文章编号:1001 - 9081(2011)08 - 02134 - 04
doi:10.3724/SP.J.1087.2011.02134

基于层次聚类的主动学习方法——HC_AL

贾俊芳
(山西大同大学 数学与计算机科学学院,山西 大同 037009)
(jiajunfang816@163.com)

摘　要:针对传统主动学习(AL)方法对大规模的无标记样本分类收敛速度过慢的问题,提出了基于层次聚类(HC)的主动学习训练算法——HC_AL 方法。通过对大规模的未标记数据进行层次聚类,并对每个层次上的类中心打标记来代替该层次上的类标记,然后将该层次上具有错误标记的类中心加入训练集。在数据集上的实验取得了较好的泛化能力和较快的收敛速度。实验结果表明通过采用分层细化、逐步求精的方法,可使主动学习的收敛速度大大提高,同时获得较为满意的学习能力。

关键词:主动学习;层次聚类;分层细化;逐步求精
中图分类号: TP183　　**文献标志码**:A

HC_AL: New active learning method based on hierarchical clustering

JIA Jun-fang
(*School of Mathematics and Computer Science, Shanxi Datong University, Datong Shanxi 037009, China*)

Abstract: Concerning the slow convergence speed of unlabeled samples classification while using the traditional Active Learning (AL) method to deal with the large-scale data, a Hierarchical Clustering Active Learning (HC_AL) algorithm was proposed. During operation in the algorithm, the majority of the unlabeled data were clustered hierarchically and the center of each cluster was labeled to replace the category label of this hierarchy. Then the wrong labeled data were added into the training data sets. The experimental results at the data sets show that the proposed algorithm improves the generalization ability and the convergence speed. Moreover, it can greatly improve the active learning convergence speed and obtain relatively satisfactory learning ability by using the method of hierarchical refinement and stepwise refinement.

Key words: Active Learning (AL); Hierarchical Clustering (HC); hierarchical refinement; stepwise refinement

0　引言

主动学习(Active Learning, AL)是针对具有大量未标记样本的一类通用有效的机器学习方法[1],目前已成为机器学习的研究热点,并在很多领域如文本分类、图像检索、人脸检测及障碍物检测[2-6]等得到成功的应用。主动学习的本质是通过对较少的样本进行标记来达到对大规模样本进行自动分类的目的。对于这类问题,已存在一些行之有效的算法,如支持向量机(Support Vector Machine, SVM)主动学习(Active Learning)算法[7]、QBBAG(Query-by-Bagging)算法[8]、Query-by-Boosting 算法[9]以及委员会投票选择(Query by Committee, QBC)[10]算法等。对于样本不太多的情况,这些方法一般是很有效的,但在实际应用中往往会遇到许多大规模数据集,而大多数经典算法由于是对每个样本依次进行人工标记来修正超平面,收敛速度往往较慢,有时根本无法处理这种大数据量的问题。

主动学习是一个从备选的大规模无标记样本集中循环选择样本进行人工标记的过程。首先,候选的样本集中的所有样本均没有标记,根据一些先验知识或者随机地选择一部分样本作为初始样本交由专家进行标记,然后利用这些已标记的样本来构造初始分类器,并利用某种启发式规则,从剩余的无标记候选样本集中选择最有利于提高分类器性能的样本,再进行人工标记并放入训练集当中,重新训练分类器。如此循环往复,直到得到的分类器对于训练集的分类效果达到预计目标为止。主动学习方法的关键步骤就是设计样本抽取的启发式规则,因为无标记样本的抽取直接决定着主动学习器最终所能达到的性能和达到该性能所需要的学习速率。

