计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
49(6): 1306-1312, 2012

一种利用近邻和信息熵的主动文本标注方法

朱 岩 景丽萍 于 剑
(北京交通大学计算机科学系 北京 100044)
(yasmine_zhu@hotmail.com)

An Active Labeling Method for Text Data Based on Nearest Neighbor and Information Entropy

Zhu Yan, Jing Liping, and Yu Jian
(*Department of Computer Science, Beijing Jiaotong University, Beijing* 100044)

Abstract As it is quite time-consuming to label text documents on a large scale, a kind of text classification with a few labeled data is needed. Thus, semi-supervised text classification emerges and develops rapidly. Different from traditional classification, semi-supervised text classification only requires a small set of labeled data and a large set of unlabeled data to train a classifier. The small set of labeled data is used to initialize the classification model in most cases. Its rationality will affect the performance of the final classifier. In order to make the distribution of the labeled data more consistent with the distribution of the original data, a sampling method is proposed to avoid selecting the K nearest neighbors of the labeled data to be new candidate labeled data. With the help of this method, the data located in various regions will have more opportunities to be labeled. Moreover, in order to obtain more category information from the very few labeled data, this method compares the information entropy of the candidate labeled data and the datum with the highest information entropy is chosen as the next datum to be labeled manually. Experiments on real text data sets suggest that this approach is very effective.

Key words semi-supervised text classification; active learning; nearest neighbor; information entropy; labeling strategy

摘 要 由于大规模标注文本数据费时费力,利用少量标注样本和大量未标注样本的半监督文本分类发展迅速.在半监督文本分类中,少量标注样本主要用来初始化分类模型,其合理性将影响最终分类模型的性能.为了使标注样本尽可能吻合原始数据的分布,提出一种避开选择已标注样本的 K 近邻来抽取下一组候选标注样本的方法,使得分布在不同区域的样本有更多的标注机会.在此基础上,为了获得更多的类别信息,在候选标注样本中选择信息熵最大的样本作为最终的标注样本.真实文本数据上的实验表明了提出方法的有效性.

关键词 半监督文本分类;主动学习;近邻;信息熵;标注方法

中图法分类号 TP18

收稿日期:2011-03-11;修回日期:2011-11-16
基金项目:中央高校基金科研业务费专项资金项目(2009YJS026);北京交通大学优秀博士生科技创新基金项目(141097522);国家自然科学基金项目(0905028,90820013,60875031)

随着计算机技术的不断发展,越来越多的数据以电子文档的形式加以存储,因此,有效地对这些海量数据进行组织管理成为一项紧迫的任务.作为一种组织和管理海量文档的手段,文本分类技术得到了前所未有的关注.传统的文本分类又称为有监督的文本分类,它的分类过程一般分为学习和预测2个阶段.在学习阶段,分类器从大量已标注类别的训练文本中学习一个分类模型;在预测阶段,分类器利用学到的分类模型对主题未知的文档进行类别判断.然而,传统的分类技术需要大量有类别标注的训练样本,在很多现实应用中,人们并没有如此多的训练样本,或者得到大量的训练样本会耗费太多的人力物力,因此,越来越多的研究者将目光投向了半监督文本分类.

与有监督分类技术不同,半监督文本分类并不对人为提供的监督信息量有严格的要求.在很多情况下,只要提供非常少量的标注样本和大量未标注样本,半监督学习就可以学到一个很好的分类模型.目前,半监督学习方法主要包括生成模型式的方法,如基于期望最大化的贝叶斯方法[1];基于类标传播的方法,如GRF方法[2],Consistency方法[3]和LNP方法[4];基于低密度分离的方法,如TSVM[5],Laplacian SVM[6]和MeanS3VM[7];协同训练的方法,如Co-training[8-9],Tri-training[10]和集成方法[11]等等.

