在线序列主动学习方法

翟俊海[1] 臧立光[2] 张素芳[3]
(河北大学数学与信息科学学院河北省机器学习与计算智能重点实验室 保定 071002)[1]
(河北大学计算机科学与技术学院 保定 071002)[2] (中国气象局气象干部培训学院河北分院 保定 071000)[3]

摘 要 现实世界中存在着大量无类标的数据,如各种医疗图像数据、网页数据等。在大数据时代,这种情况更加突出。标注这些无类标的数据需要付出巨大的代价。主动学习是解决这一问题的有效手段,也是近几年机器学习和数据挖掘领域中的一个研究热点。提出了一种基于在线序列极限学习机的主动学习算法,该算法利用在线序列极限学习机增量学习的特点,可显著提高学习系统的效率。另外,该算法用样例熵作为启发式度量无类标样例的重要性,用K-近邻分类器作为Oracle标注选出的无类标样例的类别。实验结果显示,提出的算法具有学习速度快、标注准确的特点。

关键词 主动学习,极限学习机,在线序列学习,样例熵,K-近邻
中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.01.007

Online Sequential Active Learning Approach

ZHAI Jun-hai[1] ZANG Li-guang[2] ZHANG Su-fang[3]
(Key Laboratory of Machine Learning and Computational Intelligence, College of Mathematics and Information Science, Hebei University, Baoding 071002, China)[1]
(College of Computer Science and Technology, Hebei University, Baoding 071002, China)[2]
(Hebei Branch of China Meteorological Administration Training Center, China Meteorological Administration, Baoding 071000, China)[3]

Abstract In the real world, there are a lot of unlabelled data, such as various medical images and web data, etc. In the era of big data, this situation is more prominent. It is expensive to label large amount of unlabelled data. Active learning is an effective method to solve this problem, and it is one of the hot research topics in the field of machine learning and data mining. Based on online sequential extreme learning machine, an active learning algorithm was proposed in this paper. Due to the nature of incremental learning embedded in online sequential extreme learning machine, the proposed algorithm can significantly improve the efficiency of learning system. Furthermore, the proposed algorithm uses instance entropy as heuristic to measure the importance of the unlabeled instances, and uses K-nearest neighbor classifier as Oracle to label the selected instances. The experimental results show that the proposed algorithm has fast learning speed with exact labeling.

Keywords Active learning, Extreme learning machine, Online sequential learning, Instance entropy, K-nearest neighbors

1 引言

主动学习是一种有监督学习。与传统的被动学习不同,在主动学习中,学习器不是被动地接受、处理人类提供的所有数据,而是主动地选取它所认为最有价值的数据,并交由领域专家进行标注。主动学习的目标是在可接受精度的前提下,选取尽可能少的样例以减小标注和学习的代价。

主动学习可以追溯到20世纪80年代末、90年代初,最早的主动学习算法是Angluin于1988年提出的Query算法[1],它是贝叶斯框架下的一种主动学习算法。分类器根据当前的版本空间产生了一个新的样例并由专家标注其类别,分类器根据专家标注的类别更新版本空间,进行下一轮的学习,如此循环直至学习到目标概念。其他代表性工作包括Seung等提出的基于委员会的样例选取方法[2]、Lewis等人提出的基于最大不确定性的样例选取方法[3]和Cohn等人提出的基于期望误差缩减的样例选取方法[4]。基于委员会的样例选取算法是对Query算法的改进,分类器通过Query算法得到委员会分歧最大的样例,交给专家标注其类别,得到答案后,分类器修改版本空间,得到一个新的委员会,重复上述过程,直至学习到目标概念。Lewis等提出的基于最大不确定性的样例选取方法与Query学习机制不同,它从无类标的样例集中选取当前分类器分类不确定性最大的样例,进行标注后再进行学习。在Lewis等工作的基础上,产生了很多基于不确定性的样例选取算法,如Schohn等提出了基于不确定性的SVM样例选取算法[4],该算法选择距离超平面最近的样例,即最小化Margin的样例。Wang X Z等人提出了基于最大不可指定性的样例选取算法[5],并将其用于训练模糊决策树。Tong等提出了基于平分版本空间的SVM样例选取方法,

到稿日期:2015-08-31 返修日期:2015-11-02 本文受国家自然科学基金项目(71371063),河北省自然科学基金项目(F2013201220),河北省高等学校科学技术研究重点项目(ZD20131028),河北省高等学校科学技术研究项目(QN20131153)资助。
翟俊海(1964—),男,博士,教授,主要研究方向为机器学习与数据挖掘,E-mail:mczjh@126.com;臧立光(1990—),男,硕士生,主要研究方向为数据挖掘;张素芳(1966—),女,副教授,主要研究方向为机器学习。

并将其成功应用于文本分类[6]。Cohn 等人提出了一种算法[7],它利用期望误差缩减最大作为样例选取的标准来选取样例,并应用于高斯模型和局部加权回归两种统计模型中。

