第33卷 第15期 计 算 机 工 程 2007年8月
Vol.33 No.15 Computer Engineering August 2007
·人工智能及识别技术· 文章编号：1000—3428(2007)15—0217—03 文献标识码：A 中图分类号：TP391

基于LSI和SVM的文本分类研究

刘美茹
（哈尔滨铁道职业技术学院计算机教研室，哈尔滨 150086）

摘 要：文本分类技术是文本数据挖掘的基础和核心，是基于自然语言处理技术和机器学习算法的一个具体应用。特征选择和分类算法是文本分类中两个最关键的技术，该文提出了利用潜在语义索引进行特征提取和降维，并结合支持向量机(SVM)算法进行多类分类，实验结果显示与向量空间模型(VSM)结合SVM方法和LSI结合K近邻(KNN)方法相比，取得了更好的效果，在文本类别数较少、类别划分比较清晰的情况下可以达到实用效果。

关键词：特征提取；潜在语义索引；支持向量机

Research on Text Classification Based on LSI and SVM

LIU Mei-ru
(Staff Room of Computer, Harbin Railway Technical College, Harbin 150086)

【Abstract】 Text classification is the foundation and crucial problem of text data mining, it is an application based on the technology of natural language processing and machine learning. Feature extraction and categorization algorithm are the most crucial technologies for this problem. This paper proposes that latent semantic indexing (LSI) is used for feature extraction and dimensionality reduction, support vector machine(SVM) is used for text classification. The result shows that compared with the classifier based on vector space model combined SVM and the classifier based on LSI combined K-nearest neighbor (KNN), better performance is acheived. It shows that while the number of categories is small, and the categories are divided distinctly, the method can be used for practical application.

【Key words】 feature extraction; latent semantic index(LSI); support vector machine(SVM)

在因特网上大多数的信息表现形式是文本形式，而其中的文本数据缺乏结构化、组织化的规整性，大大降低了网络文本信息的利用率。文本的自动分类技术正是解决这个问题的最好方法。运用文本分类技术不仅能降低网络查询时间，提高网络搜索质量，从而快速有效地提取文本信息，同时也节省了人工分类所要消耗的大量精力和时间。同所有文本分类的研究一样，本文的目的也是建立一个快速准确的文本分类系统，有效地进行文本的类别自动标识。

1 相关工作

文本分类是一个有指导的学习过程，它根据一个已经被标注的训练文本集合，找到文本属性和文本类别之间的关系模型，然后利用这种学习得到的关系模式对新的文本进行类别判断，自动标识出文本的类别信息。

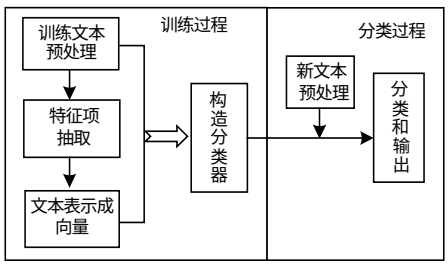

图1 文本分类系统框架

文本分类系统的框架如图1所示，文本分类总体上分为训练过程和分类过程。在训练过程中，需要进行文本预处理，特征抽取，然后将文本表示成由特征项和权重表示的向量，进而通过分类算法够造出一个分类器。在分类过程中，新的文本在预处理后，分类器根据训练过程中得到的特征信息将新文本表示成向量，然后进行分类，并输出结果。

1.1 文本预处理

文本内容是自然语言表示的，计算机难以理解其语义。文本信息源的这些特殊性使得现有的数据挖掘技术不能直接应用于其上，需要先对文本进行预处理和特征表示。在中文文本分类中，预处理主要指中文的分词。将原文本通过分词表示成词的序列，将所有的词作为候选的特征项，也有些学者提出对文本进行按定长切词，选用1~4个字长的相邻字段作为特征项。有实验表明，在很多情况下，采用两个字长的字段比单纯采用词作为特征项效果要好[1]。本文依然采用分词技术对文本进行预处理。

