参数嵌入算法在文本分类可视化中的应用

张 莹,王耀南,万 琴
ZHANG Ying,WANG Yao-nan,WAN Qin

湖南大学 电气与信息工程学院,长沙 410082
College of Electrical and Information Engineering,Hunan University,Changsha 410082,China
E-mail:gdutzy@hotmail.com

ZHANG Ying,WANG Yao-nan,WAN Qin.Application of parametric embedding algorithm to text classifier visualization. Computer Engineering and Applications,2009,45(16):31-35.

Abstract: How to visualize the text classifier result is one of the focus field in pattern recognition.On the assumption that each class can be represented by a Gaussian distribution in the embedding space,through Naive Bayes classification algorithms posterior probability for data over classes was got,Parametric Embedding(PE) algorithm was applied into the visualization of classification result in low-dimensional.PE algorithm tries to preserve the structure in an embedding space by minimizing a sum of Kullback-Leibler divergences in high-dimensional space.Data that are located at the center of cluster are typical data for the class, and data that are located between clusters have multiple topics,different data are located in the cluster of different classes.The outstanding advantage is that computing complexity is just the type of data and the corresponding number of the product,is well suited to large volume of data,fewer types of classified data visualization.Experimental result on 20 Newsgroups data sets and MiniNewsgroups data sets proves the effectiveness of the method.

Key words: Naive Bayes classifier;parametric embedding;text classification;posterior probability;classification visualization

摘 要:如何对文本分类的结果进行可视化研究一直是模式识别中研究的重点。在假设文本类别在低维嵌入空间服从高斯分布的前提下,通过朴素贝叶斯分类算法得到数据类别属性的后验概率矩阵,然后运用参数嵌入算法在低维空间可视化文本分类结果。参数嵌入算法是使嵌入空间数据的类后验概率与高维空间的条件概率 Kullback Leibler 散度和最小化的算法,属于同一类的数据在低维空间中分布较为集中,性质相似的数据之间的距离较近,而不同性质的数据之间距离则较大。其优点在于计算复杂度是数据的类别和相应个数的乘积,非常适合于数据量大,类别数较少的数据分类可视化。20 新闻组数据集和微型新闻组数据集的实验结果证明了该算法的有效性。

关键词:朴素贝叶斯分类;参数嵌入;文本分类;后验概率;分类可视化
DOI:10.3778/j.issn.1002-8331.2009.16.008 **文章编号:**1002-8331(2009)16-0031-05 **文献标识码:**A **中图分类号:**TP391.1

1 引言

文本分类指根据一些预先定义好类标签的训练文档集合来对新文档分类，其目的就是对文本集进行合理处理和组织，使得这些文本能够按照类别区分开来。从数学的角度来看是一个映射过程,它将未知类别的文本映射到已有的类别中。文本分类技术已经应用于信息检索、信息抽取、信息过滤、数据组织、网上信息快速定位等多个领域。

常用的文本分类算法[1]主要包括三大类:第一类基于特征提取的分类算法，其基本思想是利用 TFIDF 权值公式计算单词在文档中的重要性，然后用 Cosine 距离公式计算两个文档的相似度，这类算法包括 Rocchio 算法、TFIDF 算法、K 近邻算法等;另一类是基于概率和信息理论的分类算法,如朴素贝叶斯算法(Naive Bayes)、最大熵算法(Maximum Entropy)等;第三类是基于学习的分类算法,如决策树(Decision Tree)、人工神经网络(Artificial Neural Networks)、支持向量机(Support Vector Machine)等。其中 K 近邻、朴素贝叶斯和支持向量机由于分类效果比较好成为近几年人们研究的热点。

目前在文本分类领域通常使用查全率(Recall)和查准率(Precision)来衡量分类器的性能,无法得到分类后文本在低维空间的分布情况,不利于分类结果的直观显示。故本文采用参数嵌入算法(paremetric embedding)将文本分类结果投影到低维空间进行显示，不但可以增加对分类结果直观性的认识,还可以根据低维空间分布对分类算法进行改进,有助于分类准确率的改善和特征词的选择。