主动学习大致可以分为两大类:基于流的主动学习和基于池的主动学习。基于流的主动学习根据分类器当前学习到的知识产生样例,分类器只能决定当前的样例是标注还是不标注,如果决定标注,则将它交由导师标注,否则将它舍弃。基于流的主动学习产生的样例可能根本不存在,或者无意义,这使得专家无法对其进行标注。与基于流的主动学习不同,基于池的主动学习不是由分类器产生新的样例,而是从样例池中选取样例。所谓的样例池就是无类标的样例的集合,供主动学习中的分类器选取样例。沿着这一方向,近年来,主动学习的研究重点主要集中在新的样例选取策略、主动学习的停止准则及应用研究上。Wang 等提出了基于不一致性的样例选取策略,并用选出的样例训练支持向量机[8]。Lughofer 提出了基于无监督准则和有监督准则相结合的混合主动学习方法[9]。Wang 等提出了基于正则化技术的主动学习方法[10]。针对极稀疏有类标样例问题,Sun 等提出了一种新的主动学习方法[11]。He 等[12]提出了一种基于马尔科夫等价类有向无环图的主动学习方法。Hoi 等[13]提出了分类模型不确定性的批量样例选取方法。高新波等[14,15]提出了基于嵌入式 Bootstrap 的主动学习示例选取方法。张长水等[16]采用多级学习方法来改进传统的半监督学习算法,该方法对含有不确定性的数据仍然有效,并给出选取最具价值的模糊示例的方法。Yu 等人提出了一种基于极限学习机的主动学习算法[17],该算法用极限学习机确定无类标样例的不确定性。基于元认知极限学习机,Zhang 和 Er 提出了一种序列主动学习算法[18],该算法由认知单元和元认知单元构成,认知单元是一种在线序列学习算法,元认知单元控制认知单元的学习过程。Gu 等人提出了一种组合不确定性和多样性的主动学习算法[19],并将其用于图像分类。Long 等人提出了一种针对排序学习的期望损失框架[20],并将其应用于主动学习算法。Huang 等人将信息量和代表性结合起来,提出了一种选择富含信息量且具有代表性的无类标样例的算法[21]。Hu 等人提出了一种基于支持向量机的主动学习方法[22],该方法选择最接近支持向量机超平面的样例进行标注,可有效降低标注无类标样例的代价。

本文提出了一种基于池的主动学习算法,该算法将在线序列极限学习机作为分类器,将样例熵作为选择样例的启发式,将 K-NN 作为专家标注样例。本文主要基于以下 3 点进行研究:首先,众所周知,主动学习是一个迭代学习的过程,每一次迭代都将专家标注的样例加入到训练集中,然后在新的训练集中重复学习。实际上,主动学习的过程是一种增量学习。而在线序列极限学习机(Online Sequential Extreme Learning Machine,OSELM)正好适应了主动学习中要不断添加数据到训练集的思路。其次,对 OSELM 的输出结果进行软最大化处理,是为了计算需要标注样例的信息熵。最后,使用K-NN算法代替专家进行标注,可实现完全的机器学习。本文提出的算法具有学习速度快、标注准确的特点。与 3 种相关算法进行了实验比较,实验结果显示本文提出的算法优于这 3 种算法。

2 基础知识

2.1 主动学习

主动学习是一类特殊的机器学习。本质上,主动学习是一种迭代采样技术,它从无类标的数据中按某种启发式选择重要的样例交给专家标注其类别。主动学习的目标是用最小的标注代价产生一个训练集,并用该训练集训练一个泛化能力强的分类器。主动学习问题可以表示为一个五元组 $AL=(L,U,C,S,Q)$。其中,$L=\{(x_i,y_i)\mid x_i\in R^d,y_i\in D\}(1\leqslant i\leqslant l)$ 是初始有类别的训练集;D 是类别的集合,如果样例属于 k 类,则 D 可以表示为 $D=\{1,2,\cdots,k\}$。$U=\{x_i\mid x_i\in R^d,y_i\in D\}(l+1\leqslant i\leqslant l+n)$ 是无类别样例的集合。一般地,n 远远大于 l。C 是一个分类器,即一个分类算法。S 是一个专家,其职责是正确标注选出的无类标样例的类别。Q 是一个启发式,用于度量无类标样例的重要性。主动学习流程如图 1 所示。

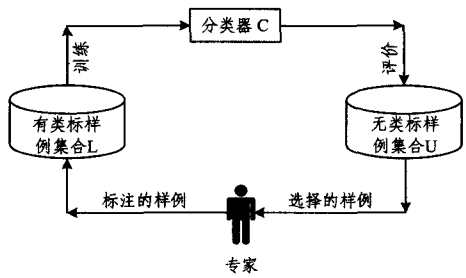

图 1 主动学习流程图

2.2 极限学习机

极限学习机[23]是一种用于训练单隐含层前馈神经网络的有效算法。给定数据集 $D=\{(x_i,y_i)\mid x_i\in R^d,y_i\in R^k\}$,$i=1,2,\cdots,n$,具有 m 个隐含结点的单隐含层前馈神经网络可以表示为:

$$f(x_i)=\sum_{j=1}^{m}\beta_j g(w_j\cdot x_i+b_j),i=1,2,\cdots,n \tag{1}$$

其中,$w_j=(w_{j1},w_{j2},\cdots,w_{jd})^{\mathrm{T}}$ 是连接第 j 个隐含结点和输入结点的权值向量,$\beta_j=(\beta_{j1},\beta_{j2},\cdots,\beta_{jm})^{\mathrm{T}}$ 是连接第 j 个隐含结点和输出结点的权值向量,b_j 是第 j 个隐含结点的偏置。

