基于二叉树多类支持向量机的文本分类研究①

Research on Text Classification Based on Binary Tree Multiclass Support Vector Machines

龙 军 王 易 刘高嵩 (中南大学 信息科学与工程学院 湖南 长沙 410075)

摘 要: 文本分类是文本数据挖掘的基础和核心，为解决在文本分类中二值支持向量机不能进行多类分类的问题，论文提出采用二叉树对多个二值支持向量机(SVM)子分类器进行组合，并运用聚类分析中类距离方法规范二叉树生成过程的基于二叉树的多类支持向量机(MSVM)分类算法。实验数据表明，相对于KNN 算法和朴素贝叶斯算法，基于二叉树的 MSVM 算法在文本分类上更具优越性。该算法已应用于科技奖励信息检索系统中，取得了良好的效果。

关键词: 向量空间模型 特征表示 特征提取 多类支持向量机 二叉树 文本分类

1 引言

20 世纪 90 年代以来，Internet 以惊人的速度发展起来，它容纳了海量的各种类型的原始信息。如何在浩若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。文本自动分类最初是应信息检索(IR)系统的要求而出现的。随着全球互联网络的普及，文本自动分类对于信息处理的意义变得更加重要。

文本分类的任务是基于内容将自然语言文本自动分配给预定义的类别,文本分类既是一种文本挖掘任务,也是对文本进行深层次挖掘的预处理步骤。本文将深入地探讨中文文本分类的关键技术,并研究当前应用于文本分类效果较好的多类支持向量机算法,将其与传统的 KNN 算法和贝叶斯算法(Bayes)相比较,证实支持向量机在文本分类上的优越性。

2 文本预处理

由于文档都是非结构化的,而且文档的内容是人类所使用的自然语言,计算机很难处理其语义,因此要进行必要的文本预处理。由于西文文本词与词之间有明显的间隔符分开的,而中文没有,中文是连续的字串,因此对中文文本预处理时还要进行文本的切分。通常采用词或者 n-grams(N-元长度为 n 的有序单词集合)法来做中文句子的切分[1]。我国对自动分词的相关研究已进行了十几年,清华大学计算机系、北京大学计算语言学研究所和哈尔滨工业大学信息检索研究室等都有接近实用的实验系统,它们的切分准确率一般可以超过 90%。

2.1 文本的特征表示

由于传统的算法不适用于处理文本信息这种非结构的数据，因此，必须将其进行结构化转换。文本的特征表示是指用文本的特征信息集合来代表原来的文本。在文本分类中采用最多的是向量空间模型(vector space model，VSM)。将文档 d_i 看作是由一组特征项 $(t_1,t_2,....t_m)$ 和相应的权重 $(w_{i1},w_{i2},.....w_{im})$ 构成的。文本和整个特征集共同组成一个矩阵 $A_{mn}=(w_{ij})_{0<i<n,0<j<m}$，每一列代表一个特征项，既为分词后的词条，每一行代表一个经过特征筛选的文档， w_{ij} 的值代表第 j 个特征(0≤j≤m-1)在第 i 个文档(0≤i≤n-1)上的权重。而权重一般采用 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。

$$w(t,d)=\frac{(1+\log_2 tf(t,d))\times\log_2(N/n(t))}{\sqrt{\sum_{f\in d}[(1+\log_2 tf(t,d))\times\log_2(N/n(t))]^2}} \quad (1)$$

① 基金项目:国家自然科学基金项目(60873081)；博士学科点专项科研基金项目(20060533084)
收稿时间:2008-10-07

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

n 个文档的文档集经过切此后有 m 个词条，VSM 表示为：

$$A=(w_{ij})_{n\times m}=\begin{bmatrix} w_{11} & \cdots & w_{1m} \\ \cdots & \ddots & \cdots \\ w_{n1} & \cdots & w_{nm} \end{bmatrix}$$

2.2 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。

2.2.1 评估函数

特征提取的基本思想通常是构造一个评估函数，按照某种算法对特征全集的每个特征属性进行评估，然后对所有的特征属性按照其评估分的大小进行排序，选取预定数目或者分数在选取范围内的特征作为结果形成特征子集。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 等[2]。国内用于中文文本分类的评估函数为 MI，DF 和 IG。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。

2.2.2 潜在语义索引(LSI)/奇异值分解(SVD)

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。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。

