文本分类中的特征降维方法研究

张玉芳, 万斌候, 熊忠阳
(重庆大学 计算机学院, 重庆 400044)

摘　要: 特征降维是文本分类过程中的一个重要环节,为了提高特征降维的准确率,选出能有效区分文本类别的特征词,提高文本分类的效果,提出了结合文本类间集中度、文本类内分散度和词频类间集中度的特征降维方法。当获取特征词在文本集上的整体评价时,提出了一种新的全局评估函数,用最大值与次大值之差作为最终的评价函数值。实验比较了该方法与传统的特征降维方法,结果表明该方法在中文文本分类中具有较好的降维效果。

关键词: 文本分类; 特征降维; 集中度; 分散度; 评估函数
中图分类号: TP301.6　　**文献标志码**: A　　**文章编号**: 1001-3695(2012)07-2541-03
doi:10.3969/j.issn.1001-3695.2012.07.037

Research on feature dimension reduction in text classification

ZHANG Yu-fang, WAN Bin-hou, XIONG Zhong-yang
(*College of Computer Science*, *Chongqing University*, *Chongqing* 400044, *China*)

Abstract: Feature dimension reduction is an important part of the procedure of text categorization, in order to improve the accuracy of feature dimension reduction, select the words that can distinguish categories effectively, and ultimately improve the effect of text classification, this paper proposed a new approach for feature selection by comprehensively taking account of text concentration among classes, dispersion within the text classes and word frequency concentration among classes. While getting overall assessment of the word in text set, it proposed new function of overall assessment by using the final assessment value, which was the difference of the maximum and the second largest value. The test compared this method with the traditional feature dimension method, results indicate better effect in Chinese text categorization.

Key words: text categorization; feature dimension reduction; concentration; dispersion; assessment function

0　引言

互联网技术的普及和高速发展,使得网络上的电子文档迅速增加,在给用户带来大量信息的同时,也给用户查找、过滤和管理这些海量信息带来困难。因此,文本分类技术的研究引起了人们的持续关注。文本分类是指依据文本的内容,由计算机根据某种自动分类算法,把文本划分到预先定义好的类别[1]。

文本分类的流程大致分为五个环节,即文本预处理、特征降维、特征加权、分类器训练、分类器性能评估。一个文本集经过分词、去除停用词等文本预处理后,得到了文本集的原始特征词集合;之后进行特征降维,选出对文本类别区分能力较强的特征词,继而利用特征加权公式计算降维后的各个特征词的权重,根据向量空间模型将文本表示成了由一定数量特征词构成的空间向量;然后进行分类器训练得到分类器;最后评估分类器性能,利用性能指标对分类结果进行评价。

1　特征降维技术研究

1.1　向量空间模型

在文本分类过程中,需要利用向量空间模型(vector space model,VSM)将文本表示成由一定数量特征词构成的空间向量。向量空间模型最早由 Slaton 教授在 20 世纪 70 年代提出,该模型具有简单、高效的特点,自提出以来便成为信息检索领域最为经典的计算模型。

在向量空间模型中,文本集合的空间被视为是由一组特征词条所构成的向量空间,每个文档 d 可以由一个空间向量表示[2]。具体来说,就是对于文本集合 D,它的特征集合 T 包含了 n 个的特征词,每一个特征词 $T_i(1\leqslant i\leqslant n)$ 都是向量空间模型的一维,文本 D_i 就被表示为一个维数是 n 的空间向量 $\{W_{i1}, W_{i2}, W_{i3}, \cdots, W_{in}\}$,其中 W_{ij} 的值表示第 $j(1\leqslant j\leqslant n)$ 维特征在文本 D_i 中的权重。

1.2　特征降维的必要性

在向量空间模型中,用词条作为表示文本的特征项。空间向量的维度包含了文本集合中的所有词汇,这导致通常一个文本分类问题所对应的文本特征空间会高达几万维,甚至更高。于此同时单独一篇文本可能仅由几百个词或词组组成,由于文本特征空间高达几万维,所以一篇文本表示成的向量中很多维上的值都为 0。这造成了文本分类的两大难题,即文本特征空间的高维性和稀疏性。这两个问题对文本分类系统的分类时间和分类精度有着直接的影响,并且许多分类学习算法在计算

收稿日期: 2011-12-06; **修回日期**: 2012-01-12

作者简介: 张玉芳(1965-),女,教授,主要研究方向为数据挖掘、信息网络与系统、现代远程教育技术(zhangyf@cqu.edu.cn);万斌候(1986-),男,硕士研究生,主要研究方向为文本分类、互联网技术;熊忠阳(1962-),男,教授,博导,主要研究方向为网格与并行处理技术、数据挖掘技术与应用、互联网应用关键技术.