对于给定的训练数据集 D,其参数 $\beta_j(j=1,2,\cdots,m)$ 可以根据最小二乘法来进行估计。可以得到如下公式:

$$f(x_i)=\sum_{j=1}^{m}\beta_j g(w_j\cdot x_i+b_j)=y_i \tag{2}$$

式(2)的矩阵表示为:

$$H\beta=Y \tag{3}$$

其中:

$$H=\begin{pmatrix} g(w_1\cdot x_1+b_1) & \cdots & g(w_m\cdot x_1+b_m)\\ \vdots & \cdots & \vdots\\ g(w_1\cdot x_n+b_1) & \cdots & g(w_m\cdot x_n+b_m)\end{pmatrix} \tag{4}$$

$$\beta=(\beta_1^{\mathrm{T}},\cdots,\beta_m^{\mathrm{T}})^{\mathrm{T}} \tag{5}$$

$$Y=(y_1^{\mathrm{T}},\cdots,y_n^{\mathrm{T}}) \tag{6}$$

H 是隐含层输出矩阵,它通常不是方阵,一般无法得到式(3)的精确解。但是通过求解 H 的 Moore-Penrose 的广义逆,可以得到式(3)的最小范数最小二乘解:

$$\hat{\beta}=H^{\dagger}Y \tag{7}$$

其中,$H^{\dagger}$ 是矩阵 H 的 Moore-Penrose 广义逆。

极限学习机算法如算法 1 所示[23]。

算法 1 极限学习机算法

Input:训练集 $D=\{(x_i,y_i)\mid x_i\in R^d,y_i\in R^k,1\leqslant i\leqslant n\}$;激活函数 g;隐含层结点数 m

Output:权值矩阵 $\hat{\beta}$

1. for (j=1;i≤m;j=j+1) do
2. 随机给定输入权值 w_j 和偏置 b_j;
3. end

4. 计算隐含层输出矩阵 H；

5. 计算矩阵 H 的广义逆矩阵 $H^{\dagger}$；

6. 计算权矩阵 $\hat{\beta}=H^{\dagger}Y$；

7. 输出权矩阵 $\hat{\beta}$.

2.3 在线序列极限学习机

在线序列极限学习机是极限学习机的一种变型[24]，它采用序列学习策略。设矩阵 H 的秩，也就是隐含结点的个数为 m。在这种情况下，可以得到：

$$H^{\dagger}=(H^{T}H)^{-1}H^{T} \tag{8}$$

将式(8)代入式(7)，得到：

$$\hat{\beta}=(H^{T}H)^{-1}H^{T}Y \tag{9}$$

给定一批训练数据 $D_0=\{(x_i,y_i)\mid x_i\in R^d,y_i\in R^k\}$，$i=1,2,\cdots,n_0,n_0\geqslant m$。根据极限学习机算法，对于数据集 D_0，只需要考虑以下的优化问题：

$$\min_{\beta}\|H_0\beta-Y_0\| \tag{10}$$

$$H_0=\begin{pmatrix} g(w_1\cdot x_1+b_1) & \cdots & g(w_m\cdot x_1+b_m)\\ \vdots & \cdots & \vdots\\ g(w_1\cdot x_{n_0}+b_1) & \cdots & g(w_m\cdot x_{n_0}+b_m)\end{pmatrix} \tag{11}$$

$$Y_0=(y_1^{T},\cdots,y_{n_0}^{T}) \tag{12}$$

式(10)的解表示如下：

$$\beta^{(0)}=K_0^{-1}H_0^{T}Y_0 \tag{13}$$

其中，$K_0=H_0^{T}H_0$。

假设现有另一批数据 $D_1=\{(x_i,y_i)\mid x_i\in R^d,y_i\in R^k\}$，$i=p,\cdots,q$，其中 $p=n_0+1,q=n_0+n_1$，n_1 是这批样例集中样例的个数。相对的最优化问题变成：

$$\min_{\beta}\left\|\begin{bmatrix}H_0\\H_1\end{bmatrix}\beta-\begin{bmatrix}Y_0\\Y_1\end{bmatrix}\right\| \tag{14}$$

其中：

$$H_1=\begin{pmatrix} g(w_1\cdot x_p+b_1) & \cdots & g(w_m\cdot x_p+b_m)\\ \vdots & \cdots & \vdots\\ g(w_1\cdot x_q+b_1) & \cdots & g(w_m\cdot x_q+b_m)\end{pmatrix} \tag{15}$$

$$Y_1=(y_p^{T},\cdots,y_q^{T}) \tag{16}$$

输出权值变成：

$$\beta^{(1)}=K_1^{-1}\begin{bmatrix}H_0\\H_1\end{bmatrix}^{T}\begin{bmatrix}Y_0\\Y_1\end{bmatrix} \tag{17}$$