除了对半监督算法进行研究外,主动选取合适的样本进行人工标注也是提高分类器准确度的一个有效手段.目前,对数据进行主动标注的方法包括:与分类器交互的增量标注方法、独立于分类器的批量标注方法.与分类器交互的增量标注方法主要出现在主动学习算法中.比如基于协同训练的主动学习把数据分成若干视图,在不同视图上分别建立分类器,然后选择预测差别最大的样本进行人工标注[12-14];还有一些主动学习方法考虑数据的分布信息,选择当前分类器边界或者中心的样本进行标注[15-16].然而,上述方法要求分类器的每次运行都有领域专家的参与,增加了领域专家的负担.此外,由于选取的标注样本依赖于所用的分类器,样本标注质量也受分类器性能的影响.与增量标注方法不同,独立于分类器的批量标注方法将人工标注和训练分类器独立起来,领域专家可以在分类器运行之前选择样本进行标注,其标注质量不受具体分类器的影响.尽管独立于分类器的批量标注方法有很大的优势,其相关研究并不多见,常用的方法就是随机标注和基于FFT(farthest-first traversal)的策略[17-18].FFT策略每次选取与已标注样本最小距离最大的样本,从而避免标注相似的样本.然而FFT并没有充分考虑样本的分布,也易受野值点的影响.

鉴于此,本文提出一种基于近邻和信息熵的文本数据类别标注方法.这种标注方法避免选取已标注样本的K近邻作为新一轮的候选标注样本,使得分布在不同区域的样本有更多的标注机会.在此基础上,该方法比较候选标注样本的信息熵,找出信息熵最大的样本进行标注.

1 基于近邻和信息熵的主动文本标注方法

对半监督学习来说,分类模型的计算主要依赖于少量标注了类别的样本和大量无类别标注的普通数据.其中,少量标注了类别的样本主要用于分类模型的初始化.其合理性将影响最终分类模型的性能.在半监督分类算法固定的情况下,尽可能按数据分布抽取及标注样本,尽可能标注信息量大的样本是提高初始分类模型准确度的一个有效手段.因此,本文提出一种独立于分类器的主动样本标注方法.该方法避免选择已标注样本的K近邻作为新一轮的候选标注样本,使得分布在不同区域的样本有更多的标注机会.此外,该方法比较候选标注样本的信息熵,选择信息量最大的样本为最终标注的样本.

1.1 利用近邻信息抽取候选标注样本

文本数据的一大特点就是高维性.对于高维数据来说,尽可能按原始数据的分布成比例地抽取样本进行人工标注并不容易.本节通过避免选择已标注样本的K近邻作为新一轮的候选标注样本,使得分布在不同区域的样本有更多的标注机会.

假设$X=\{x_1,x_2,\cdots,x_N\}$表示包含N个样本的文档集合.x_i的K近邻定义为与x_i相似度最大的K个文档.为了使标注的数据与原始数据的分布尽可能一致,当标注样本x_i后,x_i和x_i的K近邻失去成为候选标注样本的机会.新的候选标注样本将从其他未标注样本中随机产生.以一个简单的人工数据S_1为例(如图1所示),S_1包含12个样本点,分布在3个类中.如果有3次标注机会,且定义近邻数为3,基于近邻信息的标注过程可以分为以下3步:1)随机选取1个样本进行标注,该样本的3近邻失去标注机会;2)从其余8个样本中随机选取1个

样本进行标注,新标注样本的3近邻也失去了标注机会;3)从其余4个样本中随机选取1个样本进行标注.这样标注的3个样本各自代表1个类,没有出现标注样本分布不均匀的现象.

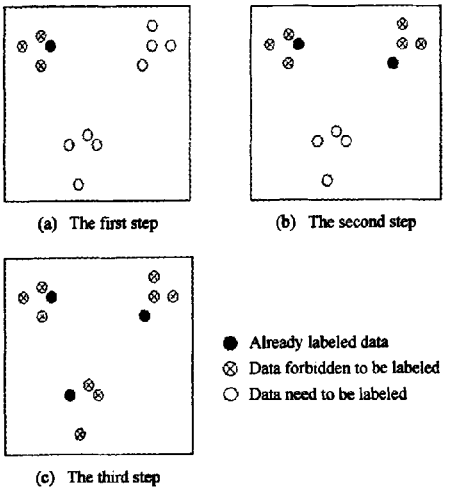

Fig. 1 Labeling process for data S_1.

图1 数据 S_1 的标注过程

真实数据往往比 S_1 复杂,但由于主动避免选择和已标注数据非常相似的样本,即使对于复杂数据,该方法也会在一定程度上减少随机标注可能出现的小范围标注稠密的情况,使得分布在数据不同区域的样本有更多的标注机会.