目前,关于主动学习已经有一些研究,并提出了一些相应的方法。具体地讲,基本上分两种:1)基于投票的方法[11]。其主要思想是首先使用多个分类器对样本就行分类,若不同的分类器对都认为样本属于某一类,则该样本所含有的信息量就小;反之,若某个样本被不同的分类器判断为属于不同类别,则该样本含有的信息量往往较大,需要进行人工标注或者做其他进一步的判断。代表的算法如 QBC 和 CBS(Committee Sampling)[12-13]。2)基于置信度的方法[14-15]。通过定义一个合适的指标如通过后验概率、期望熵等衡量样本所含的信息量,以选出不确定样本,进行人工标注。目前常用的包括选择性采样法、不确定性采样法等。但是,这些方法都有一个共有的缺点,即收敛速度较慢,不能够快速地迭代得到分类结果。

本文针对传统主动学习方法分类精度低、收敛速度慢的问题,提出了一种基于层次聚类的主动学习(Hierarchical Clustering Active Learning, HC_AL)方法,通过层次聚类模型对数据进行聚类预处理,然后以类中心标记代替该层次上的类标记,从而将该层次上具有错误标记的类中心加入训练集。采用这种分层细化、逐步求精的方法既提高学习器的学习效

收稿日期:2011-01-10;**修回日期**:2011-03-14。　**基金项目**:山西省青年科学基金资助项目(2011021013-2)。
作者简介:贾俊芳(1976-),女,山西左云人,讲师,主要研究方向:人工智能、机器学习。

率,又获得令人满意的泛化能力。

1　基于层次聚类的主动学习方法

1.1　层次聚类模型

层次聚类(Hierarchical Clustering, HC)的方法是将数据对象在不同的层次阶段上按照一定的规则聚为不同规模的类,并在类的不断分解或合并过程当中改善聚类效果,从而达到逐步细化、完善求精的目的[16]。常见的层次聚类模型可分为聚合型和分解型。聚合型层次聚类采用的是自底向上的方法,即将低层的规模较小的类逐渐合并为高层规模较大的类;而分解型则是采用自顶向下的方法,即将高层的规模较大的类按照一定的规则逐渐分解成低层规模较小的类。本文采用的是分解型层次聚类方法。

1.2　基于 HC 的主动学习模型

基于层次聚类的主动学习模型打破了传统主动学习方法中将初始近似分类面附近的无标记数据点当成含有信息较多的点进行更新的思想,而是通过层次聚类,逐层将含有信息量较大的类代表性点选出来加入训练集参与训练,在每一个聚类层次上,都至少存在一个与父类不属于同一类的子类加入到训练集,以在比父类更细的层次上对超平面进行改进,如图 1 所示。A 为整个数据集,B、C 和 D 均为第一次聚类后产生的类,简单地直接采用中心打标记的方法进行标记,其中 B、D 为正类,C 为负类,此时直接进行划分的话得到的超平面是 $f^{(j)}$。但是,当进行第二次更细层次的聚类后,类 B、C 和 D 如该图右侧所示,其中 B 得到的子类依然都是正类,但是 C 和 D 得到的子类都是既有正类又有负类,此时进行划分时得到的划分超平面是 $f^{(j+1)}$,显然 $f^{(j+1)}$ 能够更逼近于最优分类超平面。而且由于是对子类整体进行操作,所以 $f^{(j+1)}$ 相对于 $f^{(j)}$ 有了较大的改进。这样,采用这种层次聚类的思想,每次都至少有一个父类划分为多个子类,超平面也能在相应的层次上进行更新,以更快的速率趋于最优分类面。因此,与传统方法相比,该方法具有更快的收敛速度及更优秀的泛化能力。

