基于CHI值特征选取和覆盖的文本分类方法

闫 屹,张燕平,耿筱媛
(安徽大学 计算机学院,安徽 合肥 230039)

摘 要:利用CHI值特征选取和前向神经网络的覆盖算法,通过对文本进行分词的预处理后,实现文本的自动分类。该方法利用CHI值进行特征选取即特征降维,应用覆盖算法进行文本分类。该方法将CHI值特征选取和覆盖算法充分结合,在提高了分类速度的同时还保证了分类的准确度。应用该方法对标准数据集中的文本进行实验,并在不同的维数上与SVM算法、朴素贝叶斯方法的实验结果进行了比较。结果表明,与SVM算法和朴素贝叶斯方法相比较,覆盖算法在准确度上更好。并且,维数的选择对分类的精确度影响很大。

关键词:文本处理;覆盖算法;文本分类
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2008)05-0079-03

Text Classification Method Based on CHI Value Feature Selection and Cover Algorithm

YAN Yi,ZHANG Yan-ping,GENG Xiao-yuan
(Computer Department,Anhui Univ., Hefei 230039,China)

Abstract:Based on CHI value feature selection and the cover algorithm of forward neural network, realizes the automatic classification of texts after the preprocessing of the texts. Based on the CHI values, the features of text set were selected firstly,namely declining dimention of features,and then text classification was processed by the cover algorithm. The method combined CHI value feature selection and the cover algorithm fully so as to promise the accurate degree of the classification at the time of raising the classification speed. Do experiment on the texts of the standard data set in this method,and compare with the experiment result of SVM and naive Bayes on the different dimention. Experiment results demonstrate that comparing with the SVM and naive Bayes, the cover algorithm do better on accurate degree. And the influence of choice of dimention to accuracy of classification is very great.

Key words:text processing; cover algorithm; text classification

0 引 言

随着互联网的发展,以其为载体的信息爆炸般地迅速增长。而文本信息占据着信息的主导地位。因而,在当前的搜索引擎和未来一代的搜索引擎的设计中,文本信息分类技术在相关的信息检索技术中具有重要的地位。文本分类是指将待分类的文本自动指定至一个或几个预定义的文本类别中,这样就大大提高了搜索的效率,节省了时间。

基于向量比较文本分类概括出来主要分为三个阶段:分词及预处理,特征降维,分类。文中在特征选择,即特征降维阶段,选择了CHI值(X^2统计量)方法,而在分类的时候利用前向神经网络的覆盖算法,通过对一组事先已经分类的文本(训练文本)的学习,构造一个文本自动分类器。再利用这个分类器去对未分类的文本进行分类。因为在分类过程中没有领域专家的干预,所以节省了大量的人力,提高了分类效率。而与CHI值特征选择的结合,既保证了分类的准确性,又提高了分类的速度。在实验阶段,比较了覆盖算法与SVM算法,朴素贝叶斯方法的准确性,并讨论了准确性随维数的变化规律。

收稿日期:2007-08-26
基金项目:国家自然科学基金(60675031,60475017);安徽省教育厅重点自然科学研究项目(2006KJ015A);安徽省教育厅自然科学研究项目(2005kj053);安徽大学211工程学术创新团队;973计划(国家重点基础研究)(2004CB318108)
作者简介:闫 屹(1982-),男,安徽合肥人,硕士研究生,研究方向为人工智能;张燕平,博士,教授,研究方向为人工智能。

1 文本分类的预处理过程

文本分类系统的任务是在给定的分类体系下,根据文本的内容自动地确定文本关联的类别。文本分类的映射规则是系统根据已经学习的每类若干样本的数

据信息，总结出分类的规律而建立的判别公式和判别规则。在遇到新文本时，可根据所得到的判别规则来确定待分类文本的类别。为完成分类任务，需要对文本进行必要的表示和预处理，在此基础上再运用分类算法进行分类。

1.1 分 词

目前在信息的分类上面，主要采用向量空间模型(VSM)。向量空间模型的基本思想是以向量来表示文本：$(W_1, W_2, \cdots, W_n)$，其中 W_i 为第 i 个特征项的权重，对于文本来说，这个特征项一般为字、词或词组。根据实验结果，普遍认为选取词作为特征项要优于字和词组，因此，要将文本表示为向量空间中的一个向量，就首先要将文本分词，将文本用词频来表示。词频分为绝对词频和相对词频，绝对词频，即用词在文本中出现的频率来表示文本；相对词频为归一化的词频，其计算方法主要运用 TF—IDF 公式，目前有多种 TF—IDF 公式，在系统中采用了一种比较普遍的 TF—IDF 公式：

$$W(t,d) = \frac{tf(t,d) \times \log(N/n_t + 0.01)}{\sqrt{\sum_{t \in d} [tf(t,d) \times \log(N/n_t + 0.01)]^2}} \tag{1}$$