1.2 选择信息量大的样本

除了让标注样本与原始数据分布尽可能吻合外,我们也希望标注的样本包含比较大的信息量.因此,本节在1.1节的基础上,将新一轮抽取的样本和其 K 近邻都作为候选标注样本,然后从中选取信息量最大的样本作为该轮最终标注的样本.度量信息量的一个常用指标就是信息熵.为了求出每篇文档的信息熵,定义样本 x_i 的文档向量为 $[x_{i1}, x_{i2}, \cdots, x_{iM}]$,$x_{im}$ 表示文档 x_i 中单词 t_m 的权重.将文档向量归一化,使得该向量各分量之和为1,那么归一化后得到的向量 $[x'_{i1}, x'_{i2}, \cdots, x'_{iM}]$ 可以反映出文档 x_i 中各个单词出现的概率,进而文档 x_i 的信息熵可以定义为

$$H(x_i) = -\sum_{m=1}^{M} x'_{im} \times lbx'_{im}. \quad (1)$$

单词出现概率的不确定性越高,样本 x_i 的信息熵越大,标注 x_i 所获得的信息就越多.根据文本数据中主要出现文本所属类别的单词和一些噪音单词的特点,我们人工模拟了一个简单的文本数据 S_2,如表1所示.S_2 包含4篇文档,文档 x_1 和 x_2 属于第1类,其代表单词为 t_1,t_2 和 t_3,文档 x_3 和 x_4 属于第2类,其代表单词为 t_4,t_5 和 t_6,单词 t_7 为各个类中都出现的噪音单词:

Table 1 Synthetic Text Data S_2

表1 人造文本数据 S_2

Document	Word						
	t_1	t_2	t_3	t_4	t_5	t_6	t_7
x_1	5	4	3	0	0	0	2
x_2	0	6	4	0	0	0	2
x_3	0	0	0	3	6	7	2
x_4	0	0	0	0	5	0	2

计算这些文档的信息熵得出 $H(x_1)=1.92$,$H(x_2)=1.46$,$H(x_3)=1.84$,$H(x_4)=0.86$.对于第1类,样本 x_1 比 x_2 的信息量大,如果标注 x_1,相当于认为 t_1,t_2,t_3 和 t_7 以一定的概率属于第1类,如果标记 x_2,相当于认为 t_2,t_3 和 t_7 以一定的概率属于第1类,漏掉了 t_1 的信息.同理,对于第2类,样本 x_3 比 x_4 的信息量大.如果标记 x_4,而不是 x_3,就失去 t_4 和 t_6 属于第2类的信息.

1.3 算法流程

本文提出的方法使标注样本与原始数据的分布尽量一致,且包含较大的信息量,对应的算法流程如下:

输入:文本集合 X、样本的近邻数 K 和需要标注的文档数 L.

输出:需要标注的文档集合 S.

Step1. 令 S 为空,各样本的 $flag$ 置为0.

Step2. 从 X 中随机选取1个样本,并确定其 K 近邻,将这 $K+1$ 个样本的 $flag$ 置为1.

Step3. 计算这 $K+1$ 个样本的信息熵,将信息熵最大的样本加入到 S 中.

Step4. $count=1$.

Step5. While $count<L$ 且有 $flag$ 为0的样本

Step5.1. 从 $flag$ 为0的样本中随机选取1个样本 x.

Step5.2. 找出 x 的 K 近邻,并将 $flag$ 为1的样本从 K 近邻中删除,x 和其 K 近邻中剩余的样本形成集合 R.

Step5.3. 将 R 中文档的 $flag$ 置为1.

Step5.4. 比较 R 中文档的信息熵,将信息熵最大的样本加入到 S 中.

Step5.5. $count=count+1$.

End

Step6. While $count<L$ 且没有 $flag$ 为 0 的样本

Step6.1. 计算已标注样本的 $K-1$ 近邻,但不是 K 近邻的样本,这样的样本形成集合 Z.

Step6.2. 从 Z 中随机抽取 $\min(L-count, |Z|)$ 个样本加入到 S 中.

Step6.3. $count=count+\min(L-count, |Z|)$.

Step6.4. $K=K-1$.

End

其中,算法的步骤 2、步骤 3 是找到第 1 个标注样本,该样本既是随机选取的,又满足信息量比其 K 近邻的信息量大.算法的步骤 5 是在已有若干标注样本的条件下,选择与已标注样本为非 K 近邻,且信息量大的样本.算法的步骤 6 是在所有样本都成为已标注样本的 K 近邻,但还需标注更多样本时,对一些样本解除标注禁止的情况.