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60775047);国家高技术研究发展计划(863)(the National High-Tech Research and Development Plan of China under Grant No.2007AA04Z244,No.2008AA04Z214)。

作者简介:张莹(1972-),男,博士研究生,主要研究方向为信息融合,图像处理等;王耀南(1956-),男,教授,博导,主要研究领域为智能控制,机器视觉,信息处理等;万琴(1981-),女,博士研究生,主要研究方向为信息融合,视频跟踪等。

收稿日期:2009-02-10 **修回日期:**2009-03-20

2　文本分类算法简介

因为原始的文本不能够被分类器直接处理,所以首先需要对文本表达成分类器能够处理的形式。目前大多数采用经典的“词袋”(bag-of-words)模型,文档被看成是由相互无关的单词构成的集合,不考虑单词之间的上下文关系,单词出现的顺序、位置以及文章的长度等,仅记录每个单词在文档中出现的次数或记录单词是否出现。通过对训练文本集进行文法分析统计出所有单词在训练文档中出现的频率,得到词频矩阵。矩阵中的某一元素就是某个单词在某篇训练文档中出现的频率,词频矩阵是文本分类算法建立分类器模型的数据基础。一般文本分类系统的体系结构如图 1 所示。

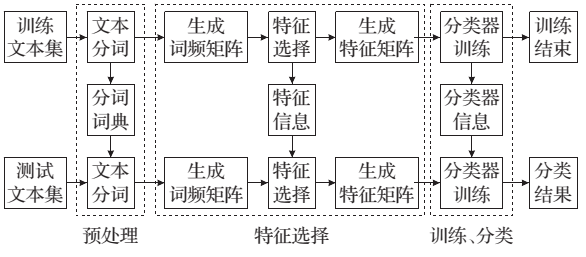

图 1　文本分类系统的体系结构

2.1　预处理

预处理是对训练文本集及待分类文本集通过分词处理分析出文档中的单词。对英文文本分析的步骤为:按空格切分出各个单词,过滤掉其中的停用词(stop words),如 a、the、that 等。然后进行词干提取(stemming),如将 played、playing 变为 play,将文档对应的词频向量中该单词的频率加 1,如果这是训练语料集中第一次遇到的新词,就存入词库,该模块中还包括保存文档的文件名、类别等工作。如果对中文文本进行预处理,需要增加中文单词切词。

2.2　特征选择

经过预处理后,文本集被表示为词频向量,同时训练文本集的词频向量构成了词频矩阵,由于词库中存在大量的词汇,把所有单词都作为特征将带来一系列问题。首先是向量的维数太大,给计算带来了非常大的压力,且存储空间大,处理速度慢。其次是这些词中实际上有很大一部分是与类别无关的,对分类作用不大。因此采用特征选择方法(互信息,信息增益,文档频率,x^2 统计量等)选择词库中那些有代表意义的单词作为特征,从而降低向量的维数,得到代表文本特征的词频矩阵和待分类文本词频向量。

2.3　训练、分类

在词频矩阵的基础上根据特定的分类算法构造分类器,对待分类文本进行分类。如 K 近邻算法对于一篇待分类的文档,系统在训练集中统计 K 个近邻中多数属于哪一类,就把待分类的文档归为那一类,然后在已分类文档集中检索与待分类文档最相似的文档子集,从而获得被测文档的类别。

3　参数嵌入算法原理

参数嵌入算法[2]是假设类别在低维嵌入空间服从高斯分布的前提下,使实验数据在嵌入空间的后验概率与训练数据的条件概率误差最小化的算法。它能反映数据和类别的整体特征,属于同一类的数据在低维空间中嵌入较为集中,性质相近类的数据之间的距离较小,而不同性质类之间数据距离较大。该算法不仅能在低维空间表示高维数据与所属类别的联系,而且能正确地揭示出数据集内部和类别集内部的关系,而且计算复杂度仅是数据集的分类数和样本个数的乘积,对数据量大,类别数较少的数据集而言,参数嵌入算法提供了一个其他基于配对距离方法所不能快速处理的、简单的、有效的可视化方法。