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。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，达到了降维的目的。采用数学中经典的奇异值分解(SVD)的方法实现 LSI。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：

$$A_{n\times m}=T_{n\times s}S_{s\times s}(D_{m\times s})^T \quad (2)$$

其中，m 为原特征空间的维数，n 为文档数，s=min(m,n)，T 和 D 都是正交矩阵。S 是一个对角矩阵，对角线上的值为 A^TA 的特征值，对角线的值为从大到小排列的非负实数。只取矩阵 T,S,D 的前 k 列(k<<min(m,n))，得到矩阵 $A_{n\times k},S_{k\times k},D_{m\times k}$。得到的 A 降维后的矩阵：

$$A_{n\times k}=T_{n\times k}S_{k\times k}(D_{k\times k})^T \quad (3)$$

特征空间从 m 维降为 k 维。通过这种方法得到一个比原始空间小得多的有效语义空间。潜在语义索引模式以其数学理论严谨、处理文本过程思路清晰得到了信息检索领域的重视，在很多实践中中都被证明是非常有效的方法[3]。

对 k 值的确定过去常常是采取实验方法，即对 k 取不同的值进行实验，观察检索的查准率与查全率。文献[3]讨论了 k 的取值。

2.3 文本间的相似度

本论文将文档 d_i,d_j 相似度 S 定义为文档向量之间的夹角余弦，此为文本挖掘中常用的相似性的度量方法。

$$sim(d_i,d_j)=\frac{\sum_{k=1}^{m}w_{ik}w_{jk}}{\sqrt{\sum_{k=1}^{m}w_{ik}^2}\sqrt{\sum_{k=1}^{m}w_{jk}^2}} \quad (4)$$

式中：m 为文档向量维数；w_{ik} 为第 k 维特征项权重。

3 多类支持向量机(SVM)

3.1 支持向量机原理

SVM 是近年来在统计学习理论的 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[4]。

SVM 的基本思想可用图 1 的两维情况来说明。图中，实心点和空心点代表两类样本，H 为分类线，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，H_1、H_2 上的点叫做支持向量。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 g(x)=x·w+b=0，对它进行归一化，使两类所有样本都满足|g(x)|≥1。使得对线性可分的样本集 (x_i,y_i)，i=1,⋯,n，$x\in R_d$，y∈{+1,−1}，满

足:

$$y_i[(w \cdot x_i) + b] - 1 \geq 0 \quad (5)$$

H_1、H_2 上的样本(叫做支持向量)使等号成立。

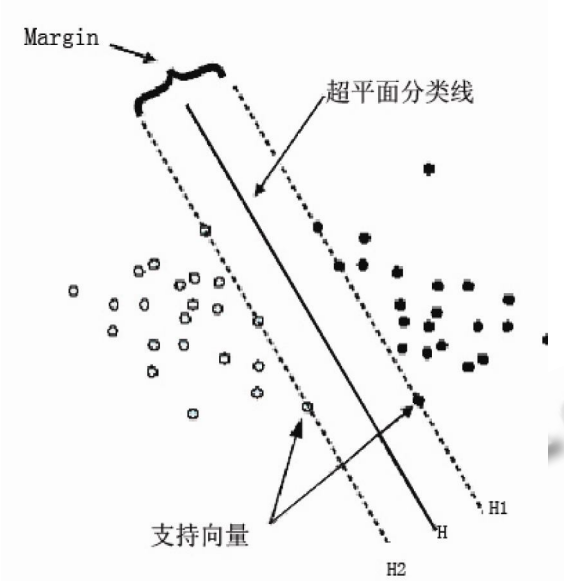

图 1　数据点集的分割

分类间隔可求得等于 2/‖w‖,使间隔最大等价于使$\|w\|^2$ 最小。满足式(5)且使$\|w\|^2/2$ 最小的分类面就叫做最优分类面,H_1、H_2 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 $K(x_i,x_j)$ 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:

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

式(6)中 a_i^* 和 b^* 为经过 Lagrange 方法求得的变量。求和实际只对支持向量进行。若 f(x)>0,则 x 属于一类别,否则就不属于该类。

3.2 基于二叉树的多类支持向量机

支持向量机作为一种分类算法已得到广泛应用。最初的 SVM 算法通过构造一个最优超平面,对两类问题进行分割。但在实际应用中，分类问题可能会多于两类样本。为了解决多类问题的分类，已经有学者提出了各种算法，这些算法的基本思想是，把多类问题划分为多个二类分类问题，通过某种方式组合这些子分类器来实现多类问题的分类。现在应用较广且性能好的算法有 one-against-one,one- against-rest 以及一次性求解等算法。[5]但这些算法会存在大量不可分区域,且训练时间和测试时间比较长[6]。

基于二叉树的多类 SVM 首先将所有类别分成两个子类，再将子类进一步划分成两个次级子类，如此循环下去，直到所有的节点都只包含一个单独的类别为止，此节点也是二叉树中的叶子，这样就得到一个倒立的二叉分类树。该方法将原有的多类问题同样分解成了一系列的两类分类问题，其中两个子类间的分类算法采用二值 svm。二叉树方法可以克服传统方法所遇到的不可分问题，并且对于 k 类分类问题只需构造 k-1 分类器，测试时并不一定需要计算所有的分类器判别函数，从而可节省测试时间。

