2}} \tag{9}$$

其中：σ_{c_i}，μ_{c_i} 分别为特征属性的标准差和平均值。对于类 $c_i,c_j\in C,(i\neq j)$，样本 x 类别的判别准则如下：

$$\begin{cases}p(c_i\mid x)\geqslant p(c_j\mid x), & x\in\{c_i\}\\ p(c_i\mid x)<p(c_j\mid x), & x\in\{c_j\}\end{cases} \tag{10}$$

其中：$C=\{c_1\cup c_2\cup\cdots\cup c_m\}$。未标记样本 $x_{l+1},x_{l+2},\cdots,x_n$ 对应的标记 $y_{l+1},y_{l+2},\cdots,y_n$ 按下式推断

$$\operatorname{sign}(c_i(x))=\begin{cases}-1,c_i(x)<c_j(x)\\+1,c_i(x)\geqslant c_j(x)\end{cases} \tag{11}$$

由于样本属性之间是相互关联的，彼此独立性假设在分类的过程中会带来一些误差，因此使用贝叶斯算法作为分类器时需对属性加权，文献［12］采用的属性加权方法如下

$$w_{A_{k,t,i}}=\frac{I(A_k=t\wedge c_i)}{I(A_k=t)} \tag{12}$$

该方法未考虑数据集中有些样本特征属性欠缺的情况，这样就会造成分母为零的情况发生，因此本文将其改进为

$$w_{A_{k,t,i}}=\frac{I(A_k=t\wedge c_i)+1}{I(A_k=t)+1} \tag{13}$$

其中：$I(A_k=t\wedge c_i)$ 代表训练集中，类 c_i，第 k 个属性取 t 的样本数，$I(A_k=t)$ 表示训练集中属性 A_k 取值为 t 的样本数。

4　融合主动学习的改进贝叶斯半监督分类算法

本文结合主动学习策略和改进加权贝叶斯分类模型，提出了基于改进加权贝叶斯算法的主动学习与半监督学习结合算法（Active Learning KL Semi－supervised，ALKLSS）。算法分两个阶段进行：第一阶段，利用改进加权贝叶斯分类算法对有标记样本进行初始分类；第二阶段计算 KL 距离对未标记样本进行主动选择，选出信息量最大的样本，交由专家标记，再用改进贝叶斯进行分类。算法的框架如下：

输入：训练样本集如式（1）所示。

输出：未标记样本 $\{x_{l+1},x_{l+2},\cdots,x_n\}$ 的类标记 $\{y_{l+1},y_{l+2},\cdots,y_n\}$。

Step 1：用公式（8）对有标记样本进行训练，得到初始分类 $\{c_i\}_{i=1,2,\cdots,p}$。

Step 2：从有标记样本 X_L 中随机选择 m 个样本，进行训练，得到 h 个分类器作为标注专家 $H=\{h_1,h_2,\cdots,h_h\}(h\leqslant l)$。

Step 3：$\forall x_i\in U$，按照公式（6）计算后验概率 p_i。

Step 4：应用主动选择策略选择未标记样本 x_i 进行标记。

（1）选取初始阈值 $\delta\geqslant 0$，用公式（7）计算未标记样本 x_i 与所有类之间的 KL 距离 $L_{i,j}$。

（2）H 主动选择满足 $L_{i,j}\leqslant\delta$ 的 x_i 进行标记。

（3）x_i 加入到 X_L 中，更新 X_L 和 X_U。

（4）反复执行（1）～（3），当 $L_{i,j}\geqslant\delta$ 时，完成对信息量最大的未标记样本选择和标记。

Step 5：根据公式（8）计算 $c_i(x)$，进行半监督分类。

Step 6：根据公式（13），确定 $x_{l+1},x_{l+2},\cdots,x_n$ 的类标号。

5　实验验证与分析

5.1　数据集、方法和参数

实验数据集选用 UCI 数据库中 9 个标准数据集，其详细描述见表 1[13]。每个训练集随机地选择 15％作为有标记样本，去除其他样本的类标号，训练初始分类器，并从此训练集中随机的选择 m 个标记样本作为新的训练集，训练得到 H 评判专家。本算法设置参数如下：$m\in\{50,100,150,200,250\}$；$H_l\in\{3,5,7,9,11\}$，$\delta\in\{0,0.01,0.02,0.05,0.09,0.1,0.2\}$。

为了说明本文算法的有效性，将其与朴素贝叶斯算法（Naive Bayesian Classifier，NBC），属性加权贝叶斯算法（Weighted Naive Bayesian，WNB）和基于QBC主动学习算法（QBC）进行比较。

表1 数据集描述

数据集	数据集规模	数据集特征
Heart	270	13
Waveform	5 000	19
Cleve	296	10
Monks	248	7
Vote	435	16
Breast	683	10
Chess	3 196	36
Sick	3 772	29
Credit－rating	690	15

5.2 实验结果

实验重复20次，依次得到在9个数据集在4个算法上的正确率（*CR*）和标准误差（*SE*），如表2所示。从表2中可以看出算法除在数据集Chess和Vote外，本文算法在其余数据集上均取得了较好的效果。分析发现Chess和Vote这两个数据集样本属性之间的相关性很强，我们提出的属性加权模型需要更复杂的形式才能很好地表达此属性对分类的影响，而其标准误差要低于其他算法。

