# 基于特征聚合与最大熵的文本分类算法

陈 光 刘宗田
(上海大学计算机工程与科学学院 上海 200072)

**摘 要** 网络信息浩如烟海又纷繁芜杂,从中掌握最有效的信息是信息处理的一大目标,而文本分类是组织和管理数据的有力手段。由于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,具有对许多问题的处理都可以达到较好的结果的优势,将最大熵模型引入到中文文本分类的研究中,并通过使用一种特征聚合的算法改进特征选择的有效性。实验表明与 Bayes、KNN和 SVM这三种性能优越的算法相比,基于最大熵的文本分类算法可取得较之更优的分类精度。

**关键词** 文本分类 最大熵模型 特征选取

# TEXT CLASSIFICATION BASED ON MAXIMUM ENTROPY AND FEATURE AGGREGATION

Chen Guang Liu Zongtian
(School of Computer Engineering and Science, Shanghai University, Shanghai 200072, China)

**Abstract** The Internet has become the main source for people to get various information. Text classification has become the key technology in document data organization and processing. Maximum Entropy Model, a probability estimation technique widely used for a variety of natural language tasks, is used for text classification. A feature aggregation algorithm is used to select efficient feature. The experimental results show that compared with Bayes, KNN and SVM, the proposed text classification algorithm achieves better performance.

**Keywords** Text classification Maximum entropy model Feature selection

## 0 引 言

随着 Internet的迅速发展,互联网上的信息极大丰富,已使它成为全球最大的分布式信息库。目前绝大多数信息均表现为文本方式,如何在浩如烟海而又纷繁芜杂的文本中掌握最有效的信息始终是信息处理的一大目标。由于分类可以在一定程度上解决网上信息杂乱的现象,方便用户准确地定位所需的信息和分流信息,因此,文本自动分类已成为一项具有较大实用价值的关键技术,是组织和管理数据的有力手段。

## 1 文本分类算法

文本自动分类的一个关键问题是如何构造分类函数(分类器),并利用此分类函数将待分类文本划分到相应的类别空间中。文本分类系统中一般采用 VSM模型,每篇文章都根据所有不同的分布模式构造成一个 n维的向量。一般采用词频或布尔值作为特征词的特征值。训练方法和分类算法是分类系统的核心,目前存在多种基于向量空间模型的训练算法和分类算法,例如:支持向量机算法,神经网络方法,KNN邻居方法和贝叶斯方法等[1]。本文将尝试采用特征词聚合与最大熵的分类算法集合的方法提高文本分类效果。

目前提取特征词的方法主要有 7种:互信息、期望交叉熵、信息增益、文本证据权、几率比、词频法和 $X^2$ 概率统计。针对 Reuters 21578和 OHSUMED等数据集合的一个测试结果证实 $X^2$ 概率统计通常比其他特征词提取方法优越。$X^2$ 的主要思想是认为词与类别之间符合 $X^2$ 分布,$X^2$ 统计量的值越高,词和类别之间的独立性越小、相关性越强,词对这一类别的贡献越大。一般取单词在所有类别中的平均值或所有类别中的最大值为其 $X^2$ 值。

## 2 基于特征聚合的特征选取

本文提出了基于特征词聚合的特征选取方法。对于每个特征词 t都能得到 C(类别总数)个 $X^2$ 值,表示为 $\chi^2_{(t,i)}$,(1≤ i≤ C)。于是,每个词可得到一条其对分类的贡献分布曲线。可认为那些有相同分布曲线的特征词是相互关联的,对分类有相同的贡献。通过下面的分析,我们希望能得到一种通用方法,可以将分类贡献分布相同的词聚合为一个模式。从而最终得到新的 m个不同的分布模式(m< N),使用单个模式作为文本向量的一个维。我们用于测试的文档有 8个大类,即 C=8,则特征词可得到 8个 $X^2$ 值。通过对词的 $X^2$ 对各类的分类贡献。我们知道"达芬奇"和"徐悲鸿"两词的贡献分布曲线基本重合,而"达芬奇"和"徐悲鸿"两个词几乎都只在艺术类中出现,同样"刘翔"和"姚明"只在体育类中出现,即认为它们只对某一类有贡献,如图 1所示。

