基于信息熵的主动学习半监督分类研究

陈锦禾,沈 洁
(扬州大学 信息中心,江苏 扬州 225009)

摘 要:针对小规模训练样本不足以支持学习器对含有大量潜在不确定因素的未标样本集分类的问题,提出了一种基于信息熵的主动学习方法,引入信息熵的离散事件概率估计理论,通过对未标文档熵值的计算,结合二阶段学习策略,主动学习利用现有知识,结合实验样本环境,主动地选取最有可能的解决问题的样本并标注它们的类别,获得新的参数,重新训练分类器,选择最有利分类器性能的样本,迭代直到未标样本集为空。实验结果表明,该方法取得了较好的分类效果。

关键词:信息熵;半监督学习;主动学习;分类
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2010)02-0110-04

Active Learning Based on Information Entropy for Semi-supervised Classification

CHEN Jin-he, SHEN Jie
(Information Center of Yangzhou University, Yangzhou 225009, China)

Abstract: Most of supervised machine learning methods led to poor performance when work on limited tagged data. Investigated a novel semi-supervised learning method based on active learning with information entropy. An optimization strategy of selecting part of instances from unlabeled examples for classifying in each iteration, based on active learning from unlabeled examples, was presented. The experiment results show that our method achieve high performance on small tagged data.

Key words: information entropy; semi-supervised learning; active learning; classification

0 引 言

文本分类技术是优化文本管理的重要环节,目的在于对未知类别的文档进行自动分类。传统的基于机器学习的分类方法主要为有监督学习和无监督学习。有监督学习主要分为两个过程:训练过程和分类过程,通过对标记样本的训练学习,计算分类器的参数,对未标文档进行分类的过程。主要有 Rocchio algorithm[1], Naive Bayes[2], support vector machines[3]等方法。有监督学习需要标记大量的训练样本,耗费大量的人力物力。无监督学习在无训练样本的情况下,针对样本分布特征进行样本标注,但是准确性较差。半监督学习克服了两者的缺点,只需提供少量的标记样本,对大量的未标样本进行学习分类,因此半监督学习得到了广泛的应用[4,5]。

基于机器学习理论的假设,机器学习方法的语料的选择是随机的,也就是说训练样本和待分类样本是独立同分布的。实际中,这一假设对于半监督学习通常难以成立。通常情形下,未标样本可能来自于与训练样本不同的分布或者不同的环境,从而标注样本数量不足以支持学习器在假设空间内搜索到目标函数。针对这一问题,笔者提出了一种基于信息熵的主动学习方法,通过对未标样本熵值的计算,使用分类样本的内在信息获取与分类相关的局部模型,使分类器有针对性地提高分类效果。

1 贝叶斯分类模型

文中主要采用贝叶斯作为分类模型。在文本分类中,贝叶斯分类模型是一个简单有效的方法。它的基本思想是,基于类别概率和词项的联合分布概率预测未知文档的类别。在很多的研究中,它被证明是一个很好的方法。首先,需要给每个类别训练文档。根据这些训练的文档,初始化一个分类器,采用这个分类器对新的文档进行分类。

设类别的集合为 $C=\{c_1,c_2,\cdots,c_n\}$,每一个文档用一系列的词表示 $V=(w_1,w_2,\cdots,w_n)$。计算文档

收稿日期:2009-06-09;**修回日期**:2009-09-20
基金项目:国家自然科学基金(60673060)
作者简介:陈锦禾(1965-),男,工程师,研究方向为信息安全与管理;沈 洁,教授,研究方向为信息安全与管理。

