基于遗传算法改进诗词风格判别的研究*)

易 勇 何中市 李良炎 周剑勇 瞿义玻
(重庆大学计算机学院 重庆400044)

摘 要 本文对诗词采用向量空间模型来表示,基于机器学习中的朴素贝叶斯等方法,首次提出了古典诗词的豪放和婉约风格判别计算模型,并用遗传算法对模型进行改进,取得较好的诗词风格判别结果。该模型已经在精典诗词语料的机器学习基础上得以实现,并且获得较好的诗词风格判别效果。
关键词 机器学习,文本分类,遗传算法,文学风格判别

A Traditional Chinese Poetry Style Identification Calculation Improvement Model

YI Yong HE Zhong-Shi LI Liang-Yan ZHOU Jian-Yong QU Yi-Bo
(College of Computer, Chongqing Univrsity, Chongqing 400044)

Abstract Based on Machine Learning methods—Naïve Bayes and Genetic Algorithm, this paper proposes a Traditional Chinese Poetry Style Identification Calculation Improvement Model to identify Bold-and-Unrestrained or Graceful-and-Restrained styles, that derive from Machine Learning Chinese Classical Ci in Song Dynasty. Feature subset selection is performed based on Genetic Algorithm and has achieved satisfactory identification results in application.
Keywords Machine learning, Text categorization, GA, Literature style identification

1 引言

目前,自然语言处理技术取得了不少进展,文学性语言的计算机处理作为自然语言处理的分支正摆在了学术界的面前,笔者在国家自然科学基金的支持下正在尝试做该领域的初步探索,本文的讨论重点是如何用机器实现诗词的风格判别。作为中华传统文化精华之一的古典诗词,含义隽永,言简情深,其风格令人陶醉,但长期以来诗词作品的风格判别多是读者凭体验、靠感觉意会对其进行风格的判别和认知,传统上并没有明确的数量化规则,更没有形式化的规则。而对机器而言只能依靠诗词的文字内容来进行风格的确定,从形式上诗词也是一段文字,其风格的判别也可视为文本风格的分类,于是此问题实质上就转化为机器学习中的文本分类问题,即在给定的风格分类体系下,根据文字的内容机器自动学习到规律,并根据规律建立分类判别器,由分类判别器自动确定文字的风格,据此本文提出了一个基于机器学习的中国古典诗词风格判别计算模型框架。本研究对于拓展机器学习的应用领域,对于传统诗词的信息化整理和发掘,对于诗词的语言理解和各种文学作品的计算机辅助研究具有重要意义。

本文第2节总结了诗词文本的表示模型和贝叶斯判别模型等相关工作。第3节介绍诗词风格判别的实验,并给出了实验结果,最后是本文的总结。

2 相关工作

2.1 确定诗词风格的标志

中国古代诗词文化发达,产生的流派众多,诗学理论的风格分类较为复杂,为了机器学习研究的方便,简化研究对象模型,本文主要按照人们最常用的风格体验,即将中国古典诗词文本的风格分为豪放风格和婉约风格两类。诗词文本的风格判别实质上转化为机器学习中的文本风格分类的问题。而对于一般的文本分类的机器学习任务,一般经典上多采用向量空间表示模型和基于统计的特征提取技术。我们的实验研究中也采用类似的技术路线。

2.2 诗词文本数字化知识表示

从已有的有关文本机器分析的文献分析可以发现:文本的机器表示方法中应用最广泛的是向量空间模型。在向量空间模型中,必须选定文本的特征表示集,这个特征为字、词、短语、概念和N-gram等,通常最常用的是字和词。文本和类别标志被表示成高维空间向量,其中每一维为一个特征。一个文本或类别向量的第1个元素表示文本或类别第1个特征的重要性,即权重值。但是中文文本不像西文文本有空格作为词语的界限,中文文本的无空格作为词的分界标志的问题造成分词的歧义和困难,而且中国古典诗词由于采用的是古代的书面语言,无法用现代语言的机器分词系统进行高效有意义的分词,即使可以分词中文词语的词语数量巨大,将造成机器学习的系统设计管理和维护的困难,同时海量的诗词经典文本由于实用和时间的原因又无法用人工进行分词,但是中国古典诗词特别讲究语言的凝练,一般通过"字"就表达了很复杂的事物和情感,根据上述原因,我们很自然地想到如果能绕过中文文本中词的正确切分的难题,将节省大量的编程,机器运行和构造、维护词表的时间,把研究重点放到如何根据"字"来对原文内容自动分类上。

