改进的最大熵权值算法在文本分类中的应用

李学相
(郑州大学软件技术学院 郑州 450002)

摘 要 由于传统算法存在着特征词不明确、分类结果有重叠、工作效率低的缺陷,为了解决上述问题,提出了一种改进的最大熵文本分类方法。最大熵模型可以综合观察到的各种相关或不相关的概率知识,对许多问题的处理都可以达到较好的结果。提出的方法充分结合了均值聚类和最大熵值算法的优点,算法首先以香农熵作为最大熵模型中的目标函数,简化分类器的表达形式,然后采用均值聚类算法对最优特征进行分类。经过实验论证,所提出的新算法能够在较短的时间内获得分类后得到的特征集,大大缩短了工作的时间,同时提高了工作的效率。

关键词 文本分类,最大熵算法,均值聚类,特征选择
中图法分类号 TP391 **文献标识码** A

Research of Text Categorization Based on Improved Maximum Entropy Algorithm

LI Xue-xiang
(School of Software Technology, Zhengzhou University, Zhengzhou 450002, China)

Abstract This paper discussed the problems in text categorization accuracy. In traditional text classification algorithm, different feature words have the same affecte on classification result, and classification accuracy is lower, causing the increase algorithm time complexity. Because the maximum entropy model can integrated various relevant or irrelevant probability knowledge observed, the processing of many issues can achieve better results. In order to solve the above problems, this paper proposed an improved maximum entropy text classification, which fully combines c-mean and maximum entropy algorithm advantages. The algorithm firstly takes shannon entropy as maximum entropy model of the objective function, simplifies classifier expression form, and then uses c-mean algorithm to classify the optimal feature. The simulation results show that the proposed method can quickly get the optimal classification feature subsets, greatly improve text classification accuracy, compared with the traditional text classification.

Keywords Text classification, Maximum entropy algorithm, C-mean, Feature selection

1 引言

随着计算机以及网络技术的不断发展,许多文本信息存在于现实中,而且数量一直持续不断地递增,所以如何对大量的数据进行有效的分类是一项非常艰巨的工作,也是目前一项重要的研究课题。文本分类是文本数据挖掘以及信息检索重要基础,当前,文本分类技术已经应用非常广泛,目前最常用的文本分类方法主要包括基于支持向量机分类器、贝叶斯方法、神经网络、模糊聚类等,这些方法通常对文本的分类精确度不高,而且过分依赖分类的模型,由于最大熵算法模型可以综合分类模型,不依赖于单一模型,因此对许多问题的处理都可以得到较好的结果。由于其应用范围较为广泛,因此成为较为热门的学科。

本文针对传统的文本分类算法存在着各特征词对分类的结果影响相同、分类准确率较低的缺陷,同时造成了算法时间复杂度的增加,提出了一种新的基于 C-均值聚类和最大熵算法相结合的文本分类方法。该方法充分结合了均值聚类和最大熵值算法的优点,首先以香农熵作为最大熵模型中的目标函数,简化分类器的表达形式,然后采用均值聚类算法对最优特征进行分类。经过实验论证,本文提出的新算法能够在较短的时间内获得分类后得到的特征集,大大缩短了工作的时间,同时提高了工作的效率,值得推广应用。

2 文本分类原理

图 1 所示是文本分类系统的主要原理。

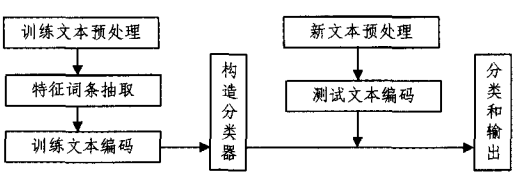

图 1 文本分类系统结构

向量空间模型是文本信息处理的一个主要特征。向量空间模型的基本思想是以向量的形式来表示文本,例如相似度公式:

到稿日期:2011-08-29 返修日期:2011-11-12 本文受国家高技术研究发展计划(2007AA010408)资助。
李学相(1965—),男,博士,副教授,主要研究方向为计算机软件与理论,E-mail:lxx@zzu.edu.cn。

$$\mathrm{Sim}(d_i C_j)=\frac{\sum_{k=1}^{m}(w_{jk}\times W_{ik})}{\sqrt{\sum_{k=1}^{m}w_{jk}^2}\times\sqrt{\sum_{k=1}^{m}W_{ik}^2}}\sum_{k=1}^{m}w_{jk}\times W_{ik} \tag{1}$$