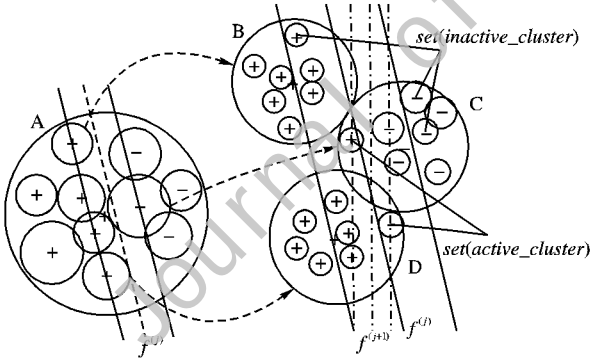


图 1　基于 HC 的主动学习模型示意图

1.3　基于 HC 的主动学习算法

为方便描述,设 l 为输入的无标记样本 X 的个数,r 为样本的特征数,则第 i 个样本可以表示为:$x_i = (x_i^1, x_i^2, \cdots, x_i^r)$,其中 $x_i \in X$。设 n_j 为第 j 层的聚类过程中每个原始的类所要划分的新的子类的个数,其中 $j \geqslant 0$(为了简化模型,本文算法假设第 j 层的聚类过程中,每个原始类所划分的新的子类的个数一致,均为 n_j)。初始状态下,所有数据 X 为一类,其第一层聚类过程中得到的子类个数为 n_0,不妨设得到的子类为 $X_1^{(0)}, \cdots, X_{n_0}^{(0)}$,则初始的聚类过程等价于寻找这样的一个聚类的映射:$X \to \{X_0^{(0)}, \cdots, X_{n_0}^{(0)}\}$,类似可得,一般地,第 j 层聚类过程中的第 i 个原始子类的聚类等价于寻找一个聚类映射:$X_i^{(j-1)} \to \{X_{i,1}^{(j)}, \cdots, X_{i,n_j}^{(j)}\}$,并设 $\mu_i^{(j)}$ 为类 $X_i^{(j)}$ 的中心。设置每次需要继续划分的类集合为活跃集,记为 $set(active\_cluster)$,而需要进行继续划分的类集为不活跃集,记为 $set(inactive\_cluster)$,记实际训练集为 $set(train)$。

1.3.1　计算初始超平面 $f^{(0)}$

计算初始分类超平面 $f^{(0)}$ 的算法的主要步骤如下。

1) 初始参数设置。给定无标记的训练集 X,初始化 HC_AL 的初始聚类个数参数。选择分类器 F 的相关参数。活跃集与不活跃集的初始设置为 $set(active\_cluster) = \{X\}$,$set(inactive\_cluster) = \varnothing$,实际训练集为 $set(train) = \varnothing$。

2) 用聚类方法得到初始的超平面 $f^{(0)}$。

a) 选择一种合适的聚类方法对活跃集进行聚类,即得到一个聚类映射:

$$X \to \{X_0^{(0)}, \cdots, X_{n_0}^{(0)}\} \tag{1}$$

b) 更新活跃集,$set(active) = \{X_0^{(0)}, \cdots, X_{n_0}^{(0)}\}$。

c) 根据式(2)计算每个子类的类中心,如第 i 个子类 $X_i^{(0)}$ 的类中心为 $u_i^{(0)}$,其中 $u_i^{(0)}$ 的第 p 个特征 $u_i^{(0),p}$ 为:

$$u_i^{(j),p} = \frac{\sum_{k=1}^{n_0} x_k^p}{|X_i^{(j)}|} \tag{2}$$

d) 根据第 i 类中的任意第 k 个数据点 $x_{i,k}^{(j)}$ 到类中心 $u_i^{(j)}$ 的距离(式(3))以及选取代表性点的式(4)来选择每个类中具有代表性的点 $x_{i,m}^{(0)}$(为简化模型,本文取距离类中心最近的点,由于有 n_0 个类,所以有 n_0 个点),交由专家打出标记 $y_i^{(0)}$,$i = 1, \cdots, n_0$。

$$d(x_{i,k}^{(j)}, u_i^{(j)}) = \sqrt{\sum_{p=1}^{r} (x_{i,k}^{(j),p} - u_i^{(j),p})^2};\ k = 1, \cdots, |X_i^{(j)}| \tag{3}$$