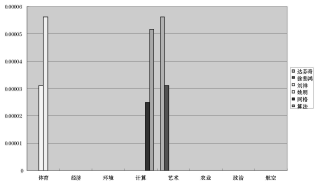

图 1 特征分布图

收稿日期:2006-11-15。陈光,硕士生,主研领域:智能信息处理,数据挖掘,自然语言处理。

通过分析8000个特征词的$X^2$分布曲线,可知每个词最大$X^2$值变化范围是10.5~66.93,非最大值大部分$X^2$值大部分低于4。我们将八个$X^2$作为该词的一个分布相关性向量$t(X_1^2, \cdots, X_8^2)$对于两词$w_1$和$w_2$定义如下的聚合条件,sim为相似度函数,$t_1$和$t_2$为$w_1$和$w_2$的分布相关性向量:

$$sim(t_1, t_2) \geqslant \sigma \quad (1)$$

当$\delta$大于一定的阈值时,我们则认为$w_1$和$w_2$两个单词的分布是非常相似的,在这种情况下,我们将这两个单词聚合为一个模式。计算单词相似度常采用距离函数,常用的距离函数是欧几里德距离、马氏距离、余弦函数和相关系数。在本文中我们全部采用余弦函数作为相似度计算函数,即:

$$sim(x, y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \quad (2)$$

当$\delta$=0.99通过对所有的18000个特征词经过上面公式并应用聚合规则后,最终得到了7036个不同的分布模式。其中最大的模式包含157个原有特征词(很多专有名词均只在一个类中出现),而"达芬奇"和"徐悲鸿"这类词应用聚合规则后,合并为一个模式。同时,模式和类别间无对应关系,如包含特征词"留学生"的模式同时对应国际新闻和教育两个类。对于聚合的新特征,我们采用原有单词的平均$X^2$作为特征选择的依据,其他方法与传统方法一致。

每篇文章都根据所有不同的分布模式构造成一个$m$维的向量。仍用词频公式作为计算特征词的特征值。并对其归一化得到的特征向量。算法的其他部分与传统方法相同。

## 3 基于最大熵的文本分类算法

最大熵原理是在1957年由E.T. Jaynes提出的,其主要思想是,在只掌握关于未知分布的部分知识时,应该选取符合这些知识但熵值最大的概率分布。因为在这种情况下,符合已知知识的概率分布可能不止一个。最大熵的原则:将已知事实作为制约条件,求得可使熵最大化的概率分布作为正确的概率分布。

自然语言处理中很多问题都可以归结为统计分类问题,很多机器学习方法在这里都能找到应用,在自然语言处理中,统计分类表现在要估计类$a$和某上下文$b$共现的概率$P(a, b)$,不同的问题,类$a$和上下文$b$的内容和含义也不相同。在词性标注中是类的含义是词性标注集中的词类标记,而上下文指的是当前被处理的词前面一个词及词类,后面一个词及词类或前后若干个词和词类。通常上下文有时是词,有时是词类标记,有时是历史决策等等[2]。

最大熵模型的优点是:在建模时,试验者只需要集中精力选择特征,而不需要花费精力考虑如何使用这些特征。而且可以很灵活地选择特征,使用各种不同类型的特征。利用最大熵模型,一般也不需要做其它方法建模中常常使用的独立性假设,参数平滑可以通过特征选择的方式加以考虑,无需专门使用平滑算法单独考虑。每个特征对概率分布的贡献则由参数$\alpha_j$决定,该参数通过GIS算法迭代训练得到[3]。

将事件集A当作类别集,将上下文环境集B当作文档集,那么我们就可以使用最大熵模型求任意一篇文档$b_i \in B$属于任意类别$a_i \in A$的概率。对于不存在兼类的分类问题,只要选择最大概率的类别就是文档的所属类别;对于存在兼类的分类问题,定义一个阈值$\varepsilon$,文档$b_i$属于所有$P(a_i | b_j) > \varepsilon$的类别。

## 4 实验及结果分析

目前为止,在中文文本分类中还没有一个标准的、普遍可接受的训练集和测试集。因此,我们采用自己制作的一套训练集和测试集文本对最大熵方法和其他分类方法进行结果对比。我们从网络新闻搜索系统来获得约7000篇文章。覆盖了新浪、搜狐和网易等400余个中文网站。去除重复新闻后,将网上新闻手工分成8个类别。训练集包含约5000篇文本;测试集包含约2000篇文本。我们采用宏平均准确率来进行对比实验[5]。

我们分别采用KNN、Bayes和SVM分类器进行性能比较。使用的$X^2$方法进行特征选择。在训练最大熵模型的参数时,我们使用了GIS算法,迭代100次。Bayes分类算发中我们采用拉普拉斯方法进行平滑[6]。KNN算法K=25,SVM选择多项式核函数,多项式核的阶数为3[4]。$\delta$取1.0、0.99和0.95。

