基于改进的最大熵均值聚类方法在文本分类中的应用*

张爱科
(柳州职业技术学院, 广西 柳州 545006)

摘 要: 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、造成算法时间复杂度增加的问题,提出了一种改进的最大熵 C-均值聚类文本分类方法。该方法充分结合了 C-均值聚类和最大熵值算法的优点,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 C-均值聚类算法对最优特征进行分类。仿真实验结果表明,与传统的文本分类方法相比,提出的方法能够快速得到最优分类特征子集,大大提高了文本分类准确率。

关键词: 文本分类; 最大熵; C-均值聚类; 特征选择
中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2012)04-1297-03
doi:10.3969/j.issn.1001-3695.2012.04.026

Application of text categorization based on improved maximum entropy means clustering algorithm

ZHANG Ai-ke
(*Liuzhou Vocational Technological College*, *Liuzhou Guangxi* 545006, *China*)

Abstract: In view of the traditional text classification algorithm has the problems of the characteristics having same influence on classification results, the low rate of classification accuracy, and the increasing of the algorithm time complexity, this paper presented an improved maximum entropy C-means clustering text classification methods. This method combined the C-means clustering algorithm and the maximum entropy algorithm, set Shannon entropy as a maximum entropy model in the target function, simplified classifier forms of expression, and then used the C-means clustering algorithm to the optimal features for classification. The simulation results show that, compared with traditional text classification methods, the proposed method can fast obtain the optimal classification feature subset, greatly improve the accuracy of text classification.

Key words: text classification; maximum entropy; C-means clustering; feature selection

随着 Web 的迅速发展,大量的文本信息存在于互联网中,如何有效地利用这些海量信息,已经成为当前数据挖掘领域研究的热点问题。文本自动分类简称文本分类(text categorization),是模式识别与自然语言处理密切结合的研究课题,是信息检索和文本挖掘的重要基础。自动文本分类技术已经应用于信息过滤、信息检索、搜索引擎、网络论坛、数字图书馆、邮件分类等多个领域[1]。目前常用的文本分类方法包括朴素贝叶斯方法(naive Bayesian classifier)[2]、基于支持向量机的分类器[3]、K-最近邻法[4~6]和 Boosting 分类法[7]等,这些方法普遍存在分类精度较低的问题。最大熵模型具有简洁、通用和易于移植等特性,已广泛应用于自然语言的文本分类中。文献[8,9]就提出了采用最大熵模型对文本进行分类。最大熵模型可以使用多种特征,且各个特征之间没有独立性假设,因而它的表达能力非常强,但该模型具有训练速度慢、资源消耗大的问题。为了能更好地提升最大熵模型的分类精度和速度,本文把 C-均值聚类引入到最大熵模型中,结合两者的优点,提出了一种改进的最大熵 C-均值聚类(improved maximum entropy C-means clustering algorithm, IMECA)文本分类方法。

1 文本分类原理

文本分类系统的任务是在给定的分类体系下,根据文本的内容自动地确定文本关联的类别。自动文本分类即根据统计模式识别思想,将文本表示成特征向量,然后用训练文本对事先选定的分类器进行训练,直接或间接地提取出蕴涵在训练文本中有关各个文本类的统计特性,并根据这些特性确定出分类准则,最后依据这些准则对未知文本进行分类决策。其系统结构如图 1 所示。

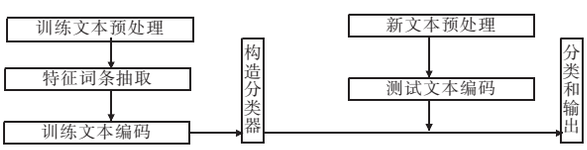

图 1 文本分类系统结构

在文本信息处理问题上,文本的表示主要采用向量空间模型(vector space model, VSM)[10]。其基本思想是以向量的形式来表示文本,如相似度公式:

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

其中:d_i 属于 $C_j(j=1,2,\cdots,m)$,d_i 表示要处理的文本;C_j 是 d_i 所属的类;m 是分类体系的类别数;W_k 为 d_i 所属的类 C_j 的第 k 个特征相的权重;n 为 C_j 包含的特征项的数目。

收稿日期: 2011-09-14; **修回日期**: 2011-10-21 **基金项目**: 广西教育厅科研项目基金资助项目(200911LX486,201106LX745)
作者简介:张爱科 (1973-),女(壮族),副教授,硕士,主要研究方向为数据挖掘、演化计算(viewfish@qq.com).