就汉字的个数而言,国标GB 2312中一二级字库中有汉字6763个,而常用的汉字在几千个左右。也就是说,尽管词的个数难以控制,但汉字的个数是相对恒定的,用字出现向量表示文本能够保证汉字对诗词文本的覆盖率。如果能从汉字中筛选出具有风格判别分类意义的汉字定义向量空间的维数,这将大大降低分类系统运行的时间和空间复杂度。因此,我们选定汉字的出现与否作为诗词文本表示的基本单元。诗词文本数字化就是将诗词文本表示为一个稀疏的向量,向量的分量是字的出现的与否的标志,用离散的数字1和0表示,如果该字在诗词中出现,则向量的该字的分量的值为1,理论上每首

*)本文受国家自然科学基金项目(基金号60173060)支持。易 勇 博士研究生,主要研究方向:自然语言处理,机器学习,数据挖掘。何中市教授,主要研究方向:自然语言处理,机器学习,概率论,容错计算。李良炎 博士研究生,主要研究方向:自然语言处理,机器学习,心理学。

诗的机器内部表示就是一个诗词语料中所出现的字的出现与否所组成的向量，在本文的实验中我们为了减少向量的维度，减少稀疏度和特征数，还采用了基于统计的特征提取技术和遗传算法，将宋词语料库中的豪放风格和婉约风格的两类诗词语料中所用的对风格分类有显著意义的汉字提取出来，最终采用了55个汉字，即作为向量的分量的维度数为55。这样每一首宋词就表示为由“人”、“花”、“春”、“天”、“云”等汉字的出现标志组成的向量模型。

2.3 风格判别学习算法

建立了诗词风格判别的文本表示的向量模型后，就可以采用自动文本分类方法，本文主要采用朴素贝叶斯(Naïve Bayes)分类方法。

贝叶斯分类利用贝叶斯决策规则进行分类，如果要决定文本 $\vec{d}$ 属于类别 $C1$还是 $C2$，首先计算出概率 $P(C1|\vec{d})$、$P(C2|\vec{d})$，即分别计算出 d 属于不同类别的概率，如果 $P(C1|\vec{d})>P(C2|\vec{d})$，则文本属于类别 $C1$，反之则属于类别 $C2$。

在实际中并不直接计算 $P(C1|\vec{d})$、$P(C2|\vec{d})$的值，而是利用贝叶斯公式：

$$P(C|\vec{d})=\frac{P(\vec{d}|C)}{P(\vec{d})}P(C)$$

根据贝叶斯决策规则，可通过下面的公式决定文本 d 的类别 $\hat{C}$：

$$\hat{C}=\arg\max P(C|\vec{d})=\arg\max P(\vec{d}|C)P(C)=\arg\max[\log P(\vec{d}|C)+\log P(C)]$$

为了计算方便，假定各特征值对给定类的影响独立于其他特征，即文本向量中的特征相互独立，即朴素贝叶斯假设：

$$P(\vec{d}|C)=P(\{d_j|d_{j\in\vec{d}}\}|C)=\prod_{d_j\in\vec{d}}P(d_j|C)$$

如果只计算单个字的分布，大大地减少了计算量：

$$P(C_j)=\frac{C_j\text{的文档个数}}{\text{总文档个数}}=\frac{N(C_j)}{\sum_k N(C_k)}\approx\frac{1+N(C_j)}{|C|+\sum_k N(C_k)}$$

$$P(w_i|C_j)=\frac{w_i\text{在}C_j\text{类别文档中出现的次数}}{C_j\text{类所有文档中出现的字的次数}}\approx\frac{1+N_{ij}}{\text{不同字个数}+\sum_k N_{kj}}$$