上是无法处理这种高维特征向量的。因此,在文本分类的过程中,特征降维就显得非常重要。利用特定的降维方法来降低文本向量空间的特征维数,不仅能够提高分类器的速度,节省存储空间,还能够过滤掉一些无关属性,减少无关信息对文本分类的干扰,从而提高分类的精度和防止过拟合。

2 常见特征降维方法研究

特征降维包括特征选择和特征抽取,常用的特征选择方法有:文档频率(DF)、互信息(MI)、信息增益(IG)、χ^2 统计量(CHI)、期望交叉熵(ECE)、文本证据权(WET)和多种方法组合等[3]。这些特征选择方法,其基本思想都是使用某种评估函数对每个特征词打分,然后把特征词按照分值从高到低排序,取分值排前的一些特征词作为降维后的特征集合。

特征选择并没有改变原始特征空间的性质,只是从原始特征空间中选择了一部分重要的特征,组成一个新的低维空间[4]。而特征抽取方法降维后得到的特征集合不是原始特征集合的子集,而是对原始集合进行合并、转换和归纳而得出的综合形式。

对于特征词 t,某个类别 C_i,各种特征选择方法的定义如下:

a)文档频率(document frequency,DF)

指在训练样本集中出现该词条的文档数。在进行特征提取时,将 DF 值高于某个特定阈值的词条提取出来,低于这个阈值的词条给予滤除。

采用 DF 作为特征降维基于如下基本假设:DF 值低于某个阈值的词条是低频词,它们不含有或含有较少的类别信息。然而这种假设与信息抽取观念有点冲突,因为在信息抽取中,有些稀有词条恰恰比那些中频词更能反映类别的特征而不应该被滤除。DF 具有相对于语料规模的线性复杂度,所以它能够容易地被用于大规模的语料特征选择[5]。

b)互信息 (mutual information,MI)

本来是信息论中的一个概念,用于表示信号之间的关系,在特征选择中用来表示词条与类别之间的关系。文本的特征 t 和类别 C_i 的互信息公式为

$$MI(t)=\sum_{i=1}^{m}P(C_i)\log\frac{P(t|C_i)}{P(t)} \tag{1}$$

采用互信息作为特征降维基于如下基本假设:在某个特定类别出现频率高,但在其他类别出现频率比较低的词条与该类的互信息比较大[6]。

c)信息增益(information gain,IG)

该模型度量了特征词在文本出现前后信息熵的改变量,它体现了某一个特征词的存在与否对类别预测的影响能力。公式为

$$IG(t)=-\sum_{i=1}^{m}P(C_i)\log P(C_i)+P(t)\sum_{i=1}^{m}P(C_i|t)\log P(C_i|t)+P(\bar{t})\sum_{i=1}^{m}P(C_i|\bar{t})\log P(C_i|\bar{t}) \tag{2}$$

信息增益的衡量标准是一个特征词 t 能够为分类系统带来多少信息,带来的信息越多,该特征词越重要。

d)χ^2 统计量

它度量了词条与文档类别之间的相关程度[7]。词条对于某个类别的 χ^2 统计量越高,表明它与该类之间的相关性越大,所携带的类别信息也就越多。公式为

$$\chi^2(t,C_i)=\frac{N\times(AD-CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \tag{3}$$

其中:A 是特征 t 和第 i 类文档共同出现的次数;B 是特征 t 出现而第 i 类文档不出现的次数;C 是第 i 类文档出现而特征 t 不出现的次数;D 是第 i 类文档和特征 t 都不出现的次数;N 为文本集合的样本总数。采用 χ^2 统计量作为特征降维基于如下基本假设:特征 t 和文档类别 C 之间符合具有一阶自由度的 χ^2 分布[8]。

3 CD 特征选择方法

本文研究了现有的特征选择方法之后,提出了一种新的特征选择方法。该方法是通过文本类间集中度、词频类间集中度和文本类内分散度三个参数来计算特征评估函数值。为方便表示,取集中度和分散度的英文单词 concentration 和 dispersion 的首字母为该特征选择方法的简写形式,该方法公式为

$$CD(t,C_i)=\frac{df_i(t)}{\sum_{i=1}^{m}df_i(t)}\times\frac{tf_i(t)}{\sum_{i=1}^{m}tf_i(t)}\times\frac{df_i(t)}{|C_i|} \tag{4}$$

其中:$df_i(t)$ 表示特征词 t 在 C_i 类出现的文本数量;$\sum_{i=1}^{m}df_i(t)$ 表示特征词 t 在所有类中出现的文本数量;$tf_i(t)$ 表示特征词 t 在 C_i 类中出现的词频数量;$\sum_{i=1}^{m}tf_i(t)$ 表示特征词 t 在所有类中出现的词频数量;$|C_i|$ 表示 C_i 类中的所有文本数量。