2 基于 C-均值聚类和最大熵文本分类

2.1 最大熵权值训练

最大熵算法最早是由 Jaynes 提出的,在根据部分信息进行推理时,应该选择的概率分布必须是服从所有已知观测数据的前提下使熵取得最大值的概率分布,这就是著名的最大熵原理(maximum entropy principle)[7]。其主要思想是按照使 x 的熵达到最大的原则来选择其概率分布 p。如果所选择的概率分布使得 x 的熵小于最大值,那么必然在求解的过程中有意或无意地添加了一些假设信息,而这些假设信息通常是没有依据的。所以,按照最大熵原则所得到的估计是在有限信息条件下最客观、最小偏见的选择。

假设有一个训练样本集合为 $\{x_1,y_1\},\{x_2,y_2\},\{x_3,y_3\}$,其中每一个 $x_i(1\leqslant i\leqslant N)$ 表示一个上下文信息,那么 $y_i(1\leqslant i\leqslant N)$ 就表示对应的结果。对于此训练样本,可以通过经验分布公式获得 (x,y) 的经验分布,其公式描述如下:

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

其中:count 表示样本在 (x,y) 出现的次数。

通过样本集合的统计数据,为上述 N 个训练样本集合建立统计模型。在模型中引入特征函数,从而使模型对上下文的信息产生依赖。假设 f_i 表示特征函数的限制条件,则有

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

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

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

获得一个最为一致的分布的模型条件,得到最优的 $p(y|x)$ 值,则条件熵作为衡量最优的标准,将其定义为

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

由于条件熵应该满足如下限制:

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

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

$$\sum_{x,y}p(x)p(y|x)f(x,y)=\sum_{x,y}p(x,y)f(x,y)\quad i\in\{1,2,\cdots,n\} \tag{9}$$

为每一个特征 f_i 引入一个拉格朗日算子,且为每个实例 x 引入一个参数 $k(x)$,求出条件熵的最大值拉格朗日函数 $\Lambda(p,\lambda)$,该函数定义如下:

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

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

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

其中:$Z_\lambda(x)$ 表示归一化因子,归一化公式为

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

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。本文采用模糊 C-均值聚类算法来计算特征参数并进行分类。

2.2 模糊 C-均值聚类分类

某文档集合经过预处理之后,可以表示为文档矩阵[8]。假设集合中包含 n 篇文档,每一文档用 s 个主题词的出现频率表示,c 表示类数目,用 w_j 来定义第 j 个主题词的权值,且满足 $\sum_{j=1}^{s}w_j=1$,则属性加权模糊 C-均值算法的目标函数可定义为

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

其中:$\sum_{k=1}^{c}u_{ik}=1,1\leqslant i\leqslant c,1\leqslant j\leqslant m$;$c$ 是类数目;β 是权重指数。

通过拉格朗日乘子法,可以得到最小化目标函数 $J(u,v,w)$ 的必要条件如下[11]:

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

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

$$w_j=D_j^{1/(1-\beta)}\Big(\sum_{j=1}^{m}D_t^{1/(1-\beta)}\Big)^{-1} \tag{16}$$

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

2.3 IMECA 的实现流程

IMECA 实现流程主要步骤如下:

a)输入文本,训练样本的选择。

b)特征模板的建立。根据文本给定信息和上下文的特定位置,建立训练特征模板。

c)构建最大熵目标函数。

d)选定聚类数目 c、模糊指数 m、最大迭代次数 T_{count}、属性权值指数 β 和阈值 ε,初始化划分矩阵 U 和属性权值矩阵 W。

e)根据目前的划分和属性权值,计算目标函数的值,如果大于迭代次数 T_{count},或它相对于上次价值函数的改变量小于阈值 ε,算法停止。

f)使用式(13)修正聚类中心 v_{ij}。

g)使用式(16)更新属性权值矩阵 W,返回步骤 b)。

3 仿真实验与结果分析

3.1 数据来源

为了验证 IMECA 的有效性,在 Pentium® 4 2.4 GHz、1 GB 内存的 PC 机上进行了大量的仿真实验,实验环境为 MATLAB7。仿真实验的数据来自高校文本分类库,该文本全部采自互联网,包含 20 个文本类别,分为训练文档集和测试文档集两个部分,去除部分重复文档和损坏文档后,得到无重复、完整文档 14 378 篇,其中训练文档 8 214 篇,测试文档 6 164 篇,每一篇文档只属于一个类别。该语料库中文档的类别分布情况是不均匀的。其中,训练文档最多的经济类有 1 369 篇,而训练文档最少的通信类有 25 篇;同时,训练文档数少于 100 篇的稀有类别共有 11 个。训练文档集和测试文档集之间互不重叠。只取前 10 个类的部分文档,其类别文档统计数如表 1 所示。

表 1 文本的类分布

类别	文件数	类别	文件数
游戏	22 843	旅游	18 471
经济	40 115	文艺	14 248
科技	53 126	时政_国际	59 130
房产	19 573	时政_国内	119 695
汽车	21 745	教育	24 405
体育	96 120	生活男女	19 382
娱乐	23 905	时政_社会	42 559
时政_军事	21 743	总计	597 060

本文以“经济”为例计算了该分类的准确度,如表 2 所示。

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	1 000	96.2
	1 916	2 000	95.8
	2 870	3 000	95.6
	3 814	4 000	95.3
	4 737	5 000	94.7

从表 2 中可以看出,本文提出的 IMECA 文本分类准确率均高达 95% 以上,为了更进一步地说明 IMECA 比传统的算法优越,比较了常见的支持向量机文本分类方法、KNN 算法和传统的最大熵文本分类算法,仿真实验结果如图 2、3 所示。图 2 显示的是五种算法不同特征数下的召回率对比,图 3 显示的是不同特征数下的准确率对比。

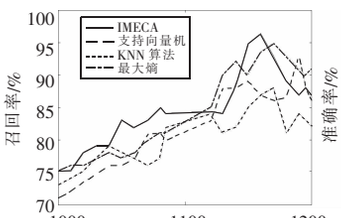

图 2　不同算法的召回率对比

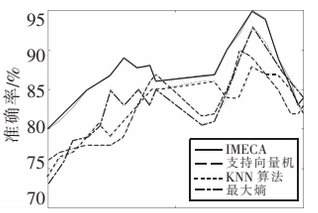

图 3　不同算法的准确率对比

3.2　结果分析

由图 2 和 3 可知,常用的文本分类算法性能相当,本文提出的 IMECA 无论在准确率还是召回率上,均具有一定的优越性。此外,从图中也可看出,并不是特征数越多,分类效果就越好。分类的效果会随着特征数的增加而变化,当特征数达到一定的数量,分类效果反而还随着特征数的增加而下降。因此,在进行特征选择时,特征数的选择要适当。如果特征项维数过高,则可能会有一些无用的干扰项,从而影响分类的准确率;但如果特征数较少,就会漏掉一些对分类有较大贡献的项。所以,对于每个给定的训练集而言,应该通过实验来确定最佳特征数。

4　结束语

文本分类是指根据文档的内容或属性,将大量的文本归到一个或多个类别的过程,在日常生活中具有非常重要的意义。本文对传统的文本分类算法进行了深入分析,结合最大熵模型和模糊 C-均值聚类算法,提出一种改进的基于最大熵 C-均值聚类算法相结合的文本分类方法。仿真实验结果证明,与传统的文本分类方法相比,提出的方法能够快速得到最优分类特征子集,大大提高了文本分类的准确率,具有一定的应用前景。

参考文献:

[1] FORMAN G. An extensive empirical study of feature selection metrics for text classification[J]. Journal of Machine Learning Research, 2003,3(1):1289-1305.