假设数据集 $X=\{x_1,\cdots,x_N\}$ 中数据 x_n 在低维嵌入空间属于类别集 $C=\{c_1,\cdots,c_k\}$ 中类别 c_k 的条件概率为 $P(c_k|x_n)$。若是有监督学习,则类别概率 $P(c_k)$ 已知;若为半监督学习或者无监督学习,则假设 $P(c_k)$ 为均匀分布。PE 算法通过寻找高维数据在低维嵌入空间的位置坐标 $R=\{r_n\}_{n=1}^N$ 和类别属性 $\phi=\{\phi_k\}_{k=1}^K$,使数据在低维空间的位置坐标和所属类别两者间后验概率 $P(c_k|r_n)$ 与条件概率 $P(c_k|x_n)$ 误差最小化,即

$$P(c_k|r_n)=\frac{P(c_k)\exp(-\frac{1}{2}\|r_n-\phi_k\|^2)}{\sum_{l=1}^{k}P(c_l)\exp(-\frac{1}{2}\|r_n-\phi_l\|^2)} \tag{1}$$

通常用 K-L 散度和来衡量两者之间的近似程度,即

$$\sum_{n=1}^{N}KL(P(c_k|x_n)\|P(c_k|r_n)) \tag{2}$$

最小化式(2)等价于式(3),因式(3)对参数 r_n 的 Hessian 矩阵为式(4),式(4)右边刚好是一个协方差矩阵,故 Hessian 矩阵是半正定矩阵。

$$E(R,\phi)=-\sum_{n=1}^{N}\sum_{k=1}^{K}P(c_k|x_n)\log P(c_k|r_n)=-\sum_{n=1}^{N}\sum_{k=1}^{K}P(c_k|x_n)\cdot\left(\log P(c_k)-\frac{1}{2}\|r_n-\phi_k\|^2-\log\sum_{l=1}^{K}P(c_l)\exp(-\frac{1}{2}\|r_n-\phi_k\|)\right)=$$
$$-\sum_{n=1}^{N}\left(\sum_{k=1}^{K}P(c_k|x_n)\log P(c_k)-\frac{1}{2}\sum_{k=1}^{K}P(c_k|x_n)\|r_n-\phi_k\|^2-\log\sum_{k=1}^{K}P(c_k)\exp(-\frac{1}{2}\|r_n-\phi_k\|^2)\right) \tag{3}$$

$$\frac{\partial^2 E}{\partial r_n\partial r_n'}=-\sum_{k=1}^{K}\phi_k\cdot\left(\frac{(r_n-\phi_k)'P(c_k)\exp(-\frac{1}{2}\|r_n-\phi_k\|^2)}{\sum_{l=1}^{K}P(c_l)\exp(-\frac{1}{2}\|r_n-\phi_l\|^2)}+\frac{P(c_k)\exp(-\frac{1}{2}\|r_n-\phi_k\|^2)\sum_{l=1}^{K}(r_n-\phi_l)'P(c_l)\exp(-\frac{1}{2}\|r_n-\phi_l\|^2)}{\left(\sum_{l=1}^{K}P(c_l)\exp(-\frac{1}{2}\|r_n-\phi_l\|^2)\right)^2}\right)=$$
$$-\sum_{k=1}^{K}P(c_k|r_n)\phi_k r_n'+\sum_{k=1}^{K}P(c_k|r_n)\phi_k\phi_k'+\sum_{k=1}^{K}P(c_k|r_n)\phi_k r_n'-\left(\sum_{k=1}^{K}P(c_k|r_n)\phi_k\right)\left(\sum_{k=1}^{K}P(c_k|r_n)\phi_k\right)'=$$
$$\sum_{k=1}^{K}P(c_k|r_n)\phi_k\phi_k'-\left(\sum_{k=1}^{K}P(c_k|r_n)\phi_k\right)\left(\sum_{k=1}^{K}P(c_k|r_n)\phi_k\right)' \tag{4}$$

对目标函数式(3)增加正规化条件得

$$J(R,\phi)=E(R,\phi)+\eta_r\sum_{n=1}^{N}\|r_n\|^2+\eta_\phi\sum_{k=1}^{K}\|\phi_k\|^2,\eta_r,\eta_\phi>0 \quad (5)$$