1.4 时间复杂度

在最坏的情况下,本文提出的方法需要计算 L 个样本的 K 近邻.计算 L 个样本的 K 近邻包括:1)计算 L 个样本和所有数据之间的 cosine 相似度;2)分别为 L 个样本找出相似度最大的 K 个样本.由于文本数据经过了 cosine 归一化,数据 cosine 相似度的计算复杂度为 $O(NML)$.分别为 L 个样本找出相似度最大的 K 个样本的计算复杂度为 $O(NKL)$.此外,本文提出的方法还需要计算最多 N 个样本的信息熵,对应的计算复杂度为 $O(NM)$.因此,方法总的计算复杂度为 $O(N(ML+KL+M))$.值得一提的是,由于文本数据具有稀疏性,实际计算相似度矩阵和信息熵的时间复杂度为 $O(NML)\times p$ 和 $O(NM)\times p$,p 为数据矩阵中的非零元素与数据矩阵中所有元素的比例.因此实际的计算复杂度为 $O(N(MLp+KL+Mp))$.本文用到的文本数据中,p 的取值在 0.003 8~0.006 6 之间.因此本文提出的标注方法可以直接应用在一般规模的文本数据上.对于超大规模的文本数据,可以先从原始数据中抽取一个中等规模的数据子集,然后在数据子集上应用本文提出的标注方法.

2 实验结果

2.1 实验数据

5 组真实的文本数据被用来评估本文提出的标注方法,如表 2 所示:

Table 2 Text Corpora

表 2 文本数据集

Data	Training/Testing	Number of Words	Number of Categories
20N	5 562/3 770	13 558	20
Reuters10	7 193/2 787	10 509	10
Ohsumed	5 584/5 578	11 101	10
Sports	4 291/4 289	21 560	7
Reviews	2 036/2 033	28 536	5

20N 是 20Newsgroup 数据的一个子集,作为预处理,其标题和冗余文章已被删去,为了加快实验速度,我们随机选取 50%的数据,其中 5 562 篇文档作为训练集,3 770 篇文档作为测试集,去掉文档频率小于 5 的单词后,20N 的单词数为13 558. Reuters10 是从 Reuters-21 578 中提取的前 10 大类文档,训练集包括 7 193 篇文档,测试集包括 2 787 篇文档,Reuters10 的预处理过程与 20N 的预处理过程类似,只不过文档频率小于 3 的单词被从文档中删去. Ohsumed,Sports 和 Reviews 是文献[19]用到并处理成向量空间模型的文本,我们随机抽取其中 50%的文本作为训练集,50%的文本作为测试集.

2.2 实验结果

为了描述标注数据与原始数据分布的吻合程度,对于数据的每一个类别,分别求出各个单词在原始数据和标注数据上的文档频率 $(df_1, df_2, \cdots, df_M)$ 及 $(df'_1, df'_2, \cdots, df'_M)$,并进行归一化,使得 $\sum_{m=1}^{M} df_m = \sum_{m=1}^{M} df'_m = 1$.那么标注数据和原始数据在这个类别上的分布不一致性可以定义为 $D = \sum_{m=1}^{M} (df_m - df'_m)^2$. D 越小标注数据与原始数据在这一类上的分布越一致.以 Reviews 数据集为例,本节用不同标注方法从 Reviews 训练集中抽取 10%的样本,然后计算在不同类别下这些样本与原始数据分布的不一致性.用到的类别标注方法包括随机标注方法(简记为 Rand);FFT 标注方法;本文提出的利用样本近邻信息,但不考虑样本信息熵的方法(简记为 NN);本文提出的利用样本近邻信息,且考虑样本信息熵的方法(简记为 NN+EN).

各标注方法得到的不一致性指标值如表 3 所示.表 3 中,黑体数字表示不一致性指标值最小,下划线数字表示不一致性指标值第 2 小.可以看出,NN 和 NN+EN 方法得到的标注数据比随机标注和 FFT 方法得到的标注数据更吻合原始数据的分布.