1.2 特征提取

目前中文文本分类主要还是选择词作为特征项，这就产生了一个特征空间维数过高的问题，如何解决维数过高和数据稀疏问题，如何筛选出最有效的特征项是目前研究文本分类最大的特点和难点之一。经常使用的特征提取的评价函数有文本频率(document frequency, DF)、chi-square (CHI)、信息增益(information gain, IG)、互信息(mutual information, MI)、term strength(TS)、GSS Coefficient、odds ratio等[2]。Yang等在Reuters21578 语料库上试验了前面 5 种方法，认为DF、CHI、IG更为有效[3,4]。国内的有些学者则认为MI>DF>IG[5]。这些方法的一个共同点特点就是假定词之间是互相独立，正

作者简介：刘美茹(1958－)，女，副教授、硕士，主研方向：自然语言理解，计算机应用

收稿日期：2007-03-13 **E-mail：**jsjwl_lmr@163.com

交的，通过计算词项和类别之间存在的某种特定关系对词进行筛选，从而达到降维的目的。

1.3 文本表示

在特征降维之后，文本需要表示成由特征项组成的向量信息，在文本分类中采用最多的是向量空间模型(VSM), 文本和整个特征集共同组成一个矩阵$\boldsymbol{A}_{mn}=(a_{ij})$，每一行代表一个特征项，每一列代表一个经过特征筛选的文档，a_{ij}的值代表第i个特征($0 \le i \le m$-1)在第j个文档($0 \le j \le n$-1)上的权重，一般采用TF·IDF进行权重计算。在计算权重之后需要进行归一化处理，式(1)表示了最终的权重计算。

$$W(t,\vec{d})=\frac{(1+\log_2 tf(t,\vec{d}))\times\log_2(N/n(t))}{\sqrt{\sum_{t\in d}\left[(1+\log_2 tf(t,\vec{d}))\times\log_2(N/n(t))\right]^2}} \tag{1}$$

其中，$tf(t,\vec{d})$是词条t在文档d 中出现的频数；N表示全部训练文档的总数；$n(t)$ 表示包含词t的文档数，称之为文档频数；而IDF= $\log_2$ (N/ $n(t)$)称为反文档频度。

1.4 分类算法

目前国内外已有很多的文本分类算法，如贝叶斯分类、K 近邻(KNN)、支持向量机(SVM)、神经网络、投票分类、决策树、线性最小方差匹配等。KNN 和 SVM 是目前使用最为广泛和相对比较有效的分类算法。

2 利用 LSI 进行特征降维

当前的VSM方法在索引和文档的检索只依赖于查询语句和文档之间的词的匹配情况，这种方法在对词的同义和近义进行扩展后会造成 2 种情况：(1)检索结果的不完整，应该检索到的没有被检索出来；(2)检索到一些无关的文本。造成这种情况主要是VSM模型只考虑词的信息，把词条当作互相独立、正交的特征，而没有考虑词和词之间在语义上的联系。前文提到的各种特征抽取的方法也是基于相同的词条之间互相独立正交的假设。事实上，文本中词条的共现情况和内在的语义结构也是重要的信息。潜在语义索引(latent semantic indexing, LSI)就是一种根据词条的共现信息探查词条之间内在的语义联系的方法[6]。通过对文档矩阵进行特殊的矩阵分解，将矩阵近似地映射到一个K维潜在语义空间上，其中，K为选择的最大的奇异值个数，映射之后的奇异值向量能最大限度的反映出词条和文档之间的依存关系。潜在语义空间实际上是把同现的词条映射到同一维空间上，而非同现的词条映射到不同的空间上，这样使得潜在语义空间相比原来的空间维数要小的多，实际上也达到了降维的目的。经过这样的映射之后，原来不包含或包含很少相同词条信息的文档之间也可能因为词条的共现关系而有较大的相似度。