其中，w_i 为一个特征，在本文中为一个汉字。

但是朴素的贝叶斯假定在一个位置上出现的字的概率独立于另外一个位置的字，文本向量中所有的字是无序的、没有结构的，字之间的没有相互依赖关系。这个假定有时并没反映文本真实情况，虽然该独立性假设很不精确，在很多情况下不成立，但这种处理带来了计算上的实用和有效，否则计算的概率项将极为庞大，而且幸运的是，在实践中朴素贝叶斯学习器在许多文本分类中性能非常好，即使独立性假设不成立，在本文的诗词风格判别实验中，我们采用朴素贝叶斯分类模型构建了诗词风格学习器和判别器，其结果也毫不比实验中采用的其他方法(决策树、KNN)逊色。

2.4 基于遗传算法的特征子集选择

在诗词风格判别的实验系统中，1327个不同的汉字出现在训练语料中，因而，最初的朴素贝叶斯分类模型的特征的数目高达上千个。然而许多汉字对诗词的风格判别无意义或用处不大。其中的一些汉字甚至可能导致风格判别模型的准确度的降低。本质上，我们面临一个去粗取精，从大特征集中进行特征选择的问题。如何选择一个最优的特征子集，这是一个NP难的问题，除了采用穷举法搜索外，不能保证得到最优解，但是一些现代启发式算法可用于该类问题而得到较优解，我们最终采用了遗传算法来进行特征子集的选择，其算法的基本过程如下：

步骤0：先将原始的特征集进行二进制编码，用基因链表示，用0或1构成的字符串表示一种特征组合，其中的数字1所对应的特征被选中，而数字0所对应的特征未被选中。

步骤1：令进化代数 $t=0$。

步骤2：初始化种群。

步骤3：对种群中的每一个个体计算适应度(其中适应度用风格判别的准确度表示)。

步骤4：形成新的种群。

步骤5：执行进化操作，在种群中进行交叉和变异。

步骤6：转到步骤3，直到进化代数或适应度达到预设值。

3 诗词风格判别的实验

3.1 基于字出现的诗词文本风格判别实验系统

我们从诗词精典共计159万字、1万9千余首宋词的《全宋词》中经专家人工选出豪放风格188首，婉约风格210首作为实验的诗词学习和训练语料，实验系统采用高级语言编程实现，系统开发和设计按研究和处理顺序分为全宋词使用汉字抽取和字频统计模块，字出现分析和字同现向量生成模块，诗词风格训练学习模块，诗词风格判别分类模块和风格判别结果评价模块。将诗词文本采用字同现向量模型进行表示，将豪放风格188首宋词标记为“豪放”，将婉约风格210首宋词标记为“婉约”，采用朴素贝叶斯分类模型进行学习，本文所用的方法从已标记风格分类的样本数据开始，尝试用模型对给定的特征值(即汉字)，判别出每个诗词样本的风格。

在典型的文本分类中涉及两个对象：学习器和分类器。在本研究中使用的学习器是朴素的贝叶斯学习器。当训练数据传给学习器学习后，产生了一个分类判别器。当新的数据传给风格分类判别器后，风格分类判别器返回判别出的诗词风格和风格倾向的概率值。

在建立了诗词风格的朴素贝叶斯学习模型和判别模型(参看表1中的方法1)后，从比较研究出发，我们用基于启发式搜索算法的爬山策略构造了一个简单的特征子集算法(参看表1中的方法2)，将汉字作为机器学习中的特征，首先从空的特征集开始，逐个测试特征，如果分类的准确度得到提高，则从最初的特征集引入该特征，直至加入的任何单个特征不能导致分类准确度的提高，则停止特征子集选择算法。此外，我们采用信息增益(information gain)评估每个特征(汉字)与风格判别的相关性，并选出了最相关的100个汉字作为特征，并用遗传算法进行特征子集选择。