由式(4)Hessian 矩阵半正定性质可得式(5)关于参数 r_n 的 Hessian 矩阵为正定矩阵,故在给定 ϕ 的情况下 R 一定可以找到全局最优解。式(5)可采用梯度下降法求解,即首先随机固定给 ϕ(或 R)一个预设值,然后对 R(或 ϕ)求导使 J 最小,反复循环直到 J 收敛。

式(3)分别对 r_n 和 ϕ_k 求导可得

$$\frac{\partial E}{\partial r_n}=\sum_{k=1}^{K}P(c_k|x_n)(r_n-\phi_k)-\frac{\sum_{k=1}^{K}(r_n-\phi_k)P(c_k)\exp(-\frac{1}{2}\|r_n-\phi_k\|^2)}{\sum_{k=1}^{K}P(c_k)\exp(-\frac{1}{2}\|r_n-\phi_k\|^2)}=\sum_{k=1}^{K}(P(c_k|x_n)-P(c_k|r_n))(r_n-\phi_k)=\sum_{k=1}^{K}a_{n,k}(r_n-\phi_k) \quad (6)$$

$$\frac{\partial E}{\partial \phi_k}=-\sum_{n=1}^{K}P(c_k|x_n)(r_n-\phi_k)+\sum_{n=1}^{N}\frac{(r_n-\phi_k)P(c_k)\exp(-\frac{1}{2}\|r_n-\phi_k\|^2)}{\sum_{k=1}^{K}P(c_k)\exp(-\frac{1}{2}\|r_n-\phi_k\|^2)}=\sum_{n=1}^{N}(P(c_k|x_n)-P(c_k|r_n))(\phi_k-r_n)=\sum_{k=1}^{K}a_{n,k}(\phi_k-r_n) \quad (7)$$

其中 $a_{n,k}=P(c_k|x_n)-P(c_k|r_n)$,式(6)、式(7)可以理解为由 $a_{n,k}$ 符号决定的在低维嵌入空间让本来远离的点尽可能分开同时使本来临近的点尽可能接近的合力总和。即可以表示属于同一类的数据在低维空间中嵌入较为集中,性质相近类的数据在投影后它们之间的距离较小,而不同性质类之间数据距离较远。

参数嵌入算法可以看作是随机邻域嵌入算法[3](SNE)的推广,用数据点 x_n 属于类别 c_k 的类条件概率 $P(c_k|x_n)$ 代替了 SNE 中的点 j 属于点 i 邻域的概率 $P(x_j|x_i)$,当 $P(c_k|x_n)$ 由非监督混合模型求得时,PE 完成非监督维数约简的运算过程与 SNE 相似。PE 单次循环的计算复杂度为 O(NK),仅与数据个数 N 和类别 K 相关,与其他降维算法[4]相比如多维尺度法(MDS)的计算复杂度 O(ND,D 是特征分解后的非零项)、等距映射法的计算复杂度 O(N^3)、SNE 的计算复杂度 O(N^2)降低许多,故 PE 计算速度优于 SNE 及其他基于距离配对的嵌入方法。

参数嵌入算法的关键在于求解高维数据的后验概率矩阵。目前后验概率的确定主要有两类方法:一类是采用贝叶斯框架理论,先求出各类的类条件概率密度,再依据贝叶斯理论求出其后验概率。另一类方法不计算类概率密度,直接拟合后验概率如支持向量机。Vapnik 将后验概率看作余弦函数和的形式,Platt 将后验概率看作 Sigmoid 函数的形式,然后采用最大似然估计准则,求出其函数的参数。在多分类器中,当所有类都是相等损失时,基于最大后验概率选择的分类就是贝叶斯最优决策。

4　朴素贝叶斯文本分类算法

贝叶斯分类分为朴素贝叶斯分类和贝叶斯网络分类,前者假设属性对类的影响相互独立,与其他分类算法相比,当假设成立时朴素贝叶斯分类是最精确的;后者则考虑属性之间的依赖程度,更能反映真实文本的情况,但计算复杂度比朴素贝叶斯高得多,目前还停留在理论的研究阶段。