本文采用数学中经典的奇异值分解的方法实现 LSI，选择这一方法的好处是分解的效果较好，且具备较强的扩展性能。

奇异值分解是将词条—文档矩阵分解为 3 个矩阵的乘积形式，即$\boldsymbol{T}_{t\times n},\boldsymbol{S}_{n\times n},\boldsymbol{D}_{d\times n}$：

$$\boldsymbol{A}_{t\times d}=\boldsymbol{T}_{t\times n}\boldsymbol{S}_{n\times n}(\boldsymbol{D}_{d\times n})^{\mathrm{T}} \tag{2}$$

其中，t为原特征空间的维数，d为文档数，n = min(t, d)，$\boldsymbol{T}$和$\boldsymbol{D}$都是正交矩阵。$\boldsymbol{S}$是一个对角矩阵，对角线的值为从大到小排列的非负实数。实际上$\boldsymbol{S}$的对角线上的值为$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$的特征值，第$i$个特征值表示在第$i$个投影轴上的偏差。只取矩阵$\boldsymbol{T}$, $\boldsymbol{S}$, $\boldsymbol{D}$的前k列，得到矩阵$\boldsymbol{T}_{t\times k},\boldsymbol{S}_{k\times k},(\boldsymbol{D}_{d\times k})^{\mathrm{T}}$。得到的A降维后的矩阵：

$$\boldsymbol{B}=\boldsymbol{S}_{k\times k}\boldsymbol{D}_{k\times d}{}^{\mathrm{T}} \tag{3}$$

特征空间从 t 维降为 k 维。

在文本分类中，存在训练文本集合扩展的问题，如果没有较好的扩展性，每次扩充训练语料集都要重新进行一次奇异值分解这显然是难以接受的。实际上，在语料库扩展或语料库过大的情况下，可以只对其中的一部分文档作 SVD 分解，而对后续的文档通过前面的计算结果进行转换，这是由以下公式推导得出的：

$$\boldsymbol{A}=\boldsymbol{TSD}^{\mathrm{T}}\Rightarrow\boldsymbol{T}^{\mathrm{T}}\boldsymbol{A}=\boldsymbol{T}^{\mathrm{T}}\boldsymbol{TSD}^{\mathrm{T}}\Rightarrow\boldsymbol{T}^{\mathrm{T}}\boldsymbol{A}=\boldsymbol{SD}^{\mathrm{T}} \tag{4}$$

对每个新的文档向量$\vec{\boldsymbol{q}}$，$\vec{\boldsymbol{q}}$在降维后的空间中表示为$\boldsymbol{T}_{t\times k}{}^{\mathrm{T}}\vec{\boldsymbol{q}}$。这样可以扩充更多的训练文本向量。而在未知文本的自动分类过程中，同样需要对经过特征表示的新文本进行相同的空间映射，使用和扩展训练文本相同的方法生成在新空间中的向量，从而使新文本在和训练文本在相同的特征空间中得到表示。

3 SVM 分类算法

SVM是近年来在统计学习理论(statistical learning theory, SLT)的VC维理论和结构风险最小原理基础上发展起来的一种新的通用学习方法。它可以根据有限的样本信息在模型的复杂性(即对特定训练样本的学习精度)和学习能力(即无错误地识别任意样本的能力)之间寻求最佳折衷,以期获得最好的推广能力(generalization ability)。SVM的基本思想可用图 2 的二维情况来说明。图中，十字和空心点代表两类样本，O_1为分线，H_1、H_2分别为过各类中离分类线最近的样本且平行于分类线的直线，它们之间的距离叫做分类间隔。所谓最优分类线就是要求分类线不但能将两类正确分开(训练错误率为0) ,而且使分类间隔最大。分类线方程为$x\bullet w+b=0$,对它进行归一化，使得对线性可分的样本集$(x_i,y_i),i=1,\ldots,n,x\in R^d,y\in\{+1,-1\}$，满足