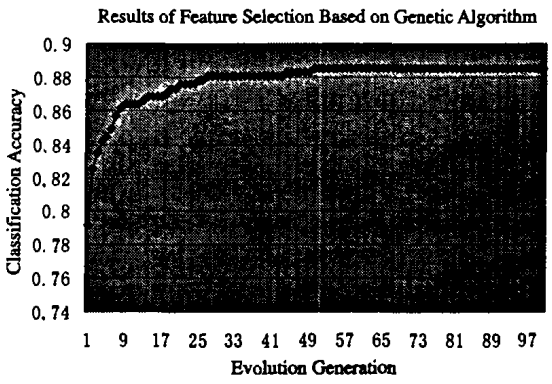

图1 基于遗传算法的特征选择结果

在最初的特征子集选择后，在遗传算法中的将其适应度定义为风格判别的准确度，并将其与朴素贝叶斯分类算法相结合(参看表1中的方法3；图1的遗传算法进化过程)。考虑到

性能评估的因素，我们使用了10交叉验证法。下文中的表格列出了对同样的188首豪放风格的宋词和210首婉约风格的宋词，使用不同方法得出的诗词风格判别的准确度以及作为特征的汉字数。

我们上述研究中采用的分类性能评估的方法是机器学习领域里比较常用的10交叉验证法，即将训练集随机划分为10个不相交的相等规模的组，每组有 $n/10$ 个宋词样本，n 为样本总数（n 在实验中为398），分类器要训练10次，每次留出1组作为验证集。

表1 10交叉验证结果

方法序号	判别模型	模型准确度	特征数(汉字数)
1	Naïve Bayes 模型	0.502	1327
2	用爬山法改进的 Bayes 模型	0.744	13 to 33
3	用遗传算法改进的 Naïve Bayes 模型	0.885	55

一种重要的分类问题的统计分析工具是接受特性(Receiver Operating Characteristic)曲线，通常简称ROC曲线，可用于比较各个分类方法的优劣，ROC曲线下包围的面积，代表了对于"二中选一"的问题正确回答的概率，ROC曲线下包围的面积越大说明对应的分类方法越好，综上所述，根据ROC曲线下包围的面积判断，基于遗传算法改进的朴素贝叶斯方法是较好的诗词风格判别方法(见表2)。

结果及分析 文学性语言的分析在自然语言处理中一直是一个难题，对诗词的计算机分析更难，本文在此作了初步的尝试。通过我们的研究表明，对古典诗词的采用以字为单位的向量空间表示模型能较好地表现诗词的内容和感情，较好地避免了诗词语句的分词难题，同时汉字字符的总数控制了问题的规模，不会使处理的无限扩大。用基于字的向量空间模型表示诗词文本后，我们将古典诗词作者判别问题转化为了文本分类的问题，利用已有的成熟理论和技术，我们建立了多种机器学习和分类器，并且成功地用朴素贝叶斯方法结合遗传算法达到了较好的风格判别结果，准确率近似达到88.5%。我们的试验系统在输入诗词文本后，其中的诗词文本风格判别系统，能给出所输入诗词的风格是豪放或婉约的评价及其风格倾向的概率，我们下一步的研究将关注建立诗词风格量化的评价准则和方法。

表2 ROC 曲线下面积的比较

判别模型	ROC 曲线下包围的面积	特征数(汉字数)
Naïve Bayes 模型	0.394	1327
爬山法改进的 Bayes 模型	0.743	13 to 33
用遗传算法改进的 Naïve Bayes 模型	0.8968	55

参考文献

1 Aas K, Eikvil L. Text Categorisation: A Survey. June 1999. 3
2 Yao Tiansun. Natural Language Understanding. Tsinghua University Press, 2002. 373
3 Han J, Kamber M. Data Mining: concepts and techniques. 2001
4 Mitchell T M. Machine Learning. McGraw-Hill Companies
5 Wen Cai. Extension Engineering Method. Science Press, 2000. 99
6 Marques de sa J P. Pattern Recognition: Concepts Methods and Applications. Springer-Verlag Berlin Heideberg New York, 2002
7 Duba R O, Hart P E, Stork D G. Pattern Classification (Second Edition). John Wiley &Sons, Inc., 2001
8 Yi Yong, He Zhongshi, Li Liangyan. Studies on Traditional Chinese Poetry Style Identification. In: the Proc. of ICMLC04. Shanghai, Aug. 2004. 2936～2939
9 唐圭璋. 全宋词. 中华书局, 1965