假设单词在给定类别下的条件概率分布是互相独立的[5],这使得分类器不需要计算单词之间的联合分布概率,可以利用单词和类之间的联合概率来估计给定文档属于类的概率,朴素贝叶斯文本分类是通过一系列关于目标函数的训练样例预测新实例的分类。

朴素贝叶斯文本分类是在已知训练文本的属性值(a_1,a_2,…,a_n)来预测新文本的分类值 C:

$$c_{map}=\arg\max_{c_j\in C}\frac{P(a_1,a_2,\cdots,a_n|c_j)P(c_j)}{P(a_1,a_2,\cdots,a_n)}=\arg\max_{c_j\in C}P(a_1,a_2,\cdots,a_n|c_j)P(c_j) \quad (8)$$

这里 $P(a_1,a_2,\cdots,a_n)$是不依赖于类别 c_j 的常量,常在计算时省略。

在单词独立性假设情况下,$P(a_1,a_2,\cdots,a_n|c_j)$等于每个属性单独概率的乘积,即

$$P(a_1,a_2,\cdots,a_n|c_j)=\prod_i P(a_i|c_j) \quad (9)$$

代入式(8),可得到朴素贝叶斯算法输出的目标值

$$C_{NB}=\arg\max_{c_j\in C}P(c_j)\prod_i P(a_i|c_j) \quad (10)$$

因此要对一个新文档进行分类,需要从训练集中估计出两组概率值:$P(c_j)$和 $P(a_i|c_j)$。

朴素贝叶斯分类器通常使用拉普拉斯公式估计各个属性在类别 c_j 上出现的概率来计算输出值:

$$P(c_j)=\frac{\sum_{i=1}^{|D|}P(c_j|d_i)}{|D|}\quad j=1,2,\cdots,|C| \quad (11)$$

$$P(a_t|c_j)=\frac{1+\sum_{i=1}^{|D|}B_{it}P(c_j|d_i)}{2+\sum_{i=1}^{|D|}P(c_j|d_i)}\quad j=1,2,\cdots,|C|;t=1,2,\cdots,n \quad (12)$$

其中 $|D|$ 是训练文档集总数,$P(c_j|d_i)\in\{0,1\}$表示训练文档 d_i 是否属于 c_j,根据实现细节的不同,$P(c_j|d_i)$一般有两种模型:

(1)多元模型

$$P(d|c_j)=\prod_{i=1}^{n}((B_{xt}P(w_t|c_j))+(1-B_{xt})(1-P(w_t|c_j))) \quad (13)$$

其中 w_t 代表第 t 个特征(即文本向量的第 t 分量),n 代表特征总数,B_{xt} 代表特征 w_t 是否在文本 d 中出现,特征项在文本中出现记为 1,否则记为 0。其中

$$P(w_t|c_j)=\frac{1+c_j\text{中包含特征}\ w_t\text{的文档数}}{2+c_j\text{中所有的文档数}} \quad (14)$$

(2)多项式模型

$$P(d|c_j)=\prod_{i=1}^{n}\frac{P(w_t|c_j)^{N_{it}}}{N_{xt}!} \quad (15)$$

$$P(w_t|c_j)=\frac{1+\sum_{j=1}^{|D|}N_{it}P(c_j|d_i)}{|C|+\sum_{s=1}^{|C|}\sum_{j=1}^{|D|}N_{is}P(c_j|d_i)} \quad (16)$$

其中 $|C|$ 表特征总数,N_{xt} 代表特征 w_t 在文本 d 中出现的次数,

N_{it} 代表特征 w_t 在文本 d_i 中的出现次数，N_{is} 代表特征 w_s 在文本 d_i 中的出现次数。

朴素贝叶斯分类器的训练过程其实就是统计每一个特征在各类中出现规律的过程，是一种简单有效的概率分类方法，其性能可与决策树、神经网络等算法相媲美，而且速度远快于其他贝叶斯算法，特别适合文本分类这种多属性的分类任务，因而能取得较好的分类效果。