Table 3 Values of D on the Reviews Data

表 3 Reviews 数据上的不一致值

Category	Method			
	Rand	FFT	NN	NN+EN
1	8.6025	25.1016	8.4766	**5.9278**
2	7.2360	5.8841	7.1121	**5.2413**
3	6.5328	11.7690	7.2710	**4.0080**
4	95.6124	107.7472	**87.4538**	121.9043
5	18.0586	89.6983	19.0334	**10.5082**

Note: the data in the table are the real values of D multiplied by 10^{-5}

为了进一步展示本文方法的优越性,用不同的标注方法分别对 5 组文本数据进行标注,在此基础上,运行基于期望最大化的贝叶斯分类器,分类器性能的好坏就体现出标注方法的好坏.

由于本文提出的标注方法是基于近邻和信息熵的,本节定义每个样本的近邻数为 10. 对于每个数据集来说,样本的标注比例从 2%变化到 10%. 为了减小随机误差,对于每个标注比例和不同的标注方法,都抽取并标注样本 10 次,然后运行 10 次基于期望最大化的贝叶斯分类器,计算 10 次分类结果的平均准确率. 因为训练集和测试集中都包含大量未标注数据,我们把训练集中未标注数据的分类准确率和测试集中数据的分类准确率分开表示,如表 4 所示. 表 4 中,黑体数字表示分类准确率最高,下划线数字表示分类准确率第 2 高. 可以看出在半监督分类算法固定的情况下,本文提出的基于近邻的文本标注方法以及基于近邻和信息熵的文本标注方法可以得到比随机标注方法和 FFT 方法更好的分类结果.

Table 4 Classification Accuracy on the Text Corpora

表 4 文本数据的分类准确率

Data	Labeling Percentage/%	Unlabeled Data in the Training Set				Data in the Testing Set			
		Rand	FFT	NN	NN+EN	Rand	FFT	NN	NN+EN
20N	2	0.4521	0.4461	0.4750	**0.5067**	0.3695	0.4033	0.4073	**0.4220**
	4	0.5553	0.5763	0.5627	**0.5950**	0.4720	**0.5184**	0.4733	0.5112
	6	0.6189	0.6412	**0.6493**	0.6418	0.5269	0.5573	**0.5692**	0.5559
	8	0.6478	0.6528	**0.6759**	0.6639	0.5572	0.5485	**0.5933**	0.5778
	10	0.6577	0.7065	**0.7255**	0.6808	0.5558	0.6060	**0.6515**	0.5911
Reuters10	2	0.7140	0.5543	**0.7320**	0.6623	0.7404	0.6323	**0.7511**	0.6647
	4	**0.7495**	0.5702	0.7404	0.7139	**0.7686**	0.6364	0.7633	0.7311
	6	0.7486	0.6128	0.7472	**0.7684**	0.7715	0.6628	0.7670	**0.7920**
	8	0.7709	0.6241	0.7629	**0.7805**	0.7896	0.6694	0.7858	**0.8079**
	10	0.7824	0.6463	0.7747	**0.7900**	0.8013	0.6899	0.7935	**0.8111**
Ohsumed	2	0.4981	0.5057	**0.5099**	0.5078	0.4470	0.4520	0.4514	**0.4540**
	4	0.5663	0.5710	**0.5816**	0.5591	0.5080	0.5145	**0.5220**	0.5116
	6	0.5916	0.5922	**0.6127**	0.5781	0.5363	0.5390	**0.5515**	0.5256
	8	0.5841	0.6146	**0.6183**	0.6000	0.5300	0.5566	**0.5583**	0.5483
	10	0.6122	**0.6187**	0.6186	0.6023	0.5506	**0.5591**	0.5572	0.5514
Sports	2	0.7932	0.7926	**0.8210**	0.7938	0.7635	0.7494	**0.8020**	0.7678
	4	0.8047	0.7932	0.8461	**0.8488**	0.7803	0.7554	0.8279	**0.8306**
	6	0.8469	0.8023	0.8485	**0.8557**	0.8303	0.7548	0.8286	**0.8433**
	8	0.8605	0.8037	0.8621	**0.8673**	0.8381	0.7650	0.8448	**0.8561**
	10	0.8706	0.8005	0.8665	**0.8713**	0.8549	0.7552	0.8493	**0.8599**
Reviews	2	0.6469	**0.7266**	0.6841	0.7264	0.6114	**0.6974**	0.6514	0.6906
	4	0.7081	0.7234	**0.7682**	0.7423	0.6888	0.6803	**0.7509**	0.7337
	6	0.7011	0.7459	**0.7882**	0.7801	0.6946	0.7196	**0.7688**	0.7573
	8	0.7123	0.7452	0.7680	**0.7697**	0.6981	0.7157	0.7435	**0.7495**
	10	0.7523	0.7235	**0.8243**	0.7711	0.7325	0.6911	**0.8136**	0.7505