的后验概率 $p[c_i \mid d_j]$。c_i 表示第 i 个类,d_j 表示第 j 个文档。基于贝叶斯概率多项式模型,计算类先验概率:

$$p[c_i] = \sum_j p[c_i \mid d_j] / \mid D \mid \tag{1}$$

式中 $\mid D \mid$ 代表所有训练的文档数,$P(c_i)$,$i = 1, 2, \cdots, n$,其值为 c_i 类样本除以训练集总样本数。对于新样本 d,其属于 c_i 类的条件概率是 $P(d \mid c_i)$,对于 $P(d \mid c_i)$ 的计算转换为对 $p[w_t \mid c_i)]$ 的计算:

$$p[w_t \mid c_i] = \frac{1 + \sum_{j=1}^{\mid D \mid} N(w_t, d_j) p(c_i \mid d_j)}{\lambda + \sum_{s=1}^{\mid V \mid} \sum_{j=1}^{\mid D \mid} N(w_s, d_j) p(c_i \mid d_j)} \tag{2}$$

其中,$N(w_t, d_j)$ 表示特征 w_t 在文本 d_j 中出现的次数,$\mid v \mid$ 是 V 中词的数量。假设词之间是独立分布的,在分类过程中通过计算文本的后验概率 $p(c_i \mid d_j)$,最终将 d_j 分入使得后验概率最大的类别,由 Bayes 公式:

$$p[c_i \mid d_j)] = \frac{p[c_i] \prod_{k=1}^{\mid d_j \mid} p[w_{d,k} \mid c_i]}{\sum_{r=1}^{\mid c \mid} p[c_r] \prod_{k=1}^{\mid d_j \mid} p[w_{d_j,k} \mid c_r]} \tag{3}$$

2　基于信息熵理论主动学习

从以上贝叶斯分类模型可以看出,对于未标文档的分类,主要通过对文档中词的概率计算文档的条件概率,从而决定了文档的类别。因此对于文档的熵值的计算主要基于文档中词的熵值的计算。由于未标文本分布的不确定性,特征复杂且未知,带有很多与训练样本不同分布的特征词,从而使得训练数据不足以支持学习器对未标文档进行较好的分类。

2.1　主动学习策略

根据分类学习在训练集中的不同处理方式,分类模型可以分为两类,一类是被动学习分类,另一类是主动学习分类。被动学习分类,主要是对训练样本选择的随机性,从而当未标样本中噪音样本较多的情况下,分类性能较差。未标样本集通常含有有助于分类的样本数据,利用这些文档信息能够有效提高分类的性能。主动学习主要基于机器学习的方法,按照某种策略从待分类样本集中动态地选择样本进行分类的过程[6,7]。目前,主要有如下几种策略:基于样本的不确定性方法、询问专家委员会的方法、版本空间和边缘的方法、基于统计的方法等[8~12]。半监督的学习分类,由于训练样本规模较小,以及大量未标样本的不确定性,导致分类器的分类性能不理想。文中针对这一问题,引入了一种基于未标样本的学习模型,使得在小规模的样本环境下,获取高性能的分类效果。主要基于两种主动学习策略:

主动学习策略 1:对于未标文档是否适应于当前训练集环境下的分类;

主动学习策略 2:未标文档中哪些文档可以作为新加入的训练文档,完善分类器的学习。

在我们的学习中,主要引入了信息熵的理论,基于信息熵的主动学习。

2.2　信息熵的主动学习半监督分类系统

信息熵是 Shannon 在 1948 年基于熵的概念引入信息论中的,解决了对信息的量化度问题。它表示一个概率分布的不确定性,在一定的约束条件下,选择具有最大不确定性的分布[13]。引入信息熵的知识,判断词的概率分布,从而整体上判断某一未标样本是否用于当前分类器的分类样本。信息熵计算如下:

$$\text{entropy}(w_i) = -\sum_c P(w_i \mid c) * \log(P(w_i \mid c)) \tag{4}$$

熵值反应了特征词分布的差异情况,如果熵值较大,说明该词与训练样本数据就越有可能有类似的分布。对于一个与训练样本数据分布差异较大的特征词,它的熵值较小。基于信息熵理论,对数据的标准化处理,采用极大值标准化处理方法,得到矩阵:$\boldsymbol{Y} = \{y_{ij}\}_{H \times m}$,其中

$$y_{ij} = \frac{\text{entropy}(w_i)}{\max_{j=1,2,\cdots,\mid v \mid}(\text{entropy}(w_j))} \tag{5}$$