从 $X_i^{(0)}$ 中选择 $x_{i,m}^{(0)}$,且使

$$d(x_{i,m}^{(0)}, u_i^{(0)}) \leqslant d(x_{i,k}^{(0)}, u_i^{(0)});\ \forall k \in \{1, \cdots, |X_i^{(0)}|\} \tag{4}$$

e) 更新实际训练集,即:

$$set(train) = \{x_{i,m}^{(0)}\} \bigcup \{y_i^{(0)}\};\ i = 1, \cdots, n_0 \tag{5}$$

采用一种分类器 F 分类,得到初始分类面 $f^{(0)}$。

3) 算法结束。

1.3.2　基于层次聚类的主动学习过程

基于层次聚类的主动学习过程如下。

1) 首先设置最大层次聚类迭代次数 $epoch$ 以及初始分类面 $f^{(0)}$ 和 $set(active\_cluster)$、$set(inactive\_cluster)$ 以及 $set(train)$。

2) 对活跃集 $set(active\_cluster)$ 中的每个类 $X_i^{(j)}$ 再进行聚类,从而得到一系列聚类映射:

$$X_i^{(j)} \to \{X_{i,1}^{(j+1)}, \cdots, X_{i,n_{j+1}}^{(j+1)}\};\ i = 1, \cdots, |\ set(active\_cluster)\ | \tag{6}$$

聚类参数 n_{j+1} 通过式(7)设置:

$$n_{j+1} = \sigma(\min(|X_i^{(j)}|));\ X_i^{(j)} \in set(active\_cluster) \tag{7}$$

其中 σ 为层次聚类步幅参数。

3) 依次重置不活跃集与活跃集。其中:

$$set(inactive\_cluster) = \{X_{i,k}^{(j+1)}\};$$
$$i = 1, \cdots, |\ set(active\_cluster)\ |,\ k = 1, \cdots, n_{j+1} \tag{8}$$

而后重置 $set(active\_cluster) = \varnothing$。

4）根据式(2)求不活跃集 $set(inactive\_cluster)$ 中的每个元素 $X_{i,k}^{(j+1)}$ 的中心点 $\mu_{i,k}^{(j+1)}$，根据式(3)求相应的类中距离这些类中心 $\mu_{i,k}^{(j+1)}$ 最近的点 $x_{i,k,m}^{(j+1)}$，然后将这些点交给专家打出标记 $y_{i,k}^{(j+1)}$。

5）更新训练集，即：

$$set(train) = set(train) \bigcup \{x_{i,k,m}^{(j+1)}\} \bigcup \{y_{i,k}^{(j+1)}\} \tag{9}$$

采用分类器 F 进行分类，得到新的分类面 $f^{(j+1)}$。

6）更新活跃集与不活跃集。对于 $set(inactive\_cluster)$ 中的所有集合中对应的 $y_{i,k}^{(j+1)}$ 进行判断，如果 $y_{i,k}^{(j+1)} \neq y_i^{(j)}$，则执行如下集合运算：

$$set(active\_cluster) = set(active\_cluster) \bigcup \{X_{i,k}^{(j+1)}\} \tag{10}$$

$$set(inactive\_cluster) = set(inactive\_cluster) - \{X_{i,k}^{(j+1)}\} \tag{11}$$

7）判断是否达到结束条件，即判断是否达到最大迭代次数 $epoch$ 或者是否存在：

$$\min(|X_i^{(j)}|) = 1;\ X_i^{(j)} \in set(active\_cluster) \tag{12}$$

若达到，则算法结束，得到分类面 $f^{(j+1)}$；否则返回第2）步继续迭代执行。

2 实验及结果分析

2.1 实验数据集

为验证 HC_AL 算法的效率，本文在 5 个标准的 UCI 数据集(见表 1)上进行测试。实验中采用 SVM 来作分类器，采用高斯核函数，核参数 σ 取 1.0，正则参数 C 取 1000。

表 1 实验采用的数据集

数据集	训练集个数	测试集个数	数据维数
Banana	8 800	1 000	2
Breast_cancer	2 000	770	9
Titanic	750	10 255	3
Thyroid	2 800	1 500	5
Flare_solar	33 300	20 000	9