5　实验比较

5.1　文本分类数据集的确定

目前常用的文本分类数据集，路透社数据集 Reuters-21578（http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.tar.gz）；20 新闻组数据集（20newsgroups）（http://kdd.ics.uci.edu/databases/20newsgroups/20_newsgroups.tar.gz）及微型新闻组数据集（mini-newsgroups）（http://kdd.ics.uci.edu/databases/20newsgroups/mini-newsgorups.tar.gz）等。路透社数据集 Reuters-21578 由 21578 篇与经济相关的新闻报道构成，分为 135 主题类别。但因为其类别过多导致分类后低维显示效果不佳，故文中没有采用。

20 新闻组数据集包含 20 000 篇新闻文档（实际总数 19 996 篇），分布于 20 个不同的新闻组类别中。根据主题大致可进一步分为 6 个子大类，如表 1 所示。该新闻组有的类别比较相关，如 comp.sys.ibm.pc.hardware 和 comp.sys.mac.hardware，有的类别之间没有什么关系，如 misc.forsale 和 soc.religion.christian。

表 1　20 新闻组数据集子主题类分布

comp.graphics(g2)	rec.autos(g8)	sci.crypt(g12)
comp.os.ms-windows.misc(g3)	rec.motorcycles(g9)	sci.electronics(g13)
comp.sys.ibm.pc.hardware(g4)	rec.sport.baseball(g10)	sci.med(g14)
comp.sys.mac.hardware(g5)	rec.sport.hockey(g11)	sci.space(g15)
comp.windows.x(g6)		
misc.forsale(g7)	talk.politics.guns(g17)	alt.atheism(g1)
	talk.politics.mideast(g18)	soc.religion.christian(g16)
	talk.politics.misc(g19)	talk.religion.misc(g20)

微型新闻组数据集是从 20 新闻组数据集中构造出来的，其每类由从 20 新闻组数据集中精选的 100 篇文本构成，能够在较少的时间内完成分类算法的性能测试，适用于小规模实验。

5.2　20 新闻组数据集实验

首先选择整个数据集中反映单词与类别关系最高的单词（互信息最高）作为分类特征词，然后随机对数据集文档总数的 60%（11 997 篇）进行训练，对余下的 40%（7 999 篇）进行测试，依次选择互信息最高的 10，20，…，1 200 个特征词，采用朴素贝叶斯多项式模型方法进行分类，并取三次交叉验证（cross-validation）实验的平均值作为最终结果，得到不同个数特征词的分类准确率如图 2 所示。可以明显地看出，随着被选择特征词的增加，分类准确率逐渐上升。同时将 100 个特征词（选择 100 个特征词主要受到计算机内存限制，此时后验概率分布矩阵已经达到 19 996×100 维）的分类后验概率矩阵代入参数嵌入算法得到低维嵌入分类效果如图 3 所示。

图 3 中每个点代表一篇文档，不同颜色点代表不同的类别，分类效果好的算法分类后同一类在低维嵌入空间应聚类较好，相似类的数据距离较为接近，而不同类之间距离应该较大。从图可以看出同一子大类基本分布在同一条直线上，且相互间