(上接第134页)

它仍然在一定程度上提供了辅助评价的支持。

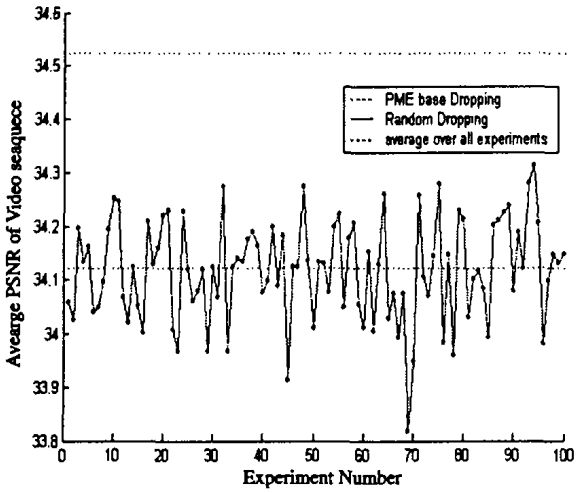

图2 视频序列平均 PSNR

总结和展望 本文中感知运动能量模型的应用使我们能够根据视频序列中运动活动的水平及视觉感知质量来对视频帧进行标记。本文将此模型与新制定的 MPEG-21数字项适配框架相结合，提出了基于 PME 的 H.264/AVC 视频适配模型及系统，实现了独立于媒体编码格式的视频适配机制。本文工作提供了一种标准化和灵活的方法来使 H.264/AVC 视频适应实际的应用环境，同时视频的主观感知质量尽可能地得到了保证。

今后的研究工作应是进一步改进 H.264/AVC 视频感知能量模型来更加准确地描述视频中的运动特性，从而更好地保证视频主观感知质量。此外，分布式视频适配，如在分布式的数个网络节点上协同地执行连续的适配操作，也将是我们的研究重点之一。

参考文献

1 Ma Y, Zhang H J. A New Perceived Motion based Shot Content Representation. 2001 IEEE Intl. Conf. on Image Processing (ICIP 2001), Greece, Oct. 2001
2 Liu T, Zhang H J, Qi F. Perceptual Frame Dropping in Adaptive Video Streaming. 2002 IEEE Intl. Symposium on Circuits and Systems (ISCAS2002), Arizona, USA, May2002
3 Ardizzone E, et al. Video Indexing Using MPEG Motion Compensation Vectors. Multimedia Computing and Systems, 1999. 725～729
4 Ardizzone E, La Casica M, Molinelli D. Motion and Color-Based Video Indexing and Retrieval. 13th International Conference on Pattern Recognition (ICPR96), Vienna, Austria, Aug, 1996
5 Wang H, Divakaran A, Vetro A, Chang S-F, Sun H. Survey of Compressed-Domain Features Used in Audio-Visual Indexing and Analysis. Journal of Visual Communication and Image Representation, June 2003, 14:150～183
6 Bormans J, Gelissen J, Perkis A. MPEG-21: The 21st century multimedia framework. IEEE Signal Processing Magazine, March 2003, 20:53～62
7 Panisa G, Huttera A, Heuera J, et al. Bitstream syntax description: a tool for multimedia resource adaptation within MPEG-21. Signal Processing: Image Communication, EURASIP, Sept. 2003, 18:721～747
8 Mukerjee D, Kuo G, Liu S, Beretta G. Motivation and Use cases for Decision-wise BSDLink, and a proposal for Usage Environment Descriptor-AdaptationQoSLinking. ISO/IEC JTC 1/SC 29/WG 11, Hewlett Packard Laboratories, April 2003
9 ITU-T Recommendation P.910. Subjective video quality assessment methods for multimedia applications. Geneva, 1996