式中，d_i 是需要处理的文本，范畴在 C_j 类中，$j=1,\cdots,m$，d_i，W_k 为 d_i 所属的类 C_j 的第 k 个特征项的权重，n 为 C_j 包含的特征项的数目。

3 基于均值聚类和最大熵文本分类

3.1 最大熵权值训练

著名科学家 E. T. Jaynes 提出了最大熵算法的概念，它的主要流程就是按照最大熵原则，在有限信息条件下选择出最客观的权值。

假设存在这样的一个训练样本，将这些集合定为 $\{x_1, y_1\}$，$\{x_2, y_2\}$，$\{x_3, y_3\}$，其中 $x_i(1\leqslant i\leqslant N)$ 表示在这些集合范围内的文本，而 $y_i(1\leqslant i\leqslant N)$ 就表示这些文本所对应的结果。按照经验公式操作这些训练样本，从而得到这些 (x,y) 的经验分布，如式(2)所示：

$$\tilde{P}(x,y)=\frac{1}{N}\times count \tag{2}$$

式中，$count$ 表示样本在 (x,y) 出现的次数。

统计样本得到统计后的集合数据，然后建立 N 的样本模型，在这个模型中引进特征建立函数，可以使上下文的信息对模型有一定的依赖性[8]。假设 f_i 表示特征函数的限制条件，则有：

$$p(f_i)=p(f_i), i\in\{1,2,\cdots,n\} \tag{3}$$

则训练样本的期望概率值为：

$$p(f)=\sum_{x,y}\tilde{p}(x)p(y|x)f(x,y) \tag{4}$$

训练样本的经验值为：

$$\tilde{p}(f)=\sum_{x,y}\tilde{p}(x)f(x,y) \tag{5}$$

根据分布一致的模型具备的条件，就可以获得最佳的 $p(y|x)$ 值，同时可以确定衡量最优的标准，这里取最优的条件熵，记作：

$$H(p)=-\sum_{x,y}\tilde{p}(x)p(y|x)\log p(y|x) \tag{6}$$

由条件熵应该满足如下限制：

$$f(x,y)=0, \sum_{y}p(y|x)=1\ \forall x \tag{7}$$

$$P(y|x)\geqslant 0, \forall x, \forall y \tag{8}$$

$$\sum_{x,y}p(x)p(y|x)f(x,y)=\sum_{x,y}p(x,y)f(x,y) \tag{9}$$

$$i\in\{1,2,\cdots,n\}$$

同时将每一个特征 f_i 与一个拉格朗日算子相对应，再对每一个具体的实例 x 增加一个参数 $k(x)$，根据上述的几个已存在的条件，将熵的最大值拉格朗日函数 $\Lambda(p,\lambda)$ 记作：

$$\Lambda(p,\lambda)=H(p)+\sum_i\lambda_i p(f_i)-p(f_i)+\sum_x k(x)(\sum_y p(y|x)-1) \tag{10}$$

采用 $p\lambda(y|x)$ 来表示 $\Lambda(p,\lambda)$ 最大时的分布 $p(y|x)$，则有：

$$p\lambda(y|x)=\frac{\exp\sum_x\lambda_i f_i(x,y)}{Z_\lambda(x)} \tag{11}$$

式中，$Z_\lambda(x)$ 表示归一化因子，归一化公式为：

$$Z_\lambda(x)=\sum_y\exp\sum_i\lambda_i f_i(x,y) \tag{12}$$

根据其特征 $\{f_i, 1\leqslant i\leqslant n\}$，就可以对其经验分布 $\tilde{p}(x)$ 和 $\tilde{p}(y|x)$ 进行统计，在本文中采用均值聚类算法，这样可以在较短的时间内获取特征参数，同时进行分类。

3.2 均值聚类分类

在预处理之后，某一指定的文档集合就可以成为文档矩阵，设定在这个文档集合中存在 n 个文档，每一个文档中出现的特征词有 s 个，而每一个特征词出现的概率为 c，那么每一个特征词的权值为 w_j，且满足条件 $\sum_{j=1}^{s}w_j=1$，那么就可以将 C 均值算法加权模糊的函数定义为：

$$J(u,v,w)=\sum_{i=1}^{c}\sum_{k=1}^{n}\sum_{j=1}^{s}u_{ik}^m w_j^\beta(x_{kj}-v_{ij})^2 \tag{13}$$