$$y_i[(w\bullet x_i)+b]-1\ge 0,i=1,...,n \tag{5}$$

此时分类间隔等于$2/\|w\|$，使间隔最大等价于使$\|w\|^2$最小。满足式(5)且使$\|w\|^2/2$最小的分类面称作最优分类面，H1、H2 上的训练样本点就称作支持向量。利用 Lagrange 优化方法可以把上述最优分类面问题转化为其对偶问题，而在最优分类面中采用适当的内积函数$K(x_i,y_j)$就可以实现某一非线性变换后的线性分类,相应的分类函数为

$$f(x)=\mathrm{sgn}(\sum_{i=1}^{n}a_i^*y_iK(x_i,x)+b^*) \tag{6}$$

式中，b^*是分类阈值，若$f(x)>0$，则x属于该类别，否则就不属于该类。

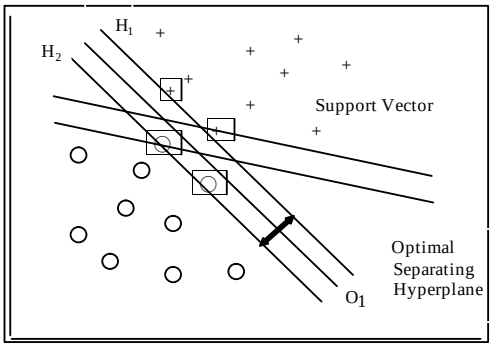

图 2 数据点集的超平面分割

4 实验结果和分析

4.1 实验数据和方案

本文采用的实验数据为复旦大学公开的一个语料库，共分 10 个类别，共 2 817 篇文本，其中训练语料 1 883 篇，测

试语料 934 篇。类别分别为：环境，计算机，交通，教育，经济，军事，体育，医药，艺术，政治。

语料库经过调用 HIT-IR 实验室的 LAS 分词系统进行分词，并去掉文档频度少于一定阈值的词条。采用 VSM+SVM 和 LSI+KNN 作为对比实验，将结果和本文的方法进行比较。VSM 模型中，采用中文分类中较好的 MI(互信息)作为特征抽取的方法，对于词条 t 和类别 c 之间的 MI 可以按下式计算：

$$MI(t,c) \approx \log \frac{A \times N}{(A+C) \times (A+B)} \tag{7}$$

其中，A 表示包含词条 t 且属于类别 c 的文档频数；B 为包含 t 但是不属于 c 的文档频数；C 表示属于 c；但是不包含 t 的文档频数，N 表示语料中文档总数。选择词条 t 和所有类别中最大的互信息值作为 t 的互信息，即

$$MI(t) = \max_{i=1}^{m} MI(t, c_i)$$

其中，m表示类别数目；c_i表示第i个类别。选择一定比例的MI值最大的词条作为最终的特征项。

在 KNN 中，通过交叉认证进行 K 值的选择，设定效果最好的 K 值。在 LSI 中，也对新空间的维数进行调整，最后设定一个相对最优值。

4.2 评价方法和结果

实验的结果如图 3 和表 1 所示。

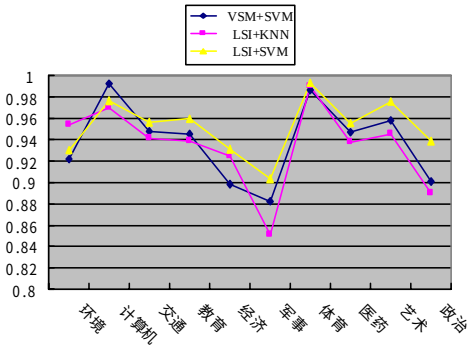

图 3　3 种方法的各小类 F1 值结果对比

表 1　3 种方法的宏平均和微平均比较

	MacroP	MacroR	MacroF1	MicroF1
VSM+SVM	0.945 68	0.932 274	0.938 93	0.938 241
LSI+KNN	0.947 861	0.924 119	0.934 267	0.932 548
LSI+SVM	0.957 362	0.947 518	0.952 414 56	0.952 891