其中：

$$K_1=\begin{bmatrix}H_0\\H_1\end{bmatrix}^{T}\begin{bmatrix}H_0\\H_1\end{bmatrix}=K_0+H_1^{T}H_1 \tag{18}$$

$$\begin{bmatrix}H_0\\H_1\end{bmatrix}^{T}\begin{bmatrix}Y_0\\Y_1\end{bmatrix}=H_0^{T}Y_0+H_1^{T}Y_1$$
$$=K_1\beta^{(0)}-H_1^{T}H_1\beta^{(0)}+H_1^{T}Y_1 \tag{19}$$

结合式(15)和式(18)，可得：

$$\beta^{(1)}=\beta^{(0)}+K_1^{-1}H_1^{T}(Y_1-H_1\beta^{(0)}) \tag{20}$$

以此类推，第 $k+1$ 批数据集 $D_{k+1}=\{(x_i,y_i)\mid x_i\in R^d,y_i\in R^k\}$，$i=l,\cdots,r$，其中 $l=(\sum_{j=0}^{k}n_j)+1$，$r=\sum_{j=0}^{k+1}n_j$，n_{k+1} 是第 $k+1$ 批样例的数目，可以得出：

$$K_{k+1}=K_k+H_{k+1}^{T}H_{k+1} \tag{21}$$

$$\beta^{(k+1)}=\beta^{(k)}+K_{k+1}^{-1}H_{k+1}^{T}(Y_{k+1}-H_{k+1}\beta^{(k)}) \tag{22}$$

其中：

$$H_{k+1}=\begin{pmatrix} g(w_1\cdot x_l+b_1) & \cdots & g(w_m\cdot x_l+b_m)\\ \vdots & \cdots & \vdots\\ g(w_1\cdot x_r+b_1) & \cdots & g(w_m\cdot x_r+b_m)\end{pmatrix} \tag{23}$$

$$Y_{k+1}=(y_l^{T},\cdots,y_r^{T}) \tag{24}$$

根据 Woodbury 定理[24]，可得：

$$K_{k+1}^{-1}=(K_k+H_{k+1}^{T}H_{k+1})^{-1}$$
$$=K_k^{-1}-K_k^{-1}H_{k+1}^{T}(I+H_{k+1}K_k^{-1}H_{k+1}^{T})^{-1}H_{k+1}K_k^{-1} \tag{25}$$

令 $P_{k+1}=K_{k+1}^{-1}$，因此更新式(21)和式(22)：

$$P_{k+1}=P_k-P_kH_{k+1}^{T}(I+H_{k+1}P_kH_{k+1}^{T})^{-1}H_{k+1}P_k \tag{26}$$

$$\beta^{(k+1)}=\beta^{(k)}+P_{k+1}H_{k+1}^{T}(Y_{k+1}-H_{k+1}\beta^{(k)}) \tag{27}$$

在线序列极限学习机算法[24]（见算法 2）包含两个阶段：1)初始化阶段；2)顺序学习阶段。

算法 2　在线序列极限学习机算法

Input：训练集 $D=\{(x_i,y_i)\mid x_i\in R^d,y_i\in R^k,1\leqslant i\leqslant n\}$；激活函数 g；隐含结点数 m

Output：权值矩阵 $\hat{\beta}$

1. //初始化阶段 S_1：利用第一块数据 D_0 训练一个单隐含层前馈神经网络.
2. for (j=1;i≤m;j=j+1) do
3. 　　随机给定输入权值 w_j 和偏置 b_j；
4. end
5. 利用式(11)计算初始的隐含层输出矩阵 H_0；
6. 利用式(13)计算权矩阵 $\beta^{(0)}$；
7. 令 k=0；
8. //序列学习阶段 S_2：利用第(k+1)块数据 D_{k+1} 训练一个单隐含层前馈神经网络.
9. 利用式(23)计算 H_{k+1}；
10. 利用式(24)计算 Y_{k+1}；
11. 利用式(26)计算 P_{k+1}；
12. 利用式(27)计算 $\beta^{(k+1)}$；
13. 令 k=k+1，转 S_2.
14. 输出最终的权值矩阵 $\hat{\beta}$.