式中，$\sum_{t=1}^{c}u_{it}=1, 1\leqslant i\leqslant c, 1\leqslant j\leqslant m$，$c$ 是类数目，β 称为权重指数。

根据拉格朗日乘子法，又可以获得最小化目标函数 $J(u,v,w)$ 的必要条件如下：

$$v_{ij}=\sum_{t=1}^{n}u_{ik}^m x_{kj}/\sum_{t=1}^{n}u_{ik}^m \tag{14}$$

$$u_{ik}=(\sum_{j=1}^{s}w_j^\beta\|x_{kj}-v_{ij}\|^2)^{\frac{1}{(1-m)}}\times(\sum_{t=1}^{c}(\sum_{j=1}^{s}w_j^\beta\|x_{kj}-v_{ij}\|)^{\frac{1}{(1-m)}})^{-1} \tag{15}$$

$$w_j=D_j^{1/(1-\beta)}(\sum_{t=1}^{m}D_t^{1/(1-\beta)})^{-1}$$

式中，$D_j=\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}^m(x_{kj}-v_{ij})^2$ (16)

整个算法的主要流程描述如下：

第一步　选择符合实验要求的训练样本文本作为实验对象。

第二步　根据文本提供的上下文信息，并结合已知的地理条件，建立出符合特征要求的模板。

第三步　计算出本文所需的最大熵函数，并验证其结果是否正确。

第四步　初始化分类函数，确定矩阵 U,W，聚类数量 C 等各方面的特征值。对取得的特征值进行验证，查看其是否收敛。

第五步　根据已有的条件，通过函数得出最终的数值，加入得到的数值，若其大于迭代次数，则算法停止，否则代入迭代公式进行修正参数，然后重新回到步骤二，进行进一步的信息提取和特征模板的制定和校验工作。

4 仿真实验与结果分析

4.1 数据来源

根据上面阐述的要求，本文搭建了 MATLAB7.0 的实验环境，实验选取 windowsXP、奔腾双核处理器。在本文所用到的实验数据均是采用某一个高校的文本库，还包含了来自互联网的文本数据。在这些文本中大致含有 20 个类别的文本，主要是用来进行测试和训练的，在去除一些可以忽视的因素外，得到了没有任何重复、特征比较明显的文本，数量是 14378 本，但是在这个文档库中每一个类别没有标清楚，存在着属于经济类时还涵盖其他的类别等情况。为了对比不同方法下本文方法的优势，选取了当中 10 个类别的文档，同时将这些文档整理归档如表 1 所列。

表 1　文本的类分布

类别	文件数	类别	文件数
游戏	22843	旅游	18471
经济	40115	文艺	14248
科技	53126	时政_国际	59130
房产	19573	时政_国内	119695
汽车	21745	教育	24405
体育	96120	生活男女	19382
娱乐	23905	时政_社会	42559
时政_军事	21743	总计	597060

本文以“经济”为例计算了该分类的准确度，如表 2 所列。

表 2　以“经济”为例的准确率

领域	正确词数	抽取到的总词数	准确率
经济	962	1000	96.2%
	1916	2000	95.8%
	2870	3000	95.6%
	3814	4000	95.3%
	4737	5000	94.7%

从表 2 中可以看出，本文提出的算法文本分类准确率都高达 95%以上，为了更进一步说明本文算法比传统的算法优越，我们比较了常见的支持向量机文本分类方法、遗传算法、聚类算法和传统的最大熵文本分类算法，比较结果如表 3 所列。

表 3　不同算法的性能比较

	时间(s)	召回率(%)	正确率(%)
聚类算法	47.52	82.32	90.43
遗传算法	26.31	89.12	93.22
支持向量机方法	24.34	90.13	93.32
传统的最大熵算法	20.12	93.21	96.75

从表 3 中可以看出本文所提算法无论在时间上还是正确率都具有一定的优越性。

结束语　将许多的文本资料根据属性或者内容划分到相对应的类别中的过程是文本分类，这在日常生活中比较普遍，但是传统的文本算法都存在着特征词不清楚、分成的类别有重叠的结果的缺陷。针对这种效率不高的文本算法，本文提出了一种新算法，其采用了最大熵算法和均值聚类算法，并将这两种算法相结合，可以降低算法的复杂度。较之传统算法而言，它可以在较短的时间内获得每一个特征子集，准确率得到了很大的提高。

参 考 文 献

[1] Huang Zhe-xue, Michael K N, Rong hong-qiang, et al. Automated Variable Weighting in k-Means Type Clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(5): 657-668