其中，$W(t,d)$ 为词 t 在文本 d 中的权重，而 $tf(t,d)$ 为词 t 在文本 d 中的词频，N 为训练文本的总数，n_t 为训练文本集中出现词 t 的文本数，分母为归一化因子。

1.2 去除停用词和稀有词

在完成分词以后，每个词就是向量中的一维。在这一步就是要去除向量中的无用维。即去除对分类作用不大的虚词和形容词，一般只保留名词和动词。并且要将一些在文本中出现频率高但是含义虚泛，多意的词放入停用词表，这些词在不同的语言环境有不同的表示。另外，如果一个词在一半以上的类别中都有出现，则认为该词也应该属于停用词。在进行文本分类前，先将文本中出现在停用词表中的词去除。除此之外，有些词条在整个文档集中出现的频率都很低，也就是所谓的稀有词，它们也不适合作为文本的特征项。

2 特征降维

经过预处理，即去除停用词和稀有词后，得到的词组(向量的维数)集依然是巨大的。而维数过大会导致分类器的运算强度过大，且不同特征对分类的影响度是不同的，因此需要采用合适的特征选择算法来进行特征降维，找出需要的词组集。特征选择[1]的目的是要从分词所得到的大量词组中找出某一真子集，选择标准是此真子集在显著降低运算复杂度的同时不会使分类的准确性产生明显下降[1]。目前常见的文本降维方法有[2,3]：文本频度、互信息方法、信息增益方法、期望交叉熵方法，文本证据权方法等。文中采用了 CHI 值，即 X^2 统计量方法来进行文本的降维。CHI 值的主要思想是认为特征项与类别之间符合 X^2 分布，X^2 统计量的值越高，特征项和类别之间的独立性越小、相关性越强，特征项对这一类别的贡献越大。X^2 统计量方法中，特征值 t_k 的 CHI 权重如公式(2)所示：

$$x^2(t_k, c_i) = \frac{n[P(t_k,c_i) \times P(\overline{t_k},\overline{c_i}) - P(t_k,\overline{c_i}) \times P(\overline{t_k},c_i)]^2}{P(t_k) \times P(c_i) \times P(\overline{t_k}) \times P(\overline{c_i})} \tag{2}$$

其中 n 表示总的文本数，$p(t_k, c_i)$ 表示学习样本集中出现特征 t_k 并属于类型 c_i 的文本的概率，$P(\overline{t_k}, \overline{c_i})$ 表示样本中不属于类型 c_i 的文本中不出现特征 t_k 的文本的概率，$P(t_k, \overline{c_i})$ 表示样本中不属于类型 c_i 的文本中出现特征 t_k 的文本的概率，$P(\overline{t_k}, c_i)$ 表示样本中属于类型 c_i 的文本中不出现特征 t_k 的文本的概率，$p(t_k)$ 指出现特征 t_k 的文本概率，$p(c_i)$ 指 c_i 类文本出现的概率。$X^2(t_k, c_i)$ 度量了 t_k、c_i 的相关程度。CHI 值越大，t_k 和 c_i 就越相关，特征 t_k 对文本分类的影响就越大。使用特征选择之后，可以得到各个词组的 CHI 值，并将其由大到小排列，根据时空复杂度的需要选取一定数量的词组作为特征词组，放入特征集中。

3 覆盖算法

经过以上步骤的预处理，一个文本就变成了用特征项(词)的权重所表示的一个向量。有了这种表示以后，就可以通过前向神经网络的覆盖算法对文本进行学习和分类。覆盖算法的主要思路是[4,5]：设给定一输入集 $K = \{x_1, x_2, \cdots, x_k, x_i \in R_n\}$，$K$ 是 n 维欧氏空间的点集，设 K 分为 s 个子集 $K_1 = \{x_1, x_2, \cdots, x_m(1)\}, \cdots, K_s = \{x_m(s-1)+1, x_m(s-1)+2, \cdots, x_k\}$。

在设计网络结构时，以每个球形领域作为一个神经元，取 $\sigma(wx - \theta)$ 为其功能函数：

$$\sigma(x) = \begin{cases} 1 & x > 0 \\ 0 & \text{其他} \end{cases}$$

功能函数就是“球形领域”的特征函数。学习过程中构造第 k 类样本 X_k 的球形领域 C_i 的方法是：任取样本中尚未被覆盖的点 $x_i \in X_k$，按下式计算：

$$d^1(\omega) = \max_{x \in X_k} \{< a_i, x >\}$$

$$d^2(\omega) = \min_{x \in X_k} \{< a_i, x > | < a_i, x > > d^1(\omega)\}$$

$$d(\omega)=\frac{1}{2}(d^1(\omega)+d^2(\omega))$$

其中，$<x,y>$表示x、y的内积，这样就可得到一个以x_i为中心，以$\theta=d(i)$为阈值的覆盖c_i，按此方法求出样本的全部覆盖。

现用一个三层网络N构造分类器，即等价于求出一组领域，这组领域将不同类的点分隔开来，使属于K_i的点的输出均为$y_i=(0,\cdots,1,0,\cdots,0)$(即第$i$个分量为1,其余分量为0的向量)，$i=1,2,\cdots,s$。覆盖算法的实质是用求出的覆盖领域作为三层网络的隐含层，其中的元件就是求得的覆盖，输入层为测试样本集，输出层为测试样本集的分类结果。