3 在线序列极限学习机主动学习算法

在本文提出的主动学习算法中，将在线序列学习机作为分类器。因为它是一种增量学习算法，所以每一次迭代都不用重新训练单隐含层前馈神经网络，这样可以显著提高算法的效率。

提出的算法需要用式(28)对神经网络的输出进行软最大化处理，如图 2 所示。软最大化处理后，神经网络的输出在区间[0,1]。此时，对于样例 x_i，第 j 个结点的输出 $p(\omega_j\mid x_i)$ 可以看作是样例 x_i 属于第 j 类的后验概率。

$$p(\omega_j\mid x_i)=\frac{e^{y_{ij}}}{\sum_{j=1}^{k}e^{y_{ij}}} \tag{28}$$

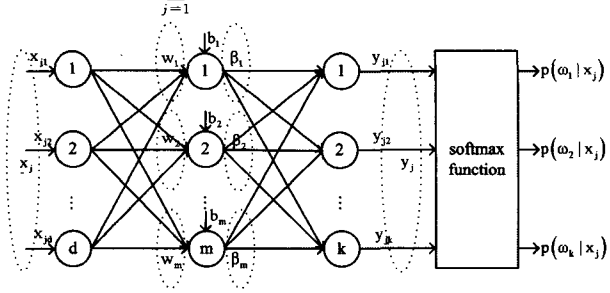

图 2　软最大化神经网络的输出

对于无类标样例集合 U 中的样例 x_i，其重要性用式(29)定义的样例熵度量。

$$H(x_i)=-\sum_{j=1}^{k}p(\omega_j|x_i)\log_2 p(\omega_j|x_i) \quad (29)$$

对于样例 $x_i\in U$，$H(x_i)$的值越大，它属于各个类别的不确定性也越大，这种样例也就越重要。本文就是用式(29)定义的样例熵作为启发式，从无类标样例集合 U 中选择样例，然后交给专家标注其类别。

在提出的算法中，用 K-近邻算法作为专家，标注选出的样例的类别。具体地，对于每一次迭代，用当前有类标的样例构成的集合 L 作为训练集，用 K-近邻算法确定选出的样例的类别。主动学习的流程如算法 3 所示。

算法 3 在线序列极限学习机主动学习算法

Input：有类标的训练集 $L=\{(x_i,y_i)|x_i\in R^d,y_i\in D,1\leqslant i\leqslant l\}$；无类别的样例集合 $U=\{x_i|x_i\in R^d,l+1\leqslant i\leqslant l+n\}$

Output：分类器 C

1. 用初始训练集 L 训练一个分类器 C；
2. while（没有达到停止条件时）do
3. for(对于每一个无类标的样例 $x\in U$)do
4. 用式(29)计算其信息熵；
5. end
6. 从集合 U 中选择信息熵大于某一个阈值的样例；
7. 将 L 作为分类器，用 K-近邻算法确定选出的样例的类别；
8. 将标注了类别的样例加入到集合 L 中；
9. 在新的训练集 L 上重新训练分类器 C；
10. end
11. 输出最终训练的分类器 C.

4 实验结果及分析

为了进一步验证本文算法的有效性，在 12 个数据库上进行了实验，其中包括 10 个 UCI 数据集和 2 个真实数据集。2 个真实数据集是 CT 和 RenRu。CT 数据集是从河北大学附属医院 212 幅脑 CT 图像经特征提取后得到的，212 幅图像中正常 CT 图像 170 幅，病变 CT 图像 42 幅。用于表示图像的特征 35 个，其中 10 个对称特征，9 个纹理特征和包括均值、方差、熵等在内的 16 个统计特征。RenRu 数据集是由河北大学智能图文实验室创建的，由 92 个汉字“人”和 56 个汉字“入”构成，每个汉字用 26 个特征进行描述。实验所用数据集的基本信息列于表 1 中。实验环境是 Intel(R) Core(TM) i5-2400 CPU @ 3.10GHz 处理器，4G 内存，32 位 Windows 操作系统，Matlab R2013a。设计了两个实验用于测试所提算法的有效性。实验 1 确定初始训练集中包含多少个样例比较合适；实验 2 与 3 种相关方法进行了实验比较。

表 1 实验所用数据集的基本信息

数据集	样例数	属性数	类别数	$\|L\|$
Banknote	1372	5	2	40
Pen	7494	16	10	250
Breast	699	11	2	80
Seed	210	8	3	50
Wine	130	13	3	25
RenRu	148	26	2	45
CT	221	36	2	25
Ionosphere	337	34	2	50
Heart	2126	22	3	90
Park	195	22	2	45
Forest	325	28	4	40
Iris	150	4	3	35

实验 1 初始训练集 L 的确定

用随机抽样的方法产生初始训练集 L，即从数据集中随机抽样 $|L|$ 个样例作为初始训练集。实验分析了 $|L|$ 对最终训练出的分类器 C 的测试精度的影响。对于每一个数据集，从原数据集中随机抽取不同个数的样例作为初始训练集，测试提出的算法对测试精度的影响。实验结果列于图 3－图 14 中，确定的合适初始训练集包含的样例数 $|L|$ 列于表 1 的最后一列。在实验 2 中，初始训练集就是按这种方法确定的。需要说明的是，在实验中隐去其他样例的类标，将其作为无类标的样例进行使用。

