第35卷 第18期 计 算 机 工 程 2009年9月
Vol.35 No.18 Computer Engineering September 2009
·人工智能及识别技术· 文章编号：1000—3428(2009)18—0182—03 文献标识码：A 中图分类号：TP181

不等式最大熵中的特征选择方法

张 永，李晓红，樊 斌
(兰州理工大学计算机与通信学院，兰州 730050)

摘 要：不等式最大熵模型较为成功地缓解了文本分类任务中的过拟合问题，但它使用的特征选择算法不能完全发挥不等式最大熵的最大优势。针对该问题提出采用改进的顺序前进式选择算法，提高文本分类任务中的识别率，试验结果证明该算法能够更准确地选出文本代表特征，对不等式最大熵模型的分类成绩有一定的改善。

关键词：不等式最大熵；特征选择；文本分类

Feature Selection Method for Inequality Maximum Entropy

ZHANG Yong, LI Xiao-hong, FAN Bin
(School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050)

【Abstract】 Inequality maximum entropy method has alleviated data sparseness with flexible modeling capability more successfully than other probabilistic models in text classification tasks, but feature selection algorithm used by the model can not fully bring its advantage. This paper proposes a new feature selection method. It improves the recognition rate in text classification. Experimental result shows that this algorithm works more effectively in selecting representative features and improves the text classification performance a lot.

【Key words】 inequality maximum entropy; feature selection; text classification

1 概述

随着Web电子文档以指数级的速度增长，为了应对信息爆炸给人们带来的挑战，自动的文本分类技术在信息检索、信息过滤等领域越来越重要，目前，文本分类也是数据挖掘领域的热门话题，而特征选择作为文本分类的前提，其重要性更是不言而喻的。

Della pietra[1]等人将最大熵方法首次引入自然语言处理(NLP)的语言模型领域之后得到了迅速的发展，它以其强健有力的建模能力、简单有效的学习算法而倍受研究人员的关注并被广泛应用。其优势已经在词义排歧[2]、POS tagging[3]、文本分类[4]、统计机器翻译等各种NLP任务中凸现。最大熵模型中的特征可以根据需求指定，不要求彼此独立，而且，最大熵方法可以灵活地将一些跨距离的特征加入到模型中，并善于将各种知识结合起来，因此，最大熵在文本分类任务中达到了较好的分类效果。

为了提高文本分类的成绩，研究学者关于最大熵做了很多的工作，例如 Nigam[5]在最大熵模型中引入了平滑技术和正则化方法来解决数据稀疏引起的过拟和问题；Kazama 等人又对照最大熵模型提出了不等式最大熵模型，其目的也是为了改善文本分类任务中出现的过拟和问题。但是，一直以来，大量的研究者都把注意力集中在如何解决数据稀疏问题和平滑技术，而忽视了特征选择方法对最大熵分类器的分类成绩的作用。所以说，寻求一种有效的特征选择方法得到低维高效的特征子集，提高文本分类的精度和效率成为最大熵分类器面对的一个重要问题。

2 不等式最大熵在文本分类中的应用

2.1 不等式约束和不等式最大熵模型

不等式约束完全打破了经验期望和模型期望相等的平衡状态，对等式约束 $E_{\tilde{p}}[f_i]-E_p[f_i]=0$ 进行彻底的松弛操作，得到如下形式：

$$-B_i \leqslant E_{\tilde{p}}[f_i]-E_p[f_i] \leqslant A_i \qquad A_i>0, B_i>0 \tag{1}$$

其中，A, B 称为约束宽度，用来反映特征的可靠性，其值可通过 single 方法或 bayes 方法求得，因此，分类问题就变成了在式(1)所示的一组不等式约束条件下求式(2)的最优解的问题。

$$P^*=\text{argmax}\ H(p) \tag{2}$$

其中，p 为所有满足式(1)的模型。

该模型与标准最大熵的区别主要有2点：(1)约束条件为不等式；(2)特征参数为2个。故对照最大熵模型，把满足上面这个不等式约束的模型称作是不等式最大熵模型[6]。最后输出的最优解具有下面的形式：

$$P_{\alpha,\beta}(y|x)=\frac{1}{Z(x)}\exp\left(\sum_i(\alpha_i-\beta_i)f_i(x,y)\right) \tag{3}$$

其中，$Z(x)$是归一化因子。另外，模型中的每个特征均有2个参数，并且要求 $\alpha_i \geqslant 0$，且 $\beta_i \geqslant 0$，α_i-β_i 的值可以是任何实数值，α_i 对应不等式的上限 A_i，β_i 对应不等式的下限 B_i。如果假设 $\lambda_i=\alpha_i-\beta_i$，则不等式最大熵就等于标准最大熵，等式约束下的解在最优点上也成立。