2.2 结果及分析

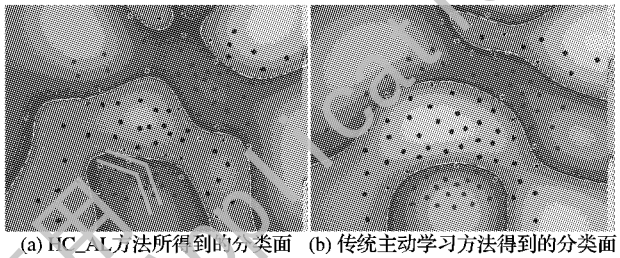
本文首先对 HC_AL 算法的有效性进行测试。实验在 1 台 PC 机(2.66 GHz CPU，1 GB 内存)上进行测试，实验平台是 Matlab 7.0。实验中，影响算法表现的主要就是层次聚类参数 n_i 的设定。为此，在 Banana 数据集上对层次聚类参数 n_i 对算法的影响进行了分析，表 2 为 HC_AL 算法在不同参数下的实验结果比较。为了简化测试，实验中采用两层聚类，即：n_1 表示第一层聚类的个数，n_2 表示第二层聚类的个数。表中带下划线的数值表示当前聚类层次下所得到的精度的最大值。

表 2 不同参数下 HC_AL 算法对 Banana 数据集的测试精度 %

第二层聚类的个数 n_2	第一层聚类的个数 n_1				
	10	20	30	40	50
0	0.78	0.79	<u>0.85</u>	0.85	0.85
5	0.83	0.79	0.85	0.85	<u>0.87</u>
10	0.83	0.85	0.88	0.87	<u>0.90</u>
15	0.85	0.87	0.87	0.90	<u>0.93</u>
20	0.88	0.93	0.90	<u>0.93</u>	0.93

由表 2 可以看出，在每个层次的聚类当中，当聚类个数增多的时候，测试精度增高；当聚类的层次由一层变为两层时，所得到的精度最大值增加。这个结果表明，本文提到的分层聚类算法在主动学习过程中是非常有用的，能够在不同的聚类次上尽可能快地提取到含有更多信息的无标记样本来交给专家标记，通过在不同层次上含有较多信息的子类提取，加快了分类面更新的步幅，从而提高了算法收敛的速度。此外，由于所得到的最终训练集规模较小，因此也有效地提高了算法的执行效率。

图 2(a)为采用 HC_AL 算法得到的分类面，而图 2(b)为采用传统主动学习方法得到的分类面(其数据集规模与 HC_AL 的最终实际训练的数据集 $set(train)$ 规模一致)。可以看到，采用 HC_AL 提取到了更多的分类面附近的含有较多分类信息的样本来参与分类，而传统主动学习方法则明显丢失了很多分类面附近的有用分类信息(其中 $n_1 = 30, n_2 = 15$)。



(a) HC_AL方法所得到的分类面 (b) 传统主动学习方法得到的分类面

图 2 两种方法对 Banana 数据的分类面

由实验结果可以看出，本文提出的 HC_AL 方法对传统的主动学习方法有了不少改进，算法的收敛速率和泛化能力都有了较大提高。图 3～6 为在其他数据集下 HC_AL 方法与同等规模数据采用传统主动学习方法所得到的实验结果对比，其中横轴为两种算法实际参与训练的数据集规模，纵轴为训练测试精度。这些实验结果同样表明本文提出的算法与传统主动学习方法相比，收敛速率和泛化能力有了较大的提高。