二叉树的结构应该怎么生成？对于不同的二叉树结构，会得到不同的分类模型，当然他们的推广性能也会不同。分割顺序不一样，每个类的分割区域是不同的，从而得到的分类模型的推广能力也会不同。越是上层节点的 SVM 子分类器的分类性能对整个分类模型的推广性影响越大。因此，在生成二叉树的过程中，应该让最易分割的类最早分割出来，即在二叉树的上层节点处分割。

基于此思想，该算法利用到聚类分析中的类距离作为二叉树的生成算法。其基本思想是让与其他类相隔最远的类最先分割出来，此时构造的最优超平面也应具有较好的推广性。

定义(最短距离法)：把类 Sp 与类 Sq 中两个最近样本向量之间的欧氏距离作为两类之间的距离，即:

$$d_{p,q} = \min\left\{\left\|x_i - x_j\right\| \middle| x_i \in S_p, x_j \in S_q\right\} \quad (7)$$

4　系统设计测试结果

本文采用的实验数据为科技奖励系统的项目信息文本资料库的一部分，从其中选择共 5 个类别，共 1817 篇文本，其中训练语料 883 篇，测试语料 934 篇。以学科来分类为：生物医学，机械工程，数学，计算机，铁路交通。

实验所采用的分词方法是北京大学计算语言研究所提供的标准,分词软件 ICTCLAS 在网上可以免费获得。通过 LSI 进行特征提取，LSI 在 MATLAB 6.5 中实现。在 LSI 中，也对新空间的维数 K 进行调整，最后设定一个相对最优值。在 SVM 算法中，选择 RBF 核函数。基于二叉树的多类 SVM 算法是在 LIBSVM 工

具包的基础上修改实现的。

为了更好的说明分类算法的效果,本试验还使用了 KNN 算法和朴素贝叶斯算法(Naive Bayesian),来和支持向量机算法进行比较。

本实验结果用常用的召回率,准确率来评估。如表 1 所示:

表 1　实验测试结果

		生物医学	机械工程	数学	计算机	铁路交通
KNN	准确率(%)	92.5	80.7	85.5	78.6	79.6
	召回率(%)	82.1	90.1	84.3	78.3	81.4
Naive Bayesian	准确率(%)	69.2	79.6	90.3	86.7	85.4
	召回率(%)	89.6	77.2	89.3	92.3	82.5
MSVM	准确率(%)	96.6	90.8	93.3	94.6	89.4
	召回率(%)	94.6	88.7	95.1	94.5	96.6

从表 1 可以看出,基于二叉树的多类 VSM 的分类效果明显优于其他两种常用算法,且达到了实用水平。

5　总结

在中文文本分类过程中,用潜在语义索引(LSI)来进行特征提取,不光达到了降维的目的,还考虑了文本中内在的语义特征,实验结果明显优于以前的特征提取。在分类过程中,运用基于二叉树的多类 SVM 算法。基于二叉树的多类 SVM 首先将所有类别分成两个子类,再将子类进一步划分成两个次级子类,如此循环下去。且规定了二叉树的结构应该的生成,该算法利用到了聚类分析中的类距离,让与其他类相隔最远的类最先分割出来。通过实验证明,改进基于二叉树的多类 SVM 算法的优越性。下一步的研究重点将放在算法中参数的选择上,有 LSI 中的维数 K 的选择,SVM 中核函数的选择以及核函数中参数的选择。

参考文献

1 周水庚,关佶红,胡运发.一个无需词典支持和切词处理的中文文档分类算法.计算机研究与发展, 2001:38.

2 Sungmoon C, Sang HO, SooYoung L. Support Vector Machines with Binary Tree Architecture for MultiClass Classification. Neural Information Processing Letters and Reviews, 2004,2(3):47－51.

3 刘海峰,王元元.基于潜在语义空间的文本检索问题研究.情报科学, 2007,25(5):748－753.

4 唐发明,王仲东,陈绵云.一种新的二叉树多类支持向量机算法.计算机工程与应用, 2005,6:24－26.

5 刘志刚,李德仁.支持向量机在多类分类问题中的推广.计算机工程与应用, 2004,40(7):10－13.

6 Atkinson-Abutridy John. Combining information extrac-tion with genetic algorithms for text mining. IEEE Intelligent Systems, 2004,4:22－30.

7 包剑,冀常鹏,李义杰.基于矢量空间模型的文本自动分类系统研究.计算机系统应用, 2005,14(3):47－49.