覆盖算法主要分为学习算法和测试算法两部分。

3.1 学习算法

设学习样本X分为m类，即$X=\{X_1,X_2,\cdots,X_m\}$，求出样本X的所有覆盖的算法为：

步骤1，求出样本X中的最大模r，并将X中的点投影到中心在原点、半径为$2r$的球面上(为避免测试集中可能出现模更大的样本，取半径为$2r$比取r合适)。

步骤2，初始化覆盖个数$j=1$，类别数$i=1$。

步骤3，取第i类的样本，构造第j个覆盖$C(j)$。

步骤4，若X_i中没有尚未覆盖的点，转步骤7;否则,任取X_i中尚未被覆盖的一点x_i。

步骤5，按公式(2)计算，作以x_i为中心、θ为阈值的覆盖$C(j)$。

步骤6，$j=j+1$，转步骤3。

步骤7，训练结束。

该算法用Matlab 6.5实现。

3.2 测试算法

在测试时，给定一个测试样本，若它属于某类覆盖的一个球形领域,即可将该测试样本分为该类;若它不属于任何类别覆盖的任何一个球形领域，则“拒识”。

具体步骤如下：

步骤1，将待测试的样本点投影到中心在原点、半径为$2r$的球面上。

步骤2，对每个样本x，计算$d(x,C_i)=<x,x_i>$，$i=1,2,\cdots,j$。其中:x_i为覆盖C_i的中心;j为覆盖总数。

步骤3，若x只属于一个覆盖，判x属于x_i所在的类,转步骤5。

步骤4，若x不属于任一覆盖，判x为“拒识”。

步骤5，统计识别的误差率。

该算法用Matlab 6.5实现。

4 实验分析

实验数据是从UCI中文自然语言理解平台上下载的文本分类语料库，含有环境、计算机、交通、教育、经济、军事、体育、医药、艺术、政治10个类别共1786篇文章。

实验将所有的语料大概按照10:1的比例一分为二，随机选择一部分作为训练样本集，剩下的一部分作为测试样本集。进行文本降维选择特征时,首先对文本内容进行分词,统计词组出现的词频。再用统计程序进行处理，其中包括去除停用词、处理稀有词等。

本实验的学习算法中共1 479篇文档。删除停用词和稀疏词前的单词数目是22 017个。停用词数目是2 200个，停用词所占的比重是0.099。稀疏词数目是10 576个,稀疏词所占的比重是0.480，删除停用词和稀疏词后的单词总数9 241个，剩下的保留词所占的比重为0.419。测试时使用文件数目是307篇文档，删除停用词和稀疏词前的单词数目是5 708个。停用词数目是401个,停用词所占比重0.0702，稀疏词数目是2 593个，稀疏词所占比重是0.4542，删除停用词和稀疏词后的单词总数2 714，剩下的保留词所占比重为0.4754。

此实验的训练样本的维数即为9 241,下一步采用使用CHI值特征降维算法进行计算,从分词结果中挑选出适当的特征,作为分类器内容部分的特征属性集。实验中选择了从10到5000之间的几个特殊的整维数进行对比实验,对降维的程度对分类准确度的影响进行了讨论。最后分别以覆盖算法,SVM算法和朴素贝叶斯方法为分类器,并对三算法的实验结果的准确性进行了讨论。结果如表1所示。

表1　交叉覆盖、SVM和朴素贝叶斯方法的比较

维数	交叉覆盖	SVM算法	朴素贝叶斯
10	71.2301%	89.0114%	83.1860%
50	82.6475%	89.2541%	87.2435%
100	92.6187%	90.8326%	89.3546%
500	94.7306%	94.0199%	93.1698%
1000	94.8012%	94.9162%	94.5392%
1500	95.3215%	95.0987%	94.8604%
2000	96.5214%	95.4621%	95.2168%
2500	96.3168%	96.0111%	95.8746%
3000	95.8910%	95.6120%	95.3871%
3500	95.6471%	95.5821%	95.3218%
4000	95.4753%	95.2914%	94.9473%