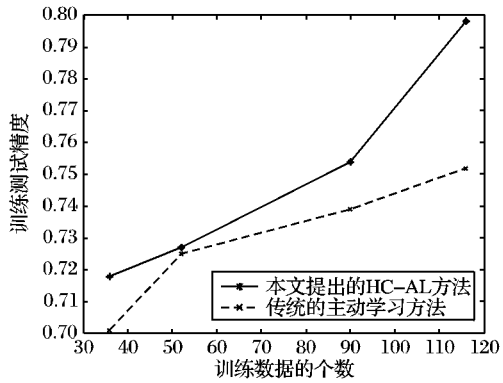


图 3 两种方法在 Breast_cancer 上的实验结果

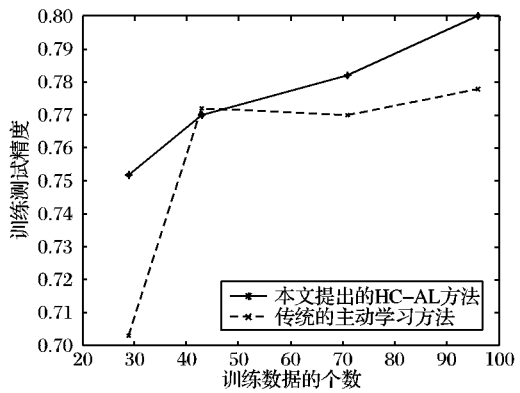


图 4 两种方法在 Titanic 上的实验结果

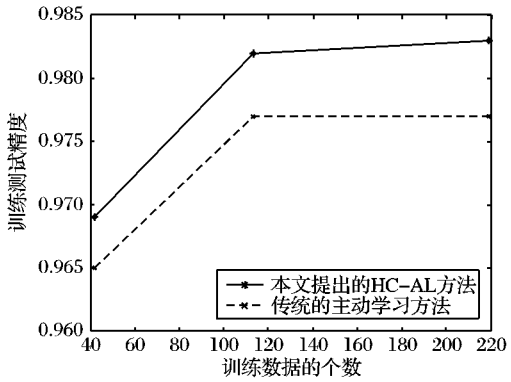


图 5　两种方法在 Thyroid 上的实验结果

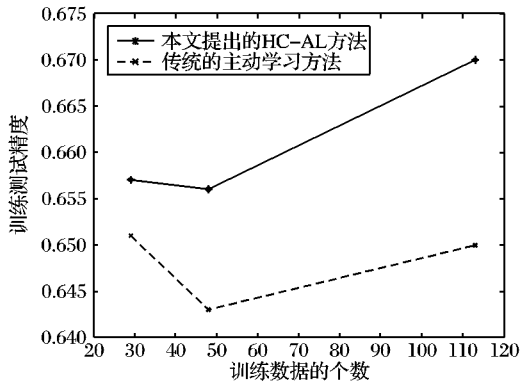


图 6　两种方法在 Flare_solar 上的实验结果

3　结语

主动学习方法由于其训练规模小、速度快的优点,在很多领域得到成功的应用,已经成为机器学习领域的一个研究热点。本文提出的 HC_AL 方法采用分层细化、逐步求精的方法,既提高学习器的学习效率,又获得满意的泛化能力。在未来的工作中,可以考虑将本文提出的 HC_AL 主动学习方法与不同的分类器如神经网络、决策树等相结合,以取得更令人满意的结果,从而使主动学习的性能得到进一步的提高。

参考文献:

[1] DIMA C, HEBERT M, STENTZ A. Enabling learning from large datasets: Applying active learning to mobile robotics [C]// ICRA 2004: Proceedings of the 2004 IEEE International Conference on Robotics and Automation. Piscataway, NJ: IEEE, 2004: 108 - 114.

[2] VLACHOS A. A stopping criterion for active learning [J]. Computer Speech and Language, 2008, 22(3): 295 - 312.

[3] 张健沛,徐华. 支持向量机(SVM)主动学习方法研究与应用[J]. 计算机应用,2004,24(1):1 - 3.

[4] CORD M, COSSELIN P H, PHILIPP-FOLIGUET S. Stochastic exploration and active learning for image retrieval [J]. Image and Vision Computing, 2007, 25(1): 14 - 23.

[5] 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.