[2] 刘斌，黄铁军，程军，等. 一种新的基于统计的自动文本分类方法[J]. 中文信息学报，2002，16(6)：18-24

[3] 朱明，王俊普. 一种最优特征集的选择算法[J]. 计算机研究与发展，1998，9(35)：803-805

[4] 胡佳妮，徐蔚然，郭军，等. 中文文本分类中的特征选择算法研究[J]. 光通信研究，2005(3)：44-46

[5] 潘有能. 一个自动分词分类系统的实现[J]. 情报学报，2002，21(1)：38-41

[6] 邓乃扬，田英杰. 数据挖掘中的新方法：支持向量机[M]. 北京：科学出版社，2004

[7] 巩知乐，张德贤，胡明明. 一种改进的支持向量机的文本分类算法[J]. 计算机仿真，2009，26(7)：165-168

[8] 解冲锋，李星. 基于序列的文本自动分类算法[J]. 软件学报，2002，13(4)：783-788

[9] 韩家炜，孟小峰，王静，等. Web 挖掘研究[J]. 计算机研究与发展，2001，38(4)：405-411

[10] 杜树新，吴铁军. 模式识别中的支持向量机方法[J]. 浙江大学学报：工学版，2003，37(5)：521-527

[11] 罗印升，李人厚，张雷，等. 人工免疫算法在函数优化中的应用[J]. 西安交通大学学报，2003，7(8)：840-843

[12] 王国才，张聪. 一种基于粗糙集的特征加权朴素贝叶斯分类器[J]. 重庆理工大学学报：自然科学版，2010，24(7)：86-90

(上接第 190 页)

结束语　本文介绍了不平衡调节因子和模糊隶属度形成的一种平衡模糊支持向量机处理高不平衡比的数据集的分类问题。模糊隶属度考虑了每个样本对分类超平面的贡献率，从所构建的隶属度的定义可以看出，这种模糊隶属度对一些野点或者异常点的影响具有很好的消除作用。当正类与负类的样本数相差较大时，引入类不平衡调节因子可以根据样本点的分散程度、样本数的多少进行调节，这对不平衡数据集的分类问题具有很好的作用。总之，平衡调节模糊支持向量机较好地处理了不平衡数据集的分类问题，提高了分类精度，但也消耗了大量的时间。

参 考 文 献

[1] Cortes C, Vapnik V. Support vector networks[J]. Machine Learning, 1995, 20(3): 273-297

[2] Vapnik V N. The nature of statistical learning theory [M]. New York: Springer, 1995

[3] Japkowicz N, Stephen S. The Class Imbalanced Problem: A Systematic Study[J]. Intelligent Data Analysis, 2002, 6(5): 429-449

[4] 邹权，郭茂祖，刘扬，等. 类别不平衡的分类方法及在生物信息学中的应用[J]. 计算机研究与发展，2010，47(8)：1407-1414

[5] 吴洪兴，彭宇，彭喜元. 适用于不平衡样本数据处理的支持向量机方法[J]. 电子学报，2006，34(12)：2395-2398

[6] Zeng Zhi-qiang, Gao Ji. Improving SVM Classification with Imbalance Data Set [J]. Neural Information Processing, 2009, 5863: 389-398

[7] 刘万里，刘三阳，王金艳. 不平衡支持向量机的调整方法[J]. 计算机科学，2009，36(3)：148-150

[8] 陶晓燕，姬红兵，马志强. 基于样本分布不平衡的近似支持向量机[J]. 计算机科学，2007，34(5)：174-176

[9] Veropoulos K, Campbell C, Cristianini N. Controlling the sensitivity of support vector machines[C]// Proc. Int. Joint Conf. Artif. Intell. Stockholm, Sweden, 1999: 55-60

[10] Batuwita R, Palade V. FSVM-CIL: Fuzzy Support Vector Machines for Class Imbalance Learning[J]. IEEE Transactions on Fuzzy Systems, 2010, 18(2): 558-572

[11] 袁亚湘. 最优化理论与方法[M]. 北京：科学出版社，1997

[12] Huang Kai-zhu, Yang Hai-qin, King I, et al. The Minimum Error Minimax Probability Machine[J]. Journal of Machine Learning Research, 2004, 5(10): 1253-1286

[13] Asuncion A, NewmanD J. UCI repository of machine learning databases [EB/OL]. Http://www. Ics. Uci edu/mlear