从表中可以看出在维数很小的时候SVM算法的准确度较高,而覆盖算法和朴素贝叶斯方法较差,而从维数上升到一定程度开始,覆盖算法就在准确度上好于SVM算法和朴素贝叶斯方法。总体而言，覆盖算

(下转第85页)

用更优化的资源。开发一个健壮、有效的和能用的编程环境一般需要各个领域的协同研究,如需要测试床来测试开发环境和关于运行时的假设。

3.3 其它新技术

近几年,通讯业爆炸式增长,无处不在的手机、PDA和其它数字设备越来越多,彻底地改变了生活方式,意味着在人们周围可理解可使用的信息会不断增长。在未来几年间,对应用开发人员来说,在网格中集成新技术、新设备和新信息源变得越来越重要。数量巨大的传感器和传感器网将嵌入到桥梁、道路、衣服等等,不断地提供巨大的数据。信息的实时分析在卫生、安全、经济和其它社会活动中扮演越来越重要的角色。对网格社区来说,新设备的集成将带来软件和应用方面的挑战,同时也会给学术研究的发展带来新的方向。

3.4 管理策略和系统性

大规模的系统都是需要组织管理才能取得成功。从Internet上的复杂系统到人类心血管系统都是有组织的层次结构,通过组织结构来协调内部实体间的交互,从而确保整个系统的稳定。同样,网格也需要有效的管理策略、组织结构和系统性(an economy)来保障其稳定性和保障个体和整体(group)的性能[4]。在未来十年,一个重要的活动就是要研究、开发和辨别有效的网格管理策略、系统性和"社会结构",以确保网格的稳定性和有效性。

4 结束语

早在1994年秋季举办的COMDEX大会上,比尔·盖茨就曾经预言:"信息随手可得"。可以想象,再过几年,会出现大量Web服务和网格服务为一体的新型服务,让计算机突破时间和空间的约束,使自动获取和处理信息成为现实。比尔·盖茨的预言是极其准确的,就像他在20世纪70年代就预测PC机将占据每个人的桌面一样。

参考文献:

[1] 张 燕,赵岳松.OGSA架构下的网格服务研究[J].常州工学院学报,2004,17(6):64-68.

[2] Foster I. A Globus Primer[EB/OL]. 2006. http://www.globus.org/.

[3] Frey J, Tannenbaum T, Foster I, et al. Condor-G: a computation management agent for multi-institutional grids [C]//Proceedings of the Tenth IEEE Symposium on High Performance Distributed Computing (HPDC). San Francisco, CA:[s.n.],2001.

[4] Berman F, Fox G, Hey T. The Grid: past, present, future [EB/OL]. 2003. http://www.grid2002.org/.

[5] IBM. An architectural blueprint for autonomic computing [M]. 4th ed. US: IBM, 2006.

(上接第81页)

法和SVM算法各有其适用的场合,而朴素贝叶斯方法的实验结果就差一些。并且,这三个算法的准确度都是随着维数的上升而提升,而当维数上升到一定程度以后,维数对准确度的影响就不那么明显了,甚至有下降的趋势,分析原因是维数过大,其中难免有些特征对分类起的作用微乎其微,可以忽略不计,甚至有些特征对分类起到了反作用。本实验测定,在维数取2000~2500之间的时候准确度最高。

5 结束语

文本自动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,使用户能够更加准确地查找到所需的信息,目前它已成为人们进行信息处理的一种主要方式。

基于覆盖算法的文本表示与实现是指通过对文本中词语的切分,保留动、名词,去除停用词和稀疏词,计算每个特征项的权重,最后以向量空间模型的形式来表示文本,进行特征降维。这样能够最大程度地反映出该文本的特征,以此输入到覆盖算法中,再对文本进行分类。由于无论是对学习文本还是对测试文本,都只进行一次扫描,所以在运行时间取得了令人满意的结果。而通过与SVM算法和朴素贝叶斯方法的比较证明了覆盖算法在准确性上的优势。

参考文献:

[1] 刘丽珍,宋瀚涛.文本分类中的特征抽取[J].计算机工程,2004,30(4):14-15.

[2] Yang Y, Pedersen J O. A Comparative Study on Feature Selection in Text Categorization[C]// KDD-2000 Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Boston, MA, USA:[s.n.],2000.

[3] Mladenic D, Grobelnk M. Feature selection for unbalanced class distribution and Naive Bayes[C]// In Proc of the 16th Int'l Conference on Machine Learning (ICML'99). San Francisco: Morgan Kaufmann Publishers, 1999:258-267.

[4] 张 铃,张 钹,殷海风.多层前向神经网络的覆盖算法设计算法[J].软件学报,1999(7):66-71.

[5] Zhang L, Zhang B. A Geometrical Representation of McCulloch Pitts Neural Model and Its Applications[J]. IEEE Trans on Neural Networks, 1999, 10(4):925-929.