$df_i(t)/\sum_{i=1}^{m}df_i(t)$ 代表的是文本类间集中度,即特征词 t 在 C_i 类中出现的文本数量与在所有类中出现的文本数量的比值,比值越大,说明特征词 t 越集中在 C_i 类中,对 C_i 类的表征作用越强。

$tf_i(t)/\sum_{i=1}^{m}tf_i(t)$ 代表的是词频类间集中度,即特征词 t 在 C_i 类中出现的频数与在所有类中出现的频数的比值,比值越大,说明特征词 t 越集中在 C_i 类中,对 C_i 类的表征作用越强。

$df_i(t)/|C_i|$ 代表的是文本类内分散度,即特征词 t 在 C_i 类中出现的文本数量与在 C_i 类中的所有文本数量的比值,比值越大,说明特征词 t 越均匀分布在 C_i 类中,对 C_i 类的表征作用越强。

对有 m 个类别的文本集而言,通过式(4)计算后,特征词集合中的每个特征词 t 得到 m 个 $CD(t,C_i)$ 值,为了得到特征 t 在文本集上的整体评价,由以下两种方式来计算:

$$CD_{\text{avg}}(t)=\sum_{i=1}^{m}P(c_i)CD(t,c_i) \tag{5}$$

$$CD_{\max}(t)=\max_{i=1}^{m}\{CD(t,c_i)\} \tag{6}$$

其中:式(5)的取值为特征词在各类别的平均 CD 值,式(6)的取值为特征词在各类别 CD 值中的最大值。

经过实验分析,在利用公式(7)计算 t 在文本集上的整体评价时有更好的效果。公式为

$$CD(t)=\max_{i=1}^{m}\{CD(t,c_i)\}-\operatorname{smax}_{i=1}^{m}\{CD(t,c_i)\} \tag{7}$$

其中:$\max_{i=1}^{m}\{CD(t,c_i)\}$ 为 m 个 $CD(t,C_i)$ 值中的最大值(假设对应的类为 A 类),$\operatorname{smax}_{i=1}^{m}\{CD(t,c_i)\}$ 为 m 个 $CD(t,C_i)$ 值中的次

大值(假设对应的类为 B 类)。

特征选择其实是挑选出能把某一类别与其他类别区分出来的特征词,这种区分性越大,这个词所获得的特征函数值也就越大。所以根据$\max_{i=1}^{m}\{CD(t,c_i)\}$的定义,特征词 t 对 A 类的表征作用最强,但这并不代表 t 对类别的区分能力强。因为如果 t 对多个类别的表征作用都很强,即 $CD(t,C_i)$ 值相差非常小,那么 t 对类别的区分能力其实并不强。

因此,如果$\max_{i=1}^{m}\{CD(t,c_i)\}$与$s\max_{i=1}^{m}\{CD(t,c_i)\}$的差值越大,即特征词 t 对 A 类的表征作用越强,同时对其他类别(包括表征作用次强的 B 类)的表征作用越小,则特征词 t 对类别的区分能力就越强,越应该保留在降维后的特征集合中。

4　实验及分析

本文采用的语料集来自于复旦大学收集的文本分类语料库,从该语料库随机抽取了 7 类共 4 851 篇文本作为文本集合,其中历史类 465 篇、计算机类 1 350 篇、农业类 870 篇、政治类 1 023 篇、体育类 681 篇、环境类 390 篇、军事类 72 篇。文本集合中训练集和测试集的比例为 2∶1。

在实验中,文本预处理阶段采用了 ICTCLAS 对文本集合中的数据进行分词,用 Lucene 建立全文索引;特征降维阶段采用了效果较好的文档频、互信息、信息增益和 CHI 与 CD 方法作比较;特征加权阶段采用了经典的 TF-IDF 加权方法[9];分类阶段采用了性能较好的 KNN 分类算法进行验证[10]。在评价分类效果时,常用的性能评价方法有:查全率、查准率和 F_1 值。其中:

查准率 = 分类的正确文本数/实际分类文本数

查全率 = 分类的正确文本数/类内应有文本数

查全率考察的是分类器分类结果的完备性,查准率考察的是分类器分类结果的正确性。将两个因素综合在一起考虑的方法就是采用 F_1 值,计算式为

$$F_1=\frac{\text{查准率}\times\text{查全率}\times 2}{\text{查准率}+\text{查全率}}$$

为便于比较,实验中采用 F_1 值作为性能评价方法。

在实验中,先比较了 CDmax(t) 、CDavg(t)和 $CD(t)$三种全局评估函数的优劣,然后采用了传统的 DF、MI、IG 等方法与 CD 方法作比较。结果分别如图 1 和 2 所示。