除了分类精度,本文也记录了提出方法在普通PC机上的时间开销.表5表示利用matlab实现的方法在不同数据上抽取10%标注样本所花费的时间.可以看出,在数据量不是特别大的情况下,本文方法的时间开销还是可以接受的.

Table 5 Running Time of the Proposed Method
表5 算法的运行时间

Data	Data Size	Running Time/s
20N	5562×13558	68.52
Reuters10	7193×10509	72.59
Ohsumed	5584×11101	55.09
Sports	4291×21560	67.36
Reviews	2036×28536	30.09

3 结论和进一步研究

本文提出了一种基于近邻和信息熵的主动文本标注方法.该方法通过避免选择已标注样本的K近邻作为新一轮候选标注样本,使得分布在数据不同区域的样本有更多的标注机会.同时,该方法还利用信息熵理论在候选标注样本中选择信息量最大的样本进行标注.实验表明本文提出的标注方法可以有效提高半监督文本分类的性能.值得一提的是,本文在选择信息量最大的样本时采用了信息熵理论,接下来的研究将会探索更有效的方法在候选标注样本中挑选信息量最大的样本.此外,今后也会从理论和实验上对与分类器交互的增量标注方法以及独立于分类器的批量标注方法进行更深入的研究比较.

参 考 文 献

[1] Nigam K, McCallum A, Thrun S, et al. Text classification from labeled and unlabeled documents using EM [J]. Machine Learning, 2000, 39(2/3): 103-134

[2] Zhu Xiaojin, Ghahramani Z, Lafferty J. Semi-supervised learning using Gaussian fields and harmonic functions [C] // Proc of the 20th Int Conf on Machine Learning. New York: ACM, 2003: 912-919

[3] Zhou Dengyong, Bousquet O, Lal N, et al. Learning with local and global consistency [C] //Advances in Neural Information Processing Systems 16. Cambridge, MA: MIT, 2004: 321-328

[4] Wang Fei, Zhang Changshui. Label propagation through linear neighborhoods [J]. IEEE Trans on Knowledge and Data Engineering, 2008, 20(1): 55-67

[5] Joachims T. Transductive inference for text classification using support vector machines [C] //Proc of the 16th Int Conf on Machine Learning. New York: ACM, 1999: 200-209

[6] Belkin M, Niyogi P, Sindhwani V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples [J]. Journal of Machine Learning Research, 2006, 7(11): 2399-2434

[7] Li Yufeng, Kwok J, Zhou Zhihua. Semi-supervised learning using label mean [C] //Proc of the 26th Int Conf on Machine Learning. New York: ACM, 2009: 633-640

[8] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training [C] //Proc of the 11th COLT. New York: ACM, 1998: 92-100

[9] Zhou Zhihua, Zhan Dechuan, Yang Qiang. Semi-supervised learning with very few labeled training examples [C] //Proc of the 22nd AAAI. Menlo Park, CA: AAAI, 2007: 675-680

[10] Zhou Zhihua, Li Ming. Tri-training: Exploiting unlabeled data using three classifiers [J]. IEEE Trans on Knowledge and Data Engineering, 2005, 17(11): 1529-1541

[11] Li Ming, Zhou Zhihua. Online semi-supervised learning with multi-kernel ensemble [J]. Journal of Computer Research and Development, 2008, 45(12): 2060-2068 (in Chinese)
(黎铭, 周志华. 基于多核集成的在线半监督学习方法[J]. 计算机研究与发展, 2008, 45(12): 2060-2068)

[12] Muslea I, Minton S, Knoblock C A. Selective sampling with redundant views [C] //Proc of the 17th National Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2000: 621-626

[13] Muslea I, Minton S, Knoblock C A. Active + semi-supervised learning=robust multi-view learning [C] //Proc of the 19th Int Conf on Machine Learning. New York: ACM, 2002: 435-442