[2] 余芳,姜云飞. 一种基于朴素贝叶斯分类的特征选择方法[J]. 中山大学学报:自然科学版,2004,43(5):118-120.

[3] 刘良斌,王小平. 基于支持向量机和输出编码的文本分类器研究[J]. 计算机应用,2004,24(8):32-34.

[4] 张文良,黄亚楼,倪维健. 一种基于聚类的文本特征选择方法[J]. 计算机应用,2007,27(1):205-206,209.

[5] 孙荣宗,苗夺谦,卫志华,等. 基于粗糙集的快速 KNN 文本分类算法[J]. 计算机工程,2010,36(24):175-177.

[6] 鲁婷,王浩,姚宏亮. 一种基于中心文档的 KNN 中文文本分类算法[J]. 计算机工程与应用,2011,47(2):127-130.

[7] 董乐红,耿国华,周明全. 基于 Boosting 算法的文本自动分类器设计[J]. 计算机应用,2007,27(2):384-386.

[8] 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.

[9] BERGER A L,PIETRA V J D,PIETRA S A D,*et al*. A maximum entropy approach to natural language processing[J]. Computational Linguistics,1996,22(1):39-71 .

[10] 刘少辉,董明楷,张海俊,等. 一种基于向量空间模型的多层次文本分类方法[J]. 中文信息学报,2001,16(3):8-14,26.

[11] SONG Feng-xi,LIU Shu-hai,YANG Jing-yu. A comparative study on text representation schemes in text categorization[J]. Pattern Analysis and Applications,2005,20(8):199-209.

(上接第 1248 页)外部协同、内部协同都是以信息协同为基础,如果信息不协同,整个企业将会处于低效而混乱不堪的境况。

4　结束语

现代物流系统协同研究从不同角度分析了物流系统协同问题,物流企业不仅需要外部合作单位、竞争对手、上下游厂商间的协同运作,更需要企业内部管理层、战略层、技术层等多方面协同运作。而所有这些复杂的协同过程,都是基于一个关键影响因素——信息协同,也就是本文研究的决定复杂物流系统的序参量。本文通过建立和求解序参量方程,得出以下结论:a)信息协同是复杂物流系统的序参量,决定着系统的涨落和平衡点;b)通过调整信息在系统内外部的共享性和开放性,可以提高系统的运作效率;c)物流企业外部协同和内部协同的效果,取决于信息协同的程度,信息协同是复杂物流信息系统的序参量,决定着系统的平衡性。

本文从较为浅显的层次分析了复杂物流系统的协同问题,一些数学公式及推论尚存在不妥之处,有待进一步研究改善。

参考文献:

[1] HAKEN H. The secret of the contracture of nature[M]. Oxford: Oxford University Press,2005:8-9.

[2] HAKEN H. 协同学引论[M]. 徐锡申,等译. 北京:原子能出版社,1984.

[3] MALONI M J,BENTON W C. Supply chain partnerships: opportunities for operations research[J]. European Journal of Operational Research,1997,101(3):419-429.

[4] LAU J S K, HUANG G Q, MAK K L, *et al*. Distributed project scheduling with information sharing in supply chains, part I: an agent based negotiation model[J]. International Journal of Production Research,2005,43(22):4813-4838.

[5] 潘开灵,白列湖. 管理协同理论及其应用[M]. 北京:经济管理出版社,2006.

[6] ANSOFF H I. Corporation strategy[M]. [S. l.]: McGraw-Hill Inc, 1965.

[7] HAKEN H. 信息与自组织[M]. 郭治安,译. 成都:四川人民出版社,1988.

[8] HAKEN H. 协同学原理及应用[M]. 吴大进,译. 武汉:华中理工大学出版社,1990.

[9] 朱涛,常国岑,张水平,等. 基于复杂网络的指挥控制信息协同模型研究[J]. 系统仿真学报,208,20(22):6058-6060,6065.

[10] 郑德俊. 基于信息生态理论的企业危机信息预警策略研究[J]. 科技与管理,2010(1):40-44.

[11] 张向先,国佳,马捷. 企业信息生态系统的信息协同模式研究[J]. 情报理论与实践,2010(4):10-13.