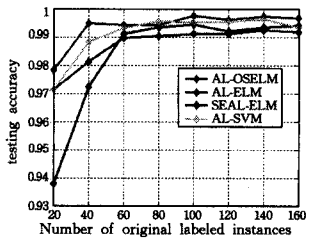

图 3 在数据集 Banknote 上的实验结果

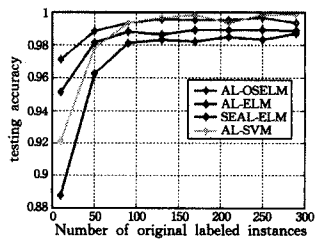

图 4 在数据集 Pen 上的实验结果

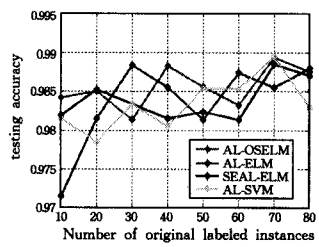

图 5 在数据集 Breast 上的实验结果

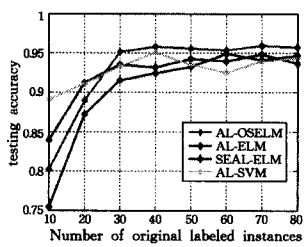

图 6 在数据集 Seed 上的实验结果

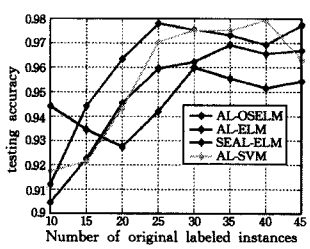

图 7 在数据集 Wine 上的实验结果

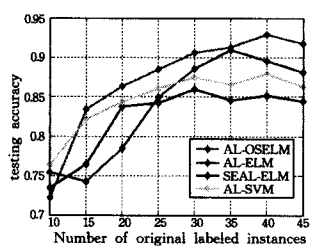

图 8 在数据集 RenRu 上的实验结果

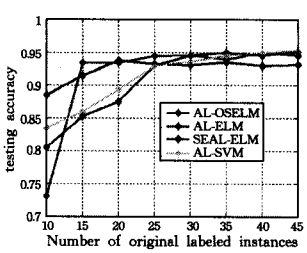

图 9 在数据集 CT 上的实验结果

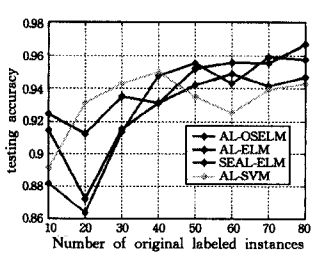

图 10 在数据集 Ionosphere 上的实验结果

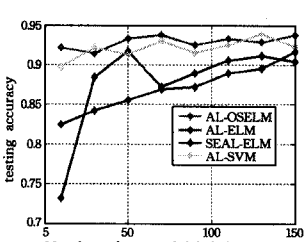

图 11 在数据集 Heart 上的实验结果

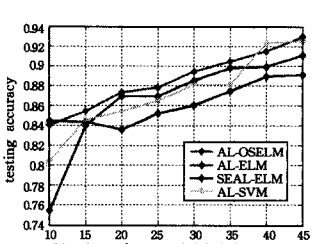

图 12 在数据集 Park 上的实验结果

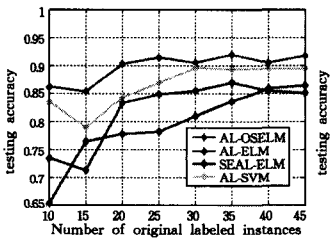

图 13　在数据集 Forest 上的实验结果

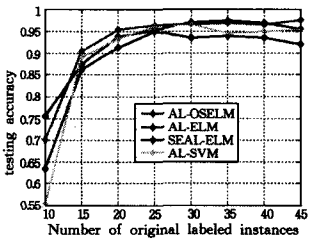

图 14　在数据集 Iris 上的实验结果

实验 2　与 3 种相关方法的比较

在该实验中，从选择的样例数和 CPU 时间两方面将所提方法与 3 种相关方法[17,18,22]进行了比较。对于提出的算法，每一次迭代，将样本熵大于某个阈值的样例选择出来进行标注，对于不同的数据集，设置的阈值是不同的，设置的原则是大于或等于平均信息熵。当没有样例可选择时，算法终止。CPU 时间比较的实验结果如表 2 所列，选择标注的样例数的实验结果如表 3 所列，需要说明的是选出的样例数不包含初始训练集中的样例。在表 2 和表 3 中，提出的算法用 AL-OSELM 表示，文献[17]中的算法用 AL-ELM 表示，文献[18]中的算法用 SEAL-ELM 表示，文献[22]中的算法用 AL-SVM 表示。从实验结果可以看出，与相关的 3 种算法相比，提出的算法选择标注的样例数基本没有变化，但所用 CPU 时间是最少的。总体来说，提出的算法具有更好的性能。