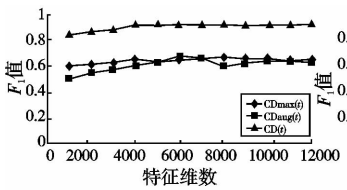

图1　三种全局评估函数对应的 F_1 值曲线

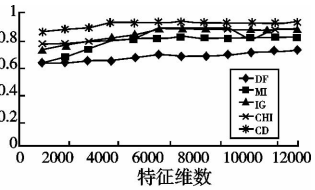

图2　各种特征选择方法对应的 F_1 值曲线

从图 1 可以看出,在相同的训练集和分类方法下,随着特征维数的逐渐增大,三种全局评估函数的 F_1 值均呈现出先逐步增大再趋于平稳的过程。其原因是如果特征维数太小,在用特征词表示文本时就不够精确,最终影响分类的效果。而如果维数超过了一定的范围,文本表示已经够精确了,分类效果并不会因为维数的增加而提高,而且高维数带来的计算开销也是不容忽视的。在 F_1 值上,采用 $CD(t)$全局评估函数得到的 F_1 值效果最好,CDmax(t) 次之;CDavg(t) 效果最差,其原因 CDavg(t)更倾向于选择在多个类别中均表现良好的特征词,而这些特征词很有可能属于多个类别,对类别的区分度不大。因此可以把 $CD(t)$全局评估函数作为最终的特征评估函数。

从图 2 可以看出,随着特征维数的逐步增大,DF、MI、IG 等方法的 F_1 值均呈现出先逐步增大再趋于平稳的过程。在 F_1 值上,DF 方法表现最差,其原因是 DF 方法仅仅考虑了特征词的词频,没有考虑特征与类别之间的关系。MI 表现一般,IG 和 CHI 表现较好且效果相当;CD 方法在初始维数很低的情况能得到较优的 F_1 值,随着维数的增加 F_1 值的波动不是很大,对特征维数的适应性很好,而且整体效果优于其他四种方法。从实验数据可知,本文提出的结合集中度和分散度的特征降维方法效果还是很好的。

5　结束语

特征降维是文本分类必须要面对的主要问题之一,是制约提高文本分类效率的瓶颈。本文提出的 CD 特征选择方法计算简单,能有效地选出区分类别的特征词。但该方法只考虑了特征词与类别之间的相关性,没有考虑特征词之间的近义和冗余关系。而特征抽取通过把测量空间的数据投影到特征空间,可以有效地解决这一问题。因此,如何有效地结合 CD 特征选择方法与特征抽取方法,使特征降维效果达到最佳将是下一步的研究内容。

参考文献:

[1] 苏金树,张博锋,徐昕. 基于机器学习的文本分类技术研究进展[J]. 软件学报,2006,17(9):1848-1859.

[2] JANA N, PETR S, MICHAL H. Conditional mutual information based feature selection for classification task[C]//Proc of the 12th International Congress on Pattern Recognition. Berlin: Springer-Verlag,2007:417-426.

[3] SANTANA L E A, De OLIVEIRA D F, CANUTO A M P, *et al*. A comparative analysis of feature selection methods for ensembles with different combination methods[C]//Proc of Internation Joint Conference on Neural Networks. Piscataway: IEEE Press, 2007:643-648.

[4] LI Fang-tao, GUAN Tao, ZHANG, Xian, *et al*. An aggressive feature selection method based on rough set theory[C]//Proc of the 2nd International Conference on Innovative Conputing, Information and Control. Washington DC: IEEE Computer Society, 2007:176-179.

[5] 徐燕,李锦涛,王斌,等. 文本分类中特征选择的约束研究[J]. 计算机研究与发展,2008,45(4):596-602.

[6] 范小丽,刘晓霞. 文本分类中互信息特征选择方法的研究[J]. 计算机工程与应用,2010,46(34): 123-125.

[7] 靖红芳,王斌,杨雅辉,等. 基于类别分布的特征选择框架[J]. 计算机研究与发展,2009,46(9):1586-1593.

[8] 代劲,何中市,胡峰. 基于云模型的文本特征自动提取算法[J]. 中南大学学报:自然科学版,2011,42(3): 714-720.

[9] 施聪莺,徐朝军,杨晓江. TFIDF 算法研究综述[J]. 计算机应用,2009,29(6):167-170.

[10] 奉国和. 四种分类方法性能比较[J]. 计算机工程与应用,2011,47(8):25-27.