评价的方法选择传统的正确率(P)，召回率(R)和 $F1$ 值的方法，其中

$$P_i = \frac{l_i}{m_i} \times 100\% \quad R_i = \frac{l_i}{n_i} \times 100\% \quad F1_i = \frac{P_i \times R_i \times 2}{P_i + R_i}$$

这里，l_i, n_i, m_i分别表示第i类的结果中正确的文本个数，结果中出现的个数和实际包含的文本个数。定义宏平均

$$MacroP = \frac{1}{m}\sum_{i=1}^{m} P_i, MacroR = \frac{1}{m}\sum_{i=1}^{m} R_i$$

$$MacroF = \frac{MacroP \times MacroR \times 2}{MacroP + MacroR}$$

由于只进行单类别分类(一个文档只给出一个预测类别)，因此微平均的 P, R 和 $F1$ 三者相等。

5　结论和下一步工作

实验表明，在各自取得最好结果的情况下，用 LSI 进行特征提取和降维结合 SVM 分类

算法相比其他方法有更好的效果。而且 SVM 相比 KNN 具备分类速度更快的优点。在训练文档比较有限(每类不到 200 篇)的情况下，本文的方法取得令人非常满意的效果。从准确率和召回率的结果中都可以看出，在类别数目较少，类别划分相对比较清晰的情况下，这个方法基本上可以达到实用的效果。

利用 SVD 进行 LSI 实现的时候，训练时间耗费相对于 VSM 要多一些，在下一步工作中，将深入研究 LSI 的其它实现方法，降低利用 LSI 进行文本分类的训练时间。

参考文献

1 Li Jingyang, Sun M, Zhang Xian. A Comparison and Semiquantitative Analysis of Words and Character-bigrams as Features in Chinese Text Categorization[C]//Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the ACL, Sydney. 2006: 545-552.

2 俞士汶. 计算语言学概论[M]. 北京: 商务印书馆, 2003.

3 Yang Yiming, Pedersen J 0. A Comparative Study on Feature Selection in Categorization[C]//Proceedings of the 14th International Conference on Machine Learning, San Francisco. 1997: 412-420.

4 Aas K, Eikvil L. Text categorization: A Survey[Z]. (1999). http://www.nlp. org. cn.

5 Manning. 统计自然语言处理基础[M]. 苑春法, 译. 北京: 电子工业出版社, 2005: 344-349.

（上接第 183 页）

4　结论

本文引入博弈论知识和分层图，以最小化网络资源占用率和最大化总体 QoS 满意度为目标，设计一种基于人工免疫算法的智能静态通信量疏导模式，支持网络提供方效用与用户效用的 Nash 均衡。仿真研究表明，该模式具有较好的性能。

参考文献

1 Rajagopalan, Luciani J, Awduche D, et al. IP over Optical Networks: A Framework[S]. IETF RFC 3717, 2004-05.

2 Jiao Yueguang, Zhou Bingkun, Zhang Hanyi, et al. Grooming of Arbitrary Traffic in Optical WDM Mesh Networks Using a Genetic Algorithm[J]. Photonic Network Communications, 2005, 10(2): 193.

3 Wen Haibo, Li Lemin, He Rongxi, et al. Dynamic Grooming Algorithms for Survivable WDM Mesh Networks[J]. Photonic Network Communications, 2003, 6(3): 253.

4 Zhu Hongyue, Zang Hui, Zhu Keyao, et al. A Novel Generic Graph Model for Traffic Grooming in Heterogeneous WDM Mesh Networks[J]. IEEE/ACM Transactions on Networking, 2003, 11(2): 285.

5 Varian H R. 微观经济学[M]. 北京: 经济科学出版社, 1997.

6 王　磊, 潘　进, 焦李成. 免疫规划[J]. 计算机学报, 2000, 23(8): 806.

7 赵建宏, 杨建宇, 雷维礼. 一种新的最短路径算法[J]. 电子科技大学学报, 2005, 34(6): 778.