不等式最大熵模型最吸引人的特点是将特征选择和Gaussian MAP估计[7]2种方法的优点结合起来，在模型估计过程中嵌入特征选择，有效地控制了模型的规模大小，并且对数据稀疏问题引起的过拟合现象也有很大的改善。

这里再简单地叙述一下不等式最大熵模型的建模过程。

作者简介：张 永(1968－)，男，副教授，主研方向：智能信息处理；李晓红，讲师、硕士研究生；樊 斌，硕士

收稿日期：2009-03-14 **E-mail**：xiaoxiao526@126.com

为了找到熵最大且分布均匀的目标模型，直接求解难度太大，故将目标函数进行了转换，形式如下式：

$$L(\alpha,\beta)=-\sum_x \tilde{p}(x)\mathrm{lb}\sum_y \exp(\sum_i(\alpha_i-\beta_i)f_i(x,y))+\sum_x \tilde{p}(x)\sum_y \tilde{p}(y|x)\sum_i(\alpha_i-\beta_i)f_i(x,y)-\sum_i \alpha_i A_i-\sum_i \beta_i B_i \quad (4)$$

由于不等式的参数受到限制，而 BLMVM 算法[8]满足参数受限的建模要求，因此求不等式最大熵的最优解时使用它，使用 BLMVM 算法对式(5)求最大值便得到如式(2)所示的概率模型。另外，如果通过在训练集上进行学习，知道了 α_i, β_i 的值，就得到了概率分布函数，完成了不等式最大熵模型的构造：

$$LL(\alpha,\beta)-\sum_i \alpha_i A_i-\sum_i \beta_i B_i \quad (5)$$

2.2 特征函数的选择

通常，特征函数是一个二值表征函数，即 $f(x,y)\rightarrow(0,1)$[9]，用来表征特征与类别之间是否相关，特征函数的选择在文本分类任务中尤其重要。可是，在使用最大熵进行文本分类时，(0, 1)的取值已经不能满足模型需要，所以选择字数作为特征值，选择“词-类别”的组合作为特征函数，表示如下：

$$f_{(w,c)}(x,y)=\begin{cases}0 & \text{if } y\neq c\\ \dfrac{N(x,w)}{N(x)} & \text{otherwise}\end{cases}$$

其中，$N(x,w)$是词条 w 在文档 x 中的出现次数；$N(x)$是文档 x 中的词条数。就文本分类而言，每一篇文档中的特征都比较多，得到的候选特征集合就比较大，但是特征函数使得当某个词条经常出现在某个类中的时候，该词条对应的特征函数具有较高的特征值，因此选出的特征也是对模型真正有用的特征。

3 特征选择

特征选择的基本任务是如何从多维特征中找出那些最有效的特征。就有效性而言，就是选择一个维数更小的特征子空间最有效地表示文档自身，因此，需要一个定量的准则来衡量特征对分类的有效性。具体来说，就是从一个高维空间中依据某种标准选择一个低维空间，使该低维空间有良好的类别可分性。既然选择有效特征的最终目的是进行文本分类，那么文本分类器错误发生率最小的那组特征子集就是最好的特征。这么一来，特征选择方法的好坏也就取决于文本分类任务中识别率的高低了。

3.1 cut-off 方法

cut-off 是最简单的特征选择方法之一，其实现原理就是在进行特征选择前先设置一个频率阈值，然后统计文本中所有词出现的频率，将那些出现频率小于预定义阈值的特征删除，是零频率遗漏思想的一种泛化。cut-off 算法思想简单、实现容易，也没有高昂的计算代价，在最大熵的很多应用中使用。

不等式最大熵模型也使用 cut-off 方法选择文本特征。cut-off 没有考虑特征之间的相互关系，只是把特征作为独立个体，通过测试其预测力来进行删除或者添加操作，同时也忽略了特征与模型之间的交互作用；而且，它们本身可能就包含不确定性，因此，本文提出使用一种改进的顺序前进式的特征选择算法(ISFS)[10-11]来实现特征筛选任务，而且，由于在建立模型的过程中不等式约束中本身也嵌入了特征选择，因此实现了更好的特征选择功能。

3.2 改进的顺序前进式算法

这是一种自下而上的搜索方法，其实现思想就是每次从未入选的特征中选择一个特征，使得它与已入选的特征组合在一起时所得的紧密度最大，然后再在剩余候选特征集合中选取若干个与此特征关联度最强的特征，并进行判断是否淘汰，一直重复此操作直到特征数增加到某一数量为止[12]。