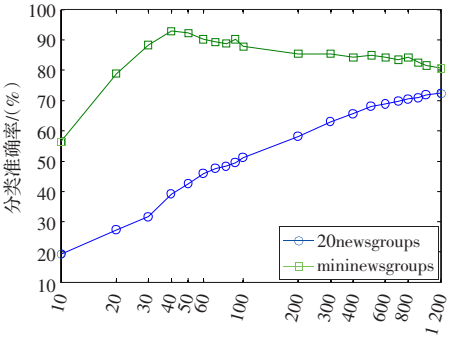

图 2　两个新闻组不同特征词个数分类准确率比较

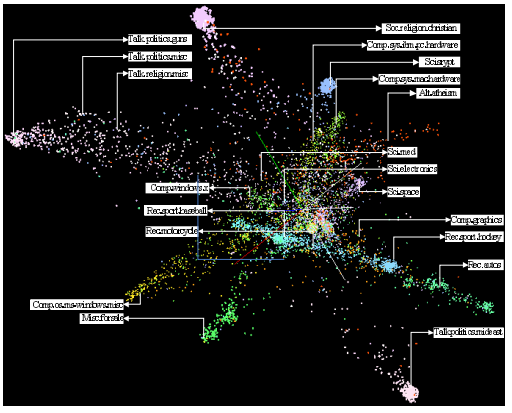

图 3　20 新闻组 100 个特征词分类效果

距离较近。不同主题类分布较远，整体结果与表 1 符合，说明参数嵌入算法能有效地显示分类结果。但是由于此时分类准确率为 51.1%，所以图 3 中有许多散点和交叉重叠的部分。增加特征词到 1 200 个时，分类准确率为 72.78%，继续增加特征词到 5 000 个和 10 000 个时分类准确度也只能达到 78.22% 和 79.14%。但由于要计算高维矩阵，实现好的分类效果图已经难以做到，故下一步在微型新闻组中进行实验。

5.3　微型新闻组数据集实验

对微型新闻组数据总数 2 000 篇的 60%（1 200 篇）进行训练，对余下的 40%（800 篇）进行测试，依次选择互信息最高的 10，20，…，1 200 个单词作为特征词，取三次交叉验证实验的分类准确率平均值作为最终结果，得到不同个数特征词的分类准确率如图 2 所示。为了比较分类低维嵌入效果，将 100 个特征词分类后验概率矩阵代入参数嵌入算法得到分类后效果如图 4 所示。

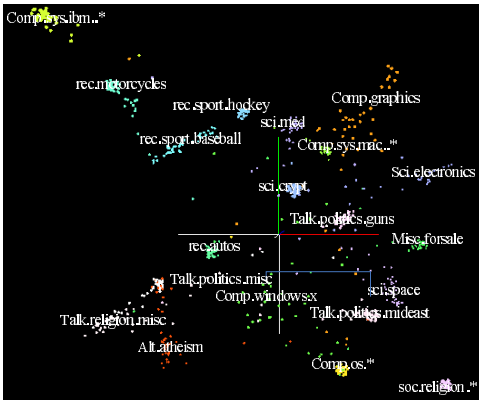

图 4　微型新闻组 100 个特征词分类效果

随着分类准确度(87.82%)的提高,分类效果出现了一些变化。不同主题子类间距离增大,同一主题类间距离变近。说明参数嵌入算法能有效地根据分类准确率显示分类结果。但是仍然有 comp.windows.x 和 talk.politics.misc 两类分布较为松散且有部分重叠,为了找出原因,进一步研究了两个新闻组各子类在 100 个特征词时分类准确率如图 5 所示。发现子类 comp.windows.x 分类准确率只有 12.7%是造成图 4 中对应投影类中数据点松散的原因, 同时图 3 中子数据类较为松散也是由于 sci.electronic 和 talk.religion.misc 分类准确率低的缘故。这是因为在分类时选取整个文本集中互信息高的词并不能完全代表某些子类所具有的特征,由于文本内容差异,在集合中有些低互信息特征词更能代表其类别特征,所以可根据结果重新选择合适的分类特征。

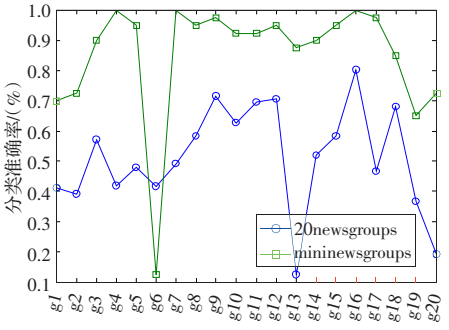

图 5　新闻组 100 个特征词时各子类分类准确率比较

5.4　最高分类准确率时分类效果

微型新闻组数据集在特征词个数为 40 时达到最高分类准确率 92.8%,此时分类效果如图 6 所示。与图 4 相比,类别间区分更加明显,散点个数减少,真实地反映了分类准确率的变化。

6　结束语

本文研究了朴素贝叶斯分类器在文本分类中的应用及特征词个数对分类准确率的影响,并通过参数嵌入算法把文本分类后效果投影到低维子空间进行直观分析,可以准确地描述子类分布情况,增加了研究的透明性,有助于对影响分类准确率的因素进行详细探讨,也可以将参数嵌入算法应用到高维数据聚类结果的可视化中。