表 2　CPU 时间比较的实验结果

数据集	AL-OSELM	AL-ELM	SEAL-ELM	AL-SVM
Banknote	7.0824	16.8481	10.1401	1135.6064
Pen	29.6246	34.7882	31.2002	2845.8734
Breast	2.3244	4.7736	3.5880	151.3253
Seed	0.8112	1.2168	0.9984	65.3345
Wine	0.3900	0.7488	0.6708	38.4479
RenRu	0.4368	0.4992	0.6240	42.4949
CT	0.5928	0.8736	0.7332	47.1812
Ionosphere	1.2948	1.4976	1.4820	112.0751
Heart	12.4333	18.2053	14.0353	1822.4934
Park	0.9672	1.0140	1.0544	103.2794
Forest	1.6536	2.4024	1.7784	81.1932
Iris	0.4213	0.5620	0.7361	15.6024

表 3　选择标注的样例数的实验结果

数据集	AL-OSELM	AL-ELM	SEAL-ELM	AL-SVM
Banknote	282	253	210	271
Pen	211	265	190	223
Breast	159	174	130	168
Seed	44	43	22	39
Wine	24	35	19	26
RenRu	25	26	16	26
CT	33	26	33	31
Ionosphere	37	37	31	43
Heart	341	333	290	287
Park	43	33	30	41
Forest	102	101	80	97
Iris	23	22	27	25

结束语　提出了一种主动学习算法，该算法利用在线序列极限学习机作为分类器，将 K-近邻分类器作为 Oracle 标注选出的无类标样例的类别，将样例熵作为启发式度量样例的重要性。提出的方法具有如下 3 个特点：1)算法思想简单，易于实现；2)算法运行速度快且标注准确；3)用样例熵作为启发式，可解释性好。算法运行速度快的原因有两点：1)在线序列极限学习机算法具有增量学习的特点，在迭代过程中不需要重新训练分类器；2)极限学习机是一种快速学习算法。

参 考 文 献

[1] ANGLUIN D. Queries and concept learning [J]. Machine Learning, 1988, 2(4): 319-342.

[2] SEUNG H, OPPER M, SOMPOLINSKY H. Query by committee [C]// Proceedings of the Fifth Annual Workshop on Computational Learning Theory. 1992: 287-294.

[3] LEWIS D, GAIL W. A sequential algorithm for training text classifiers [C] // Proceedings of the 17th ACM International Conference on Research and Development in Information Retrieval. Berlin: Springer, New York, 1994: 3-12.

[4] SCHOHN G, COHN D. Less is more: active learning with support vector machines [C]// Proceedings 17th International Conference on Machine Learning. Morgan Kaufmann, San Francisco, CA, 2000: 839-846.

[5] WANG X Z, DONG L C, YAN J H. Maximum ambiguity based sample selection in fuzzy decision tree induction [J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(8): 1491-1505.

[6] TONG S, KOLLER D. Support vector machine active learning with applications to text classification [J]. The Journal of Machine Learning Research, 2002, 2: 45-66.

[7] COHN D, ATLAS L, LADNER R. Improving generalization with active learning [J]. Machine Learning, 1994, 15(2): 201-221.

[8] WANG R, KWONG S, CHEN D G. Inconsistency-based active learning for support vector machines [J]. Pattern Recognition, 2012(45): 3751-3767.

[9] LUGHOFER E. Hybrid active learning for reducing the annotation effort of operators in classification systems [J]. Pattern Recognition, 2012(45): 884-896.

[10] WANG Z, YAN S H, ZHANG C S. Active learning with adaptive regularization [J]. Pattern Recognition, 2011, 44(10/11): 2375-2383.

[11] SUN S, HARDOON D R. Active learning with extremely sparse labeled examples [J]. Neurocomputing, 2010, 73(16-18): 2980-2988.

[12] HE Y B, GENG Z. Active Learning of Causal Networks with Intervention Experiments and Optimal Designs [J]. Journal of Machine Learning Research, 2008, 9: 2523-2547.

[13] HOI S C H, JIN R, LYU M R. Batch mode active learning with applications to text categorization and image retrieval [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1233-1248.

[14] TIAN Chun-na, GAO Xin-bo, LI Jie. An example selection method for active learning based on embedded bootstrap algorithm [J]. Journal of Computer Research and Development, 2006, 43(10): 1706-1712. (in Chinese)
田春娜，高新波，李洁. 基于嵌入式 Bootstrap 的主动学习示例选择方法 [J]. 计算机研究与发展，2006，43(10)：1706-1712.

[15] GAO X B, SU Y, LI X L, et al. A review of active appearance models [J]. IEEE Transaction on System, Man, and Cybernetics Part C: Applications and Reviews, 2010, 40(2): 145-158.

（下转第 70 页）

结束语 本文首先提出了一种新的多视图聚类集成算法，克服了单一多视图 K-means 聚类受初始参数影响大的缺点，通过集成不同的聚类划分，充分利用多视图数据的互补性和一致性，提高了聚类的准确率、稳定性和鲁棒性。其次，实现的分布式多视图聚类集成算法有良好的并行性能，能够有效地对大规模多视图数据进行聚类，为下一步在大数据背景下进行聚类分析奠定了基础。

参 考 文 献

[1] KUMAR A, DAUMÉ H. A co-training approach for multi-view spectral clustering[C]// Proceedings of the 28th International Conference on Machine Learning (ICML-11). 2011: 393-400.