这里首先形式化特征空间，把所有可能的特征都称为候补特征，然后从这个候补特征集合内选取对模型最为有用的特征集合 G。这里假设候选特征集为 F，已选定特征集为 S，对应权值 λ，p_s 为满足约束的情况下熵最大的模型。

改进的顺序前进式选择算法(improved sequential forward selection)如下：

输入 特征候选集合 F，经验分布 $P(x,y)$

输出 模型选用的特征集合 S，结合这些特征的模型 P_S

1 初始化：S=Φ(空集)；所对应的模型 P_s 均匀分布；n=0

2 对每一个候选特征 f ∈F，计算 gain(f)=L($P_{S\cup\{f\}}$)-L(P_S)

3 选择特征 f_s=max gain(f)

4 按照下面的公式计算 f_s 与剩余候选特征的综合权重：

weight(f)=wordchi*I_x+wordmi*I_{mi}

其中，wordchi 及 wordmi 分别是词条的 CHI 值、MI 值，I_x, I_{mi} 为计算 weight 时的比重，其中 $I_x \geqslant I_{mi}$，并进行排序，选出前 r 个特征，与 f_s 构成集合 G

5 使用 LMVM 算法计算 P_s，调整参数，如果满足终止条件则执行 6，否则，执行 8

6 S=S∪G; F=F-G

7 n=n+r+1

8 Goto 2

通常，一个特征对所处理问题带来的信息越多，该特征越适合引入到模型中，所以该算法进行特征选取时是由特征的信息增益作为衡量标准的，同时也考虑了特征与模型之间的交互作用；由于互信息反映的是特征之间关联度的强弱，因此该算法考虑了特征与特征之间的关系，由此可见，本文提出的算法完全克服了 cut-off 的弱点，而且算法中使用了常用的特征选择评估函数，使得特征的选择更加精准。本文中 r 取 3。

4 试验及结果分析

本文通过文本分类的试验来验证不同的特征选择算法对模型效率的影响。试验采用“路透社-21578”提供的语料库，首先将语料库中已经标注了类别的文本分为训练集和测试集两部分，其中用来做训练数据的文档为 7 048 篇，剩下的 2 991 篇作为测试集，在训练集中出现过的主题类别有 112 个。另外，本文采用的评估指标是国际上通用的分类器评价指标：召回率(R)，精确率(P)以及 F1 测度(F)[13]。

本次试验使用 LMVM 算法对标准最大熵求解，用 BLMVM 算法实现不等式最大熵的求解。而且由于 LMVM 和 BLMVM 属于同一个算法家族同一个代码库，因此就可以在较为公平的基础上对下面的实验成绩进行比较。

表 1 是不等式最大熵模型使用了改进的特征选择算法之后在不同的宽度因子下得到的特征数量和文本分类成绩。

表 1 不等式最大熵模型中控制参数的选择

W	特征数	精度		
		P	R	F
1.00E-04	1 032.4	85.12	80.72	81.31
1.43E-06	2 139.2	87.35	81.02	82.73
1.28E-07	3 836.8	89.91	82.95	83.68
1.75E-09	4 195.8	91.68	83.46	85.72

从表 1 中可以得到如下结论：随着控制参数 W 的值越来越小，活跃特征数逐渐增加，分类准确率逐步提高。所以选择合适的特征选择方法和控制参数对文本分类的影响是非常重要的。

表 2 为不等式最大熵使用了 cut-off 算法和 ISFS 算法之后得到的分类结果，比较结果如下所示。

表 2　2 种特征选择方法在不等式最大熵文本分类中的精度对比

特征选择算法	参数取值	特征子集数目	分类器精度		
			P	R	F
Cut-off	C=3, W=1.68E-11	4 208	90.38	80.74	84.46
		6 439	91.58	81.37	85.98
ISFS	r=4, W=1.68E-11	4 208	92.37	83.72	85.24
		6 439	94.29	86.49	90.22

实验结果表明，新的特征选择算法使得不等式最大熵文本分类的成绩大大提高。正如前文所说，特征选择方法的好坏也就取决于文本分类任务中识别率的高低，本文所提出的特征选择方法使得不等式最大熵分类器的准确率大大提高，因此也证明了该方法的高效性。

5　结束语

为了提高不等式最大熵分类识别算法的可靠性及效率，需要对特征进行合理的选择，以选择出对该分类器而言最有区分力的特征。本文提出的特征选择算法使得不等式最大熵文本分类的成绩有很大的提高，但是也还是需要开发更好的特征选择算法从而发挥不等式最大熵的最大优势。

未来可以朝下面这 2 个方向展开进一步的研究工作：(1)收集更有力的证据来证明不等式最大熵的优越性，比方说用别的自然语言处理任务来评估不等式最大熵；(2)探讨能否使用更好的特征选择算法进一步发挥不等式最大熵在文本分类任务中的优势。

参考文献

[1] Berger A L, Pietra S A D, Pietra V J D. A Maximum Entropy Approach to Natural Language Processing[J]. Computational Linguistic, 1996, 22(1): 39-71.