从图2中可以得到如下结论:基于最大熵模型的文本分类方法优于Bayes方法。这与文献[3]的实验结果基本吻合。当$\delta$=0.99时,基于最大熵模型的文本分类方法正确率超过KNN与SVM方法不相上下,都达到93.75%的正确率。使用特征聚合后,所有分类算法的正确率都有明显提高,这是因为去掉了一部分冗余的特征单词,引入了一部分新单词。

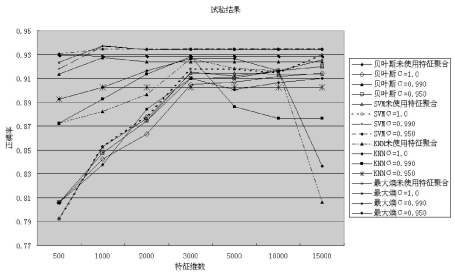


图2 不同分类方法下宏平均准确率比较图

## 5 结 论

本文中,我们使用最大熵模型进行了文本分类的研究。并且对特征选择进行改进和分析。实验结果显示,基于最大熵模型的分类器的准确率超过SVM、KNN和贝叶斯分类算法,它是一种很好的分类器。但是最大熵算法与特征聚合算法结合后,使用不同的训练文档时测试结果相差较大,尤其是在训练集不是很平衡的情况下。特征词聚合之后新$X^2$值生成上也有一定的问题。这些问题还有待于我们进一步研究。

## 参 考 文 献

[1] Yang Y, Liu X. A re-examination of text categorization methods. In 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99). Berkeley: ACM Press, 1999: 42-49.

[2] Adwait R. Maximum entropy models for natural language ambiguity resolution [PhD dissertation]. University of Pennsylvania, 1998.

[3] Kamal N, John L, Andrew M. Using maximum entropy for text classification. In: Proceedings of the IJCAI-99 Workshop on Information Filtering. Stockholm, Sweden, 1999.

(下转第277页)

随机选取初始聚类中心对聚类结果的影响，然后提出了一种从数据对象分布出发寻找初始聚类中心的思想以及基于这种思想的算法过程，并通过实验分析得出改进后的算法能够得到较高且稳定的准确率，更适用于对实际数据的聚类。

## 参 考 文 献


[1] 朱明. 数据挖掘[M]. 合肥：中国科学技术大学出版社，2002

[2] Kumiawan A, Benech N, Tao Yufei. Towards High-dimensional Clustering[J]. COMP, November 1999: 1-2

[3] MacQueen J. Some Methods for Classification and Analysis of Multivariate Observations[J]. In: Proceedings of 5th Berkeley Symp. Math. Statist. Prob., 1967, 1: 281-297

[4] Jolla L. Alternatives to the k-means algorithm that find better clustering[J]. In: Proceeding of ACM SIGMOD, 1992: 192-195

[5] Schrimpf. Migration of Processes, files and virtual devices in the MDX operating system[J]. ACM SIGOPS, 1995: 70-81


（上接第 220页）

引入速度分量，增加观测距阵的秩将提高系统的跟踪能力。

车辆的观测方程为：

$$Y_k = H_k X_k + V_k \tag{16}$$

其中，$V_k$为均值为零，方差为 $R_k$的高斯白噪声；观测距阵为：

$$H_k = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

观测矢量为 $Y_k = [x_b\ v_{gpsx}\ y_b\ v_{gpsy}]^T$。

通过地图辅助的位置和速度匹配估计，我们对观测矢量进行修正，用地图位置匹配点 $P_t(x_m, y_m)$代替 GPS测量位置转换到大地平面的坐标点 $P_t(x_b, y_b)$；用地图速度匹配估计（$v_{mx}$, $v_{my}$）代替 GPS测量速度（$v_{gpsx}$, $v_{gpsy}$）。