[2] Bickel S, Scheffer T. Multi-View Clustering [C]// ICDM. 2004: 19-26.

[3] KUMAR A, RAI P, DAUME H. Co-regularized multi-view spectral clustering[M]// Advances in Neural Information Processing Systems. 2011: 1413-1421.

[4] CAI X, NIE F, HUANG H. Multi-view k-means clustering on big data[C]// Proceedings of the Twenty-Third international Joint Conference on Artificial Intelligence. AAAI Press, 2013: 2598-2604.

[5] TZORTZIS G, LIKAS A. Kernel-based weighted multi-view clustering[C]// Proceedings of the 12th IEEE International Conference on Data Mining (ICDM). 2012: 675-684.

[6] XIIE X, SUN S. Multi-view clustering ensembles [C]// Proceedings of the IEEE 2013 International Conference on Machine Learning and Cybernetics (ICMLC). 2013: 51-56.

[7] MIZAEI H. A novel multi-view agglomerative clustering algorithm based on ensemble of partitions on different views[C]// 2010 20th International Conference on Pattern Recognition (ICPR). 2010: 1007-1010.

[8] STREHL A, GHOSH J. Cluster ensembles—a knowledge reuse framework for combining multiple partitions[J]. The Journal of Machine Learning Research, 2003, 3: 583-617.

[9] IAM-ON N, BOONGOEN T, GARRETT S. Refining pairwise similarity matrix for cluster ensemble problem with cluster relations[M] // Discovery Science. Springer Berlin Heidelberg, 2008: 222-233.

[10] DEAN J, GHEMAWAT S. MapReduce: simplified data processing on large clusters[J]. Communications of the ACM, 2008, 51 (1): 107-113.

[11] ZHAO W, MA H, HE Q. Parallel k-means clustering based on mapreduce[M]// Cloud Computing. Springer Berlin Heidelberg, 2009: 674-679.

[12] CHEN W Y, SONG Y, BAI H, et al. Parallel spectral clustering in distributed systems[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(3): 568-586.

[13] LU Wei-ming, DU Chen-yang, Wei Bao-gang, et al. Distributed affinity propagation clustering based on map reduce[J]. Journal of Computer Research and Development, 2012, 49 (8): 1762-1772. (in Chinese)
鲁伟明，杜晨阳，魏宝刚，等. 基于 MapReduce 的分布式近邻传播聚类算法[J]. 计算机研究与发展，2012，49(8)：1762-1772.

[14] ZHAO Wei-dong, MA Hui-fang, FU Yan-xiang, et al. Research on Parallel k-means Algorithm Design Based on Hadoop Platform[J]. Computer Science, 2011, 38(10): 166-168. (in Chinese)
赵卫中，马慧芳，傅燕翔，等. 基于云计算平台 Hadoop 的并行 k-means 聚类算法设计研究[J]. 计算机科学，2011，38(10)：166-168.

[15] TANG Dong-ming. Affinity propagation clustering for big data based on Hadoop[J]. Computer Engineering and Applications, 2015, 51(4): 29-34. (in Chinese)
唐东明. 基于 Hadoop 的仿射传播大数据聚类分析方法[J]. 计算机工程与应用，2015，51(4)：29-34.

[16] AMINI M R, USUNIER N, GOUTTE C. Learning from multiple partially observed views- an application to multilingual text categorization[M]// Advances in Neural Information Processing Systems (NIPS). 2009: 28-36.

[17] XIA R, PAN Y, DU L, et al. Robust multi-view spectral clustering via low-rank and sparse decomposition[C]// AAAI Conference on Artificial Intelligence. 2014: 2149-2155.

(上接第 41 页)

[16] ZHANG C S, WANG F. A multilevel approach for learning from labeled and unlabeled data on graphs [J]. Pattern Recognition, 2010, 43(6): 2301-2315.

[17] YU H, SUN C, YANG W, et al. AL-ELM: One uncertainty-based active learning algorithm using extreme learning machine [J]. Neurocomputing, 2015, 166: 140-150.

[18] YONG Z, MENG J E. Sequential active learning using meta-cognitive extreme learning machine [J]. Neurocomputing, 2016, 173: 835-844.

[19] GU Y, JIN Z, CHIU S C. Active learning combining uncertainty and diversity for multi-class image classification [J]. IET Computer Vision, 2015, 9(3): 400-407.

[20] LONG B, BIAN J, CHAPELLE O, et al. Active learning for ranking through expected loss optimization[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(5): 1180-1191.

[21] HUANG S J, JIN R, ZHOU Z H. Active Learning by Querying Informative and Representative Examples [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 36 (10): 1936-1949.

[22] 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(7): 142-160.

[23] HUANG G B, ZHU Q Y, SIEW C K. Extreme learning machine: Theory and applications [J]. Neurocomputing, 2006, 70 (1-3): 489-501.

[24] LIANG N Y, HUANG G B, SARATCHANDRAN P, et al. A fast and accurate on-line sequential learning algorithm for feedforward networks [J]. IEEE Transactions on Neural Networks, 2006, 7(6): 1411-1423.