[6] 韩光,赵春霞,胡雪蕾. 一种新的 SVM 主动学习算法及其在障碍物检测中的应用[J]. 计算机研究与发展,2009,46(11):1934 - 1941.

[7] TONG S, KOLLER D. Support vector machine active learning with applications to text classification [J]. Journal of Machine Learning Research, 2002, 2(1): 45 - 66.

[8] ABE N, MAMITSUKA H. Query learning strategies using boosting and bagging [C]// Proceedings of the 15th International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann Publishers, 1998: 1 - 9.

[9] BAEZA-YATES R, HURTADO C, MENDOZA M. Query clustering for boosting Web page ranking [C]// AWIC 2004: Proceedings of the Second International Atlantic Web Intelligence Conference, LNCS 3034. Berlin: Springer-Verlag, 2004, 3034: 164 - 175.

[10] FINE S, GILAD-BACHRACH R, SHAMIR E. Query by committee, linear separation and random walks [J]. Theoretical Computer Science, 2002, 284(1): 25 - 51.

[11] TONG JUN, YIN JIANPING, ZHU EN. An active learning method based on most possible misclassification sampling using committee [C]// MDAI'07: Proceedings of the 4th International Conference on Modeling Decisions for Artificial Intelligence. Berlin: Springer-Verlag, 2007: 104 - 113.

[12] ROVER B, JEM J R, ROSS D K. Active learning for regression based on query by committee [C]// IDEAL'07: Proceedings of the 8th International Conference on Intelligent Data Engineering and Automated Learning, LNCS 4881. Berlin: Springer-Verlag, 2007: 209 - 218.

[13] ZHU JINGBO, WANG HUIZHEN, TSOU B K, *et al*. Active learning with sampling by uncertainty and density for data annotations [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 18(6): 1323 - 1331.

[14] 陈锦禾,沈洁. 基于信息熵的主动学习半监督分类研究[J]. 计算机技术与发展,2010,20(2):110 - 113.

[15] 龙军,殷建平,祝恩,等. 选取最大可能预测错误样例的主动学习算法[J]. 计算机研究与发展,2008,45(3):472 - 478.

[16] YU H, YANG J, HAN J W, *et al*. Making SVMs scalable to large data sets using hierarchical cluster indexing [J]. Data Mining and Knowledge Discovery, 2005, 11(3): 100 - 128.

(上接第 2133 页)

[8] 柳位平,朱艳辉,栗春亮,等. 中文基础情感词词典构建方法研究[J]. 计算机应用,2009,29(10):2875 - 2877.

[9] ESULI A, SEBASTIANI F. Sentiwordnet: A publicly available lexical resource for opinion mining [C]// Proceedings of LREC-06, the 5th Conference on Language Resources and Evaluation. Genova, Italy: [s. n.], 2006: 417 - 422.

[10] 姚天昉,娄德成. 汉语情感词语义倾向判别的研究[C]//ICCC2007:第七届中文信息处理国际会议论文集. 北京:电子工业出版社,2007:221 - 225.

[11] HATZIVASSILOGLOU V, MCKEOWN K. Predicting the semantic orientation of adjectives [C]// ACL-97: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics. Madrid, Spain: [s. n.], 1997:174 - 181.

[12] TURNEY P. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews [C]// ACL-02: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, PA: [s. n.], 2002: 417 - 424.

[13] 蔺璜,郭姝慧. 程度副词的特点范围与分类[J]. 山西大学学报:哲学社会科学版,2003,26(2):71 - 74.

[14] 林炜,林世平. 中文倾向性挖掘中情感词修饰极性的研究[J]. 计算机科学,2008,35(8):208 - 210.

[15] 廖祥文. 基于博主背景的博客倾向性检索归一化策略[J]. 中文信息学报,2010,24(3):75 - 80.

[16] 应伟,王正欧,安金龙. 一种基于改进的支持向量机的多类文本分类方法[J]. 计算机工程,2006,32(16):74 - 76.