对特征词熵的加权如下:

$$q(w_i) = 1 - \frac{\text{entropy}(w_i)}{\max_{j=1,2,\cdots,\mid v \mid}(\text{entropy}(w_j))} \tag{6}$$

当 $q(w_i) = 0$,那么该特征词与训练样本的分布毫无疑问是同一分布的,当 $q(w_i) = 1$,则与训练样本数据不同一分布。在具体的计算中,$q(w_i)$ 的值通常是一个小数。通过对文档中词的熵值的计算,综合考虑该文档与训练数据的分布差异。基于贝叶斯假设理论,这里假设词之间的信息熵是独立同分布的。对于两个特征词联合熵的表达如下:

$$\text{entropy}(w_i \cup w_j) = \text{entropy}(w_i) + \text{entropy}(w_j \mid w_i) \tag{7}$$

对于 $\text{entropy}(w_j \mid w_i)$ 的计算如下:

$$\begin{aligned} &\text{entropy}(w_j \mid w_i) = E[I(w_j \mid w_i)] \\ &= \sum_c P(w_i \mid c) \text{entropy}(w_j \mid w_i) \\ &= -\sum_c P(w_i \mid c) \sum_c P(w_j \mid c) \log(P(w_j \mid c)) \end{aligned} \tag{8}$$

从而,得出对文档的熵值的计算如下:

$$\begin{aligned} &\text{entropy}(d) = \\ &\sum_{i=1}^{n} \sum_{j=1}^{n} \Big(-\sum_c P(w_i \mid c) * \log(P(w_j \mid c)) \\ &- \sum_c P(w_i \mid c) \sum_c P(w_j \mid c) \log(P(w_j \mid c)) \Big) \end{aligned} \tag{9}$$

n 代表该文档中词的数量。当 $\text{entropy}(d)$ 的值越大,则被选择为当前分类器的分类文档的概率则越大。根据未标文档的具体情况,挑选一个区别度较好的阈

值。当 entropy(d)的值在阈值之上时,则被选为分类样本,根据分类器对其进行类别的标注。并将这些被标注的文档加入到训练集中,重新训练分类器。因此,文中的主动学习是一个循环反复的过程。每次迭代,选取确定性最强的样本进行分类,并且完善训练数据。具体的算法如下:

算法 1.基于信息熵的主动学习算法

输入:D_{tra}:规模较小的训练样本集,

D_{unl}:规模较大的未标样本集。

输出:D_{unl}的分类值 $c_1, c_2, \cdots, c_n$

Step 1:D_{test}:待分类样本集,$D_{test} = \varnothing$

Step 2:基于训练集 D_{tra} 训练一个贝叶斯分类器 NB-C

Step 3:设置区别度的阈值 ε

Step 4:对于所有文档 $d_j \in D_{unl}$,执行如下循环:

4.1 计算所有的词 $w_i \in d_j$ 的信息熵权值 $q(w_i)$

4.2 计算文档的熵值 entropy(d)

4.3 调用主动学习策略 1,如果

entropy(d) $> \varepsilon$, $D_{test} = D_{test} \cup d_j$

Step 5:使用 step 2 中的 NB-C 对 D_{test} 中的样本进行分类,D_{test}':被分类器标注类别的文档集

Step 6:调用主动学习策略 2,将这些带有类别的文档并入 D_{tra} 中,$D_{tra} = D_{tra} \cup D_{test}'$

Step 7:$D_{unl} = D_{unl} - D_{test}'$,

如果 $D_{unl} = \varnothing$,终止学习,退出;