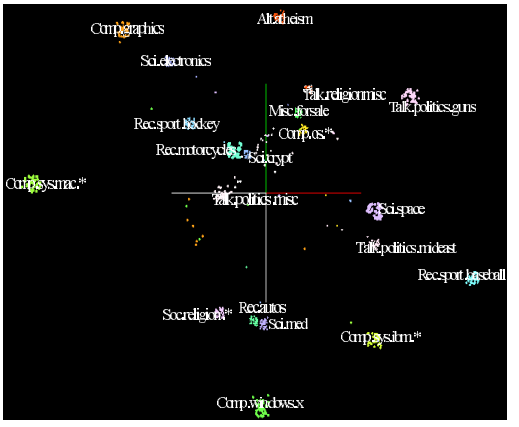

图 6　微型新闻组分类准确率最高时分类效果

参考文献:

[1] 朱望斌.自动文本分类算法研究[D].长沙:湖南大学,2005.

[2] Iwata T,Saito K,Ueda N,el al.Parametric embedding for class visualization[J].Neural Computation,2007(19):2536-2556.

[3] Hinton G E,Roweis S T.Stochastic neighbor embedding[C]//Advances in Neural Information Processing Systems.Cambridge,MA,USA:The MIT Press,2002(15):833-840.

[4] van der Maaten L J P.An introduction to dimensionality reduction using matlab[R].Maastricht University,Maastricht,The Netherlands,2007.

[5] 王小燕.文本分类相关技术与应用研究[D].西北大学,2007.

(上接 23 页)

作战计划本体等方面进行了分析。分析结果为作者进一步设计基于本体的作战计划生成和验证系统提供了有益的指导。

参考文献:

[1] 刘忠,张维明,阳东升,等.作战计划系统技术[M].北京:国防工业出版社,2007.

[2] Blythe J,Gil Y.A problem-solving method for plan evaluation and critiquing[C]//Proceedings of the 12th Banff Workshop on Knowledge Acquisition,Modeling,and Management(KAW'99),Banff,Alberta,1999.

[3] Cohen P,Schrag R,Jones E,et al.The DARPA high-performance knowledge bases project[J].AI Magazine,1998.

[4] IET.Rapid knowledge formation program[EB/OL].http://www.iet.com/Projects/RKF/.

[5] Bowman M,Lopez A M,Tecuci G.Ontology development for military applications[C]//Proceedings of the Thirty-ninth Annual ACM Southeast Conference,Athens,GA,2001:112-117.

[6] Bowman M.A methodology for modeling expert knowledge that supports teaching-based development of Agents[D].School of Information Technology and Engineering,George Mason University,Fairfax Virginia.

[7] Lenat D.CYC:A large-scale investment in knowledge infrastructure[J].Communications of the ACM,1995,38(11).

[8] Chaudhri V K,Farquhar A,Fikes R,et al.OKBC:A programmatic foundation for knowledge base interoperability[C]//AAAI Proceedings of the Fifteenth National Conference on Artificial Intelligence,Madison,Wisconsin,July 26-30,1998.

[9] Boicu M,Tecuci G,Bowman M,et al.A problem-oriented approach to ontology creation and maintenance[C]//Farquhar A,Stoffel K.Proceedings of the Sixteenth National Conference on Artificial Intelligence(AAAI-99) Workshop on "Ontology Management",July 18-19,Orlando,Florida.Menlo Park,CA:AAAI Press/MIT Press,1999.

[10] Bowman M.Center of gravity analysis:Preparing for intelligent agents[R/OL].(2001).http://www.iwar.org.uk/sigint/resources/cog-intel/Bowman_M__01.pdf.

[11] Tecuci G,Boicu M,Marcu D,et al.Development and deployment of a disciple agent for center of gravity analysis[C]//Proceedings of the Fourteenth Innovative Applications of Artificial Intelligence Conference (IAAI-2002),Edmonton,Alberta,Canada,July 28-August 1,2002.

[12] Gil Y,Blythe J.PLANET:A shareable and reusable ontology for representing plan[C]//Proceedings of the AAAI Workshop on Representational Issues for Realworld Planning Systems,2000.