### 2.3 自适应递推算法—扩展的卡尔曼滤波

状态一步预测方程：$\hat{X}_{k/k} = \hat{X}_{k/k-1} + K_k[Y_k - H_k \hat{X}_{k/k-1}]$　（17）

状态估计方程：$\hat{X}_{k/k-1} = \Phi_{k/k-1} \hat{X}_{k-1/k-1} + U_k \bar{a}_k$　（18）

滤波增益方程：$K_k = P_{k/k-1} H_k^T [H_k P_{k/k-1} H_k^T + R_k]^{-1}$　（19）

验前协方差阵：$P_{k/k-1} = \Phi_{k/k-1} P_{k-1/k-1} \Phi_{k/k-1}^T + Q_{k-1}$　（20）

验后协方差阵：$P_{k/k} = [I - K_k H_k] P_{k/k-1}$　（21）

如果把 $a_k$的一步预测 $a_{k/k-1}$看作在 k瞬时的“当前”加速度，即随机机动加速度的均值，这样就可得到加速度的均值自适应算法，设：

$$\begin{cases} \bar{a}_{kx} = a_{k/k-1x} \\ \bar{a}_{ky} = a_{k/k-1y} \end{cases} \tag{22}$$

$$\begin{bmatrix} x_{k/k-1} \\ v_{k/k-1} \\ a_{k/k-1} \end{bmatrix} = \begin{bmatrix} 1 & T & (-1+\alpha T+e^{-\alpha T})/\alpha^2 \\ 0 & 1 & (1-e^{-\alpha T})/\alpha \\ 0 & 0 & e^{-\alpha T} \end{bmatrix} \begin{bmatrix} x_{k-1/k-1} \\ v_{k-1/k-1} \\ a_{k-1/k-1} \end{bmatrix} + \begin{bmatrix} (-T+\alpha T^2+(1-e^{-\alpha T})/\alpha)/\alpha \\ T-(1-e^{-\alpha T})/\alpha \\ 1-e^{-\alpha T} \end{bmatrix} a_{k/k-1}$$

上式矩阵形式为：

$$\hat{X}_{k+1/k} = \Phi_1(T) \hat{X}_{k/k} \tag{23}$$

其中，$\Phi_1(T) = \begin{bmatrix} 1 & T & T^2/2 \\ 0 & 1 & T \\ 0 & 0 & 1 \end{bmatrix}$，这样式（18）可改写为 $\hat{X}_{k/k-1} = \Phi_{1\,k/k-1} \hat{X}_{k-1/k-1}$，其它方程不变，由以上的递推方程即可解出车辆的真实状态。

## 3 仿真试验

为了考察地图匹配估计的效果，我们在太原市火车站附近进行了跑车实验，把跑车采集的 GPS定位数据作为参考数据。在获得原始数据的基础上，叠加不同种类的噪声源，再根据本文介绍的算法，进行地图匹配估计。结果表明，借助地图匹配位置择近技术对 GPS定位数据作进一步的校正，可使 GPS定位精度在一定程度上有所提高。图 2表示汽车经过地图位置择近匹配前后在电子地图中的行驶轨迹。其中，实线代表 GPS定位的轨迹，虚线代表 GPS定位数据经过地图匹配修正后的轨迹。

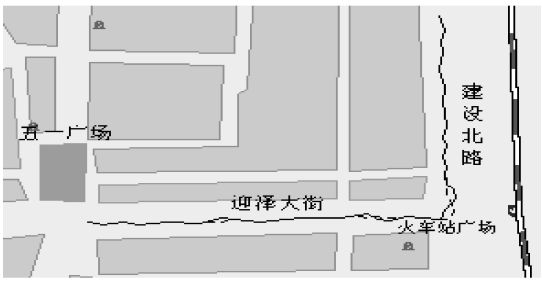


图 2 汽车位置滤波前后的行驶轨迹

## 4 小 结

本文提出采用“当前”的统计模型作为车辆的运动模型，并采用地图辅助位置择近和速度择角的算法来修正 Kalman滤波。试验结果表明，采用“当前”统计模型的 Kalman滤波算法，使滤波器滤波性能有明显的改善；特别是利用地图辅助位置择近和速度择角估计对 Kalman滤波进行修正后，使车辆在道路法线方向的误差减小，从而能够满足 ITS先进的车辆定位导航系统的需要。

## 参 考 文 献


[1] 高晖. GIS与 GPS在 TIS中的应用研究. 北京航空航天大学博士学位论文，1998.

[2] 关桂霞. 车载 GPS/DR组合导航系统研究. 华北工学院硕士学位论文，2001.


（上接第 264页）


[4] Hsu C W, Lin C J. A comparison of methods for multi-class Support Vector Machines. IEEE Transactions on Neural Networks, 2002, 13(2): 415-425.

[5] Yang Y. An evaluation of statistical approaches to text categorization. Information Retrieval, 1999, 1(1): 76-88.

[6] 黄萱菁，吴立德，郭以昆，刘秉伟. 现代汉语熵的计算及语言模型中稀疏事件的概率估计. 电子学报，2000, 28(8): 110-112