[2] Ratnaparkhi A. Maximum Entropy Models for Natural Language Ambiguity Resolution[D]. Pennsylvania, USA: University of Pennsylvania, 1998.

[3] Ratnaparkhi A. A Maximum Entropy Model for Part-of-speech Tagging[C]//Proc. of the Conference on Empirical Methods in Natural Language Processing. Pennsylvania, USA: [s. n.], 1996: 133-142.

[4] 李荣陆, 王建会, 陈晓云, 等. 使用最大熵模型进行中文文本分类[J]. 计算机研究与发展, 2005, 42(1): 94-101.

[5] Nigam K, Lafferty L, McCallum A. Using Maximum Entropy for Text Classification[C]//Proc. of Workshop on Machine Learning for Information Filtering. Stockholm, Sweden: [s. n.], 1999: 61-67.

[6] Kazama J, Tsujii J. Maximum Entropy Models with Inequality Constraints: A Case Study on Text Categorization[J]. Machine Learning, 2005, 60(3): 159-194.

[7] Chen S F, Rosenfeld R. A Gaussian Prior for Smoothing Maximum Entropy Models[R]. CMU, Tech. Rep.: CMU-CS-99-108, 1999.

[8] Benson S J, Jorge J. A Limited Memory Variable Metric Method for Bound Constraint Minimization[R]. Argonne National Laboratory, Tech. Rep.: ANL/MCS-909- 0901, 2001.

[9] 贾　宁, 张　全. 基于最大熵模型的中文姓名识别[J]. 计算机工程, 2007, 33(9): 31-33.

[10] 秦　进, 陈笑蓉. 文本分类中的特征抽取[J]. 计算机应用, 2003, 23(2): 45-46.

[11] 贾丽洁. 基于最大熵模型的分词技术研究[D]. 济南: 山东师范大学, 2007.

[12] 许高建, 路　遥, 胡学钢, 等. 一种改进的文本特征选择方法的研究与设计[J]. 苏州大学学报: 工科版, 2008, 28(2): 18-22.

[13] Yang Yiming. An Evaluation of Statistical Approaches to Text Categorization[J]. Information Retrieval, 1999, 1(1): 76-88.

编辑　任吉慧

~~~~~~

(上接第 181 页)

5　结束语

本文设计了 3 种类型的用户满意度评估方法，提出一种基于用户满意度的遗传算法 USGA。通过实验证明，USGA 与 SGA 和 Min-Min 算法相比，能保证较优的调度时间跨度和资源负载均衡，并且在服务质量方面有很大的提高。

参考文献

[1] Foster I, Kesselman C. 网格计算[M]. 2 版. 金　海, 译. 北京: 电子工业出版社, 2004.

[2] Wu Minyou, Wei Shu, Zhang Hong. Segmented Min-Min: A Static Mapping Algorithm for Meta-tasks on Heterogeneous Computing Systems[C]//Proc. of Heterogeneous Computing Workshop. [S. l.]: IEEE Press, 2000.

[3] Song Shanshan, Kwok Yu-Kwong, Hwang Kai. Security-driven Heuristics and a Fast Genetic Algorithm for Trusted Grid Job Scheduling[C]//Proceedings of the 19th IEEE International Parallel and Distributed Processing Symposium. Denver, CA, USA: IEEE Press, 2005.

[4] 王天擎, 谢　军, 曾　洲. 基于蚁群算法的网格资源调度策略研究[J]. 计算机工程与设计, 2007, 28(15): 60-61.

[5] 陈国良, 王东生. 遗传算法及应用[M]. 北京: 人民邮电出版社, 2001.

[6] 林剑柠, 吴慧中. 基于遗传算法的网格资源调度算法[J]. 计算机研究与发展, 2004, 41(12): 2195-2199.

[7] Buyya R, Murshed M. GridSim: A Toolkit for the Modeling and Simulation of Distributed Resource Management and Scheduling for Grid Computing[J]. The Journal of Concurrency and Computation: Practice and Experience, 2002, 14(13-15): 121-134.

编辑　张正兴
~~~~~~