否则,跳到 step 1。

3 实验设计与结果分析

3.1 数据集

为了模拟未标样本与训练样本数据环境的不一致,以及分布未知情形,实验的数据通过 spider 从五个结构不同的网站搜集网页,分别是 http://www.163.com、http://www.sina.com、http://www.digitalchina.com、http://digital.pconline.com.cn 和 http://www.jwdigital.com。类别主要是关于数码方面的网页,主要分为电脑、MP3/MP4、打印机、数码相机、电视、手机共六个类,经过预处理,转换成统一的文本形式。它们的类别和分布见表 1。

表 1 选择的实验文档及其分布

类别	电脑	MP3/MP4	打印机	数码相机	电视	手机
163	101	83	120	164	74	88
sina	107	79	85	93	95	90
Digitalchina	96	68	79	81	53	110
pconline	93	91	78	72	86	65
jwdigital	109	117	90	121	100	87
总数	516	438	452	531	408	440

3.2 实验结果

将这些文档分成两个集合,一个作为训练集,一个作为未标文档集。文中主要研究基于小规模样本环境下的分类性能,因此对于数据集的分别如下:分别从各个网站中选取 50% 的文档作为训练集,剩下的文档与其他网站的文档构成未标文档集,进行六次不同的实验。

表 2 为六组不同的训练样本每次迭代所产生的标注样本数。从中可见虽然从各个网站选择的样本类别是一致的,但是各个网站对于各个产品的布局风格有差异,大量的未标文本集与训练数据的分布差异等导致训练数据不足以支持学习器对其进行分类。通过对未标样本内部信息的主动学习,选取最有可能作为当前分类器分类的样本进行类别的标注。在每次的迭代中,对未标样本循序渐进地学习和标注,利用未标样本数据的信息来解决未标样本集的分类问题。通过对训练集的有效扩充,提升分类器的分类性能。随着标注样本数量的增大,学习器不断适应当前环境下的分类。基于实验数据规模,经过九次左右的迭代学习,完成了对所有未标样本的学习标注。

表 2 六组数据每次迭代产生的标注样本数

训练数据	163	sina	Digitalchina	pconline	jwdigital
一	12	24	19	17	23
二	21	33	30	29	47
三	45	51	73	55	89
四	92	163	162	98	201
五	213	347	397	201	442
六	451	789	1050	447	908
七	948	1765	2559	976	2023
八	1546	2266		1879	2307
九	2527			2520	

表 3 列出了在每次迭代的过程中,对于学习器选择的待分类样本的分类效果,主要通过综合指标 F 值评价分类的性能。F 值的定义如下:

$F = \frac{2pr}{p+r}$, p 为精率度, r 为召回率。

表 3 在不同标注样本数下的分类性能

F 值	163	sina	Digitalchina	pconline	jwdigital
一	77.03	72.73	77.38	75.43	73.21
二	77.34	72.91	77.54	75.10	73.46
三	78.41	74.08	77.87	75.87	73.58
四	79.33	76.45	78.09	76.25	74.64
五	80.19	76.77	78.94	76.37	74.89
六	80.38	76.82	79.76	78.22	75.31
七	80.57	77.01	80.34	78.91	75.98
八	80.60	77.12		80.10	76.04
九	80.88			80.32	

F 值近似平坦,因为其分类性能仅与本次待分类的样本数有关,与未标样本数无关。基于每次的迭代选择策略,选择最大可能的适应于当前训练数据下的学习器的分类样本。随着训练数据的不断扩充,待分

类样本的规模增大,分类的性能稳步上升,分别从77.03、72.73 等上升到 80.88、77.12 等。最终完成了对未标样本的分类。

为了验证在基于小规模样本环境下主动学习的优势,以经典的受监督学习方法贝叶斯分类器进行分类。分别以某一网站的 50% 数据作为训练集,剩下的数据与其他网站的数据作为分类样本。实验结果见表 4。

表 4　简单分类器下的分类性能

	163	sina	Digitalchina	pconline	jwdigital
F 值	42.06	34.96	42.12	52.20	44.87

由于训练样本数据与分类样本具有交大的差异,分类样本中不确定性样本的比例较大,导致学习器在对未知数据进行分类时分类性能不理想。这说明训练样本数据的选择对于未标分类样本的分类有很大的影响。