[14] Muslea I, Minton S, Knoblock C A. Active learning with multiple views [J]. Journal of Artificial Intelligence Research, 2006, 27: 203-233

[15] Nguyen T, Smeulders A. Active learning using pre-clustering [C] //Proc of the 21st Int Conf on Machine Learning. New York: ACM, 2004: 623-630

[16] Xu Zhao, Yu Kai, Tresp V, et al. Representative sampling for text classification using support vector machines [G] // LNCS 2633: Proc of the 25th European Conf on Information Retrieval Research. Berlin: Spinger, 2003: 393-407

[17] Zhang Xue, Zhao Dongyan, Chen Liwei, et al. Batch mode active learning based multi-view text classification [C] //Proc of the 6th Int Conf on Fuzzy Systems and Knowledge Discovery. Los Alamitos, CA: IEEE Computer Society, 2009: 472-476

[18] Basu S, Banerjee A, Mooney R J. Active semi-supervision for pairwise constrained clustering [C] //Proc of the SIAM Int Conf on Data Mining. Philadelphia, PA: SIAM, 2004: 333-344

[19] Zhao Ying, Karypis G. Empirical and theoretical comparisons of selected criterion functions for document Clustering [J]. Machine Learning, 2004, 55(3): 311-331

Zhu Yan, born in 1982. PhD candidate in the Department of Computer Science, Beijing Jiaotong University. Her current research interests include machine learning and text mining.

Jing Liping, born in 1978. Associate professor in the Department of Computer Science, Beijing Jiaotong University. Her current research interests include machine learning and text mining (lpjinghk@gmail.com).

Yu Jian, born in 1969. Professor in the Department of Computer Science, Beijing Jiaotong University. His current research interests include machine learning, text mining and image processing (jianyu@bjtu.edu.cn).

一种利用近邻和信息熵的主动文本标注方法

作者： 朱岩， 景丽萍， 于剑， Zhu Yan， Jing Liping， Yu Jian
作者单位： 北京交通大学计算机科学系 北京100044
刊名： 计算机研究与发展 ISTIC EI PKU
英文刊名： Journal of Computer Research and Development
年，卷(期)： 2012,49(6)

参考文献(19条)

1. Nigam K;McCallum A;Thrun S Text classification from labeled and unlabeled documents using EM[外文期刊] 2000(2/3)

2. Zhu Xiaojin;Ghahramani Z;Lafferty J Semi-supervised learning using Gsussisn fields and harmonic functions 2003

3. Zhou Dengyong;Bousquet O;Lal N Learning with local and global consistency 2004

4. Wang Fei;Zhang Changshui Label propagation through linear neighborhoods[外文期刊] 2008(01)

5. Joachims T Transductive inference for text classification using support vector machines 1999

6. Belkin M;Niyogi P;Sindhwani V Manifold regularization:A geometric framework for learning from labeled end unlabeled examples 2006(11)

7. Li Yufeng;Kwok J;Zhou Zhihua Semi-supervised learning using label mean 2009

8. Blum A;Mitchell T Combining labeled and unlabeled data with co-training 1998

9. Zhou Zhihua;Zhan Dechuan;Yang Qiang Semi-supervised learning with very few labeled training examples 2007

10. Zhou Zhihua;Li Ming Tri-training:Exploiting unlabeled data using three classifiers[外文期刊] 2005(11)

11. 黎铭;周志华 基于多核集成的在线半监督学习方法[期刊论文]-计算机研究与发展 2008(12)

12. Muslea I;Minton S;Knoblock C A Selective sampling with redundant views 2000

13. Muslea I;Minton S;Knoblock C A Active + semisupervised learning=robust multi-view learning 2002

14. Muslea I;Minton S;Knoblock C A Active learning with multiple views 2006

15. Nguyen T;Smeulders A Active learning using predustering 2004

16. Xu Zhao;Yu Kai;Tresp V Representative sampling for text classification using support vector machines 2003

17. Zhang Xue;Zhao Dongyan;Chen Liwei Batch mode active learning based multi-view text classification 2009

18. Basu S;Banerjee A;Mooney R J Active semi-supervision for pairwise constrained clustering 2004

19. Zhao Ying;Karypis G Empirical and theoretical comparisons of selected criterion functions for document Clustering[外文期刊] 2004(03)