表2 算法正确率和标准误差比较结果

Datasets	NBC		WNB		QBC		ALKLSS	
	CR	*SE*	*CR*	*SE*	*CR*	*SE*	*CR*	*SE*
Heart	83.01	2.79	84.28	2.54	83.45	3.02	85.50	1.64
Waveform	79.56	3.10	81.64	2.28	81.69	2.79	83.67	2.15
Cleve	82.44	1.56	83.05	2.19	83.56	1.68	84.59	1.40
Monks	77.90	2.59	78.20	2.54	78.80	2.60	79.06	2.39
Vote	90.27	3.12	91.21	2.89	91.13	3.03	91.37	1.19
Breast	96.02	2.22	96.89	2.11	97.09	2.21	98.18	1.45
Chess	77.89	1.76	78.67	1.57	77.90	0.98	78.85	0.87
Sick	92.61	2.35	93.69	2.24	93.53	1.99	95.43	0.98
Credit－rating	77.69	3.23	78.78	2.94	79.87	2.81	83.10	2.44

6 结束语

本文提出的基于改进贝叶斯的主动学习与半监督学习结合算法，集中了半监督学习和主动学习算法的优势，避免了由于被动接受数据而带来的分类效果不理想问题，同时，得到的分类器也可以自动预测和控制分类，实验表明本文算法与朴素贝叶斯、加权贝叶斯和基于QBC主动学习算法相比，分类效果更好，分类精度更高，解决了人工智能领域的一些问题。下一步工作，将改进半监督学习的效率，提高算法执行效率，并将其应用到多分类问题中。

参考文献：

[1] Letouzey F, Denis F, Gilleron R. Learning from positive and unlabeled examples [A]. Proceedings of the 11th International Conference on Algorithmic Learning Theory [C]. Sydney, Australia, 2000, 71－85.

[2] Hu L S, Lu S X, Wang X Z. A new and informative active learning approach for support vector machine [J]. Information Sciences, 2013, 244 (9): 142－160.

[3] Liu L, Xie Y G, Wang Z L, et al. A new graph based active learning method [J]. Procedia Engineering, 2012, 29: 2610－2620.

[4] Leng Y, Xu X Y, Qi G H. Combining active learning and semi－supervised learning to construct SVM classifier [J]. Knowledge－Based Systems, 2013, 44 (5): 121－131.

[5] Shen F R, Yu H, Sakuirai K. An incremental online semi－supervised active learning algorithm based on self－organizing incremental neural network [J]. Neural computing & Application, 2011, 20 (7): 1061－1074.

[6] Nigam K, Mccallum A, Thrun S, et al. Text classification from labeled and unlabeled documents using EM [A]. International Conference on Machine Learning [C]. Stanford, USA, 2000, 39 (2－3): 103－134.

[7] 邓桂骞，赵跃龙，刘 霖，等. 一种优化的贝叶斯分类算法 [J]. 计算机测量与控制，2012，20 (1)：199－201.

[8] 杨 帆，张彩丽. 基于粗糙集理论的贝叶斯故障诊断方法研究 [J]. 计算测量与控制，2007，15 (11)：1470－1477.

[9] Webb G I, Pazzan J. Adjusted probability Naive Bayesian induction [A]. Proceedings of the 11th Australian Joint Conference on Artificial Intelligence [C]. Berlin: Springer Verlag, 1998: 285－295.

[10] Hall M. A decision tree－based attribute weighting filter for Naive Bayes [J]. Knowledge－Based Systems, 2007, 20 (2): 120－126.

[11] 许 震. 基于KL距离的半监督分类算法 [D]. 上海：复旦大学，2010.

[12] 秦 锋，任诗流，程泽凯，等. 基于属性加权的朴素贝叶斯分类算法 [J]. 计算机工程与应用，2008，44 (6)：107－109.

[13] Blake C L, Merz C J. UCI repository of machine learning databases [R/OL]. University of California, Irvine, Department of Information and Computer Science, 1998. http://www.ics.uci.edu/～mlearn/MLRepository.html.

（上接第1937页）

参考文献：

[1] 陈彩华，龙卫兵，刘 彬. 基于ARM－Linux的家用网络平台设计与实现 [J]. 计算机测量与控制，2010，18 (9)：2176－2177.

[2] 刘 余，孟小华. 嵌入式智能家居终端通信模块的设计与实现 [J]. 计算机工程与设计，2010，(8)：1689－1692.

[3] 张云川，王正勇，卿粼波，等. 基于ARM的便携式视频解码终端设计与实现 [J]. 计算机工程，2009，35 (4).

[4] 戢卫平，罗 飞，曹建忠，等. 基于ARM平台和GPRS的远程监控系统 [J]. 计算机应用研究，2006，23 (6)：196－198.

[5] 吴金华，郑 耿，李驹光. 基于ARM9的无线数据终端的设计与实现 [J]. 计算机工程，2008，34 (14)：253－255.

[6] 赵晓军，苏海霞，任明伟，等. 基于ARM9和CAN总线的远程监控系统 [J]. 计算机工程，2010，36 (5)：231－233.

[7] 陶 永，鄢 萍，郭建兴，等. 基于MIPS体系的嵌入式Linux引导装载系统的设计与实现 [J]. 计算机应用，2004，24 (11)：159－161.

[8] 鲁 力，张 波. 嵌入式TCP/IP协议的高速电网络数据采集系统 [J]. 仪器仪表学报，2009，30 (2)：405－409.