针对这一问题,采用文中的信息熵主动学习方法,通过引入信息熵的信息量化度理论,每次迭代选取最大适应于当前分类器的样本进行分类同时完善训练数据,优化分类器的参数。有效利用了未标样本内部信息,提高了分类器的性能。

4　结束语

主要针对小规模训练样本环境下,如何对含有大量潜在不确定因素的未标样本集进行分类。通过引入主动学习的策略,基于信息熵的信息量化理论,利用未标样本的内在信息进行学习分类。实验结果表明,该方法具有较好的分类性能。

参考文献:

[1] Rocchio J. Relevant feedback in information retrieval[M]//In Salton G. The smart retrieval system - experiments in automatic document processing. Englewood Cliffs, NJ: [s. n.], 1971.

[2] McCallum A, Nigam K. A comparison of event models for naive Bayes text classification[C]//AAAI - 98 Workshop on Learning for Text Categorization. [s. l.]: AAAI Press, 1998.

[3] Guyon I, Boser B, Vapnik V. Automatic capacity tuning of very large Vcdimension classifiers[J]. Advances in Neural Information Processing Systems, 1993(5): 147 - 155.

[4] Igam K, McCallum A, Thrun S, et al. Learning to classify text from labeled and unlabeled documents[C]//In: Mostow J, Madison C R. Proceedings of the 15th National Conference on Artificial Intelligence. Wisconsin: AAAI Press, 1998: 792 - 799.

[5] 刘　晶,郭　雷,聂晶鑫. 基于 SVM 的一种新的分类器设计方法[J]. 计算机应用研究,2006,23(7):181 - 183.

[6] Engelbrecht A P, Cloete I. Incremental Learning Using Sensitivity Analysis[C]// Neural Networks, 1999. IJCNN apos; 99. International Joint Conference. [s. l.]: IEEE Press, 1999: 1350 - 1355.

[7] 陈耀东,王　挺,陈火旺. 半监督学习和主动学习相结合的浅层语义分析[J]. 中文学习学报,2008,22(2):70 - 75.

[8] Thompson C A, Califf M E, Mooney R J. Active Learning for Natural Language Parsing and Information Extraction[C]// In: Proceedings of the sixteenth Internatlonal Machine Learning Conference. Slovenia: [s. n.], 1999.

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

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

[11] Liere R, Tadepalli P. Active learning with committees for text categorizeation[C]//In Proceedings of the Fourteenth National Conference on Articial Intelligence. Providence, RI: [s. n.], 1997: 591 - 596.

[12] McCallum A, Nigam K. Employing EM and pool - based active learning for text classification[C]//In Machine Learning: Proceedings of the Fifteenth International Conference (ICML '98). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc, 1998: 359 - 367.

[13] Carter T. An introduction to information theory and entropy [EB/OL]. 2007 - 06. http://astarte. csustan - edu/tom/SFICSSS.

(上接第 109 页)

[4] Brodie B C, Taylor D E, Cytron R K. A Scalable Architecture For High - Throughput Regular - Expression Pattern Matching[J]. Computer Architecture, 2006, 11(1): 191 - 202.

[5] 陆　虎,宋余庆,薛万宇,等. 一种基于正则表达式匹配的协议分析异常检测方法[J]. 计算机应用与软件,2008,25(3):118 - 122.

[6] 李哲夫. 正则表达式在电信业务处理中的应用研究[M]. 广州:暨南大学出版社,2008.

[7] 余石泉,周肆清. 正则表达式在编程题自动阅卷中的应用[J]. 计算机技术与发展,2007,17(7):109 - 112.

[8] 王功明,吴华瑞,赵春江,等. 正则表达式在电子政务客户端校验中的应用[J]. 计算机工程,2007,33(9):140 - 143.

[9] Singh C T, Maulik U. A framework for an artificial immunity and speech based navigation for mobile robots[J]. Evolutionary Computation, 2008, 3(4): 1247 - 1251.