文章编号:1003 - 0077(2005)06 - 0021 - 07

基于类别特征域的文本分类特征选择方法*

赵世奇[1],张 宇,刘 挺,陈毅恒,黄永光,李 生

(哈尔滨工业大学 信息检索研究室,黑龙江 哈尔滨 150001)

摘要:特征选择是文本分类的关键问题之一,而噪音与数据稀疏则是特征选择过程中遇到的主要障碍。本文介绍了一种基于类别特征域的特征选择方法。该方法首先利用"组合特征抽取"[1]的方法去除原始特征空间中的噪音,从中抽取出候选特征。这里,"组合特征抽取"是指先利用文档频率(DF)的方法去掉一部分低频词,再用互信息的方法选择出候选特征。接下来,本方法为分类体系中的每个类别构建一个类别特征域,对出现在类别特征域中的候选特征进行特征的合并和强化,从而解决数据稀疏的问题。实验表明,这种新的方法较之各种传统方法在特征选择的效果上有着明显改善,并能显著提高文本分类系统的性能。

关键词:计算机应用;中文信息处理;文本分类;特征选择;类别特征域

中图分类号:TP391 **文献标识码**:A

A Feature Selection Method Based on Class Feature Domains for Text Categorization

ZHAO Shi-qi, ZHANG Yu, LIU Ting, CHEN Yi-heng, HUANG Yong-guang, LI Sheng

(Information Retrieval Laboratory, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China)

Abstract: Feature selection is one of the key problems in text categorization. The chief obstacles to feature selection are noise and sparseness. This paper presents a novel feature selection method which is based on class feature domains. First, we will make use of the combined feature selection method[1] to remove noisy features from the original feature space and extract candidate features. That is, we'll take off low frequency words using Document Frequency method firstly and then select candidate features using Mutual Information method. Then, we will construct a class feature domain for each class and conquer the sparseness of trainning datas by merging and strengthening the candidate features which appear in the class feature domains. Experiments show that our method is much better than kinds of traditional feature selection methods and it can improve the performance of text categorization systems markedly.

Key words: computer application; Chinese information processing; text categorization; feature selection; class feature domains

1 引言

文本分类是指在给定的分类体系下,对未知类别的文档进行自动处理,并根据文档特征来判断其所属类别的过程。近年来,随着网上电子文档的数量以指数级的速度增长,文本分类技术在信息检索、信息过滤以及内容管理等各项应用中变得越来越重要。

* **收稿日期**:2004 - 11 - 24 **定稿日期**:2005 - 06 - 20
基金项目:国家自然科学基金重点资助项目(60435020)
作者简介:赵世奇(1981—),男,硕士生,主要研究方向为信息检索.

文档特征的选择和提取是文本分类的必要前提。最基本的特征选择方法是将一篇文档中出现的全部词条作为这篇文档的特征。这样,原始的特征空间便是由可能出现在文档中的所有词条构成。由于中文的词条总数超过二十万[1],因此原始特征空间的维数将十分巨大,这种特征空间的高维性将直接带来数据稀疏的问题。同时,原始特征空间中还含有许多起干扰作用的噪音特征。因此,找到一种更为有效的特征选择方法显得至关重要。

针对上述问题,本文提出了一种新的特征选择方法——基于类别特征域的特征选择方法。该方法首先利用文档频率和互信息等方法移除原始特征空间中的噪音特征,而将剩余的部分作为候选特征保留下来。在此基础上,再为每个类别构建一个"类别特征域",对于那些包含于各个类别特征域的候选特征,我们将对其进行合并和强化,其目的是使那些对于分类更为重要的特征得以突显,以避免数据稀疏。我们通过大量实验对这种新的方法与一些传统的特征选择方法进行了比较,实验结果表明:该方法在特征抽取方面的表现明显好于其它几种方法,而且文本分类系统在应用了这种特征选择方法之后其分类效果也有了显著的提高。

2 特征选择方法的相关研究

在文本分类中常用的特征选择方法有:文档频率 DF、信息增益 IG、互信息 MI 以及 CHI 统计等[1~3]。其中:文档频率的方法最为简单,该方法首先计算每个词条的 DF 值(一个词条的 DF 值即为训练语料中包含该词条的文档数),然后将 DF 值低于某一特定阈值的词条从原始特征空间中移除,从而达到降维和特征选择的目的。信息增益的方法考察的是每个词条对于整个分类所能提供的信息量。一个词条的信息增益越大,则说明该词条对于分类提供的信息量越大。该方法也正是选择那些信息增益高于特定阈值的词条作为文本分类的特征。互信息的方法则是考察每个词条与各个类别的相关度,一个词条与一个类别的互信息越大,则说明二者关联越紧密。通常的做法是:首先分别计算一个词条与每个类别的互信息,然后选择其中最大的一个作为这个词条的互信息值。特征选择时依然是设定一个阈值,将互信息值高于这一阈值的词条作为特征保留下来。CHI 统计是通过度量词条与类别间的独立性的缺乏程度来进行特征的选择。简单说,词条对于某类的 CHI 统计值越高,它与该类的独立性越小,相关性越大。

上述几种特征选择方法有一个共同的特点,即它们都能够较为有效地去除特征空间中的噪音,但在解决数据稀疏问题上却没有令人满意的表现。鉴于这个原因,有人尝试将隐性语义标引 LSI[4]以及词聚类[5]等方法用于特征选择的问题当中来。这两种方法的共同点在于都是力求挖掘出词条与词条间的潜在的语义关系,以解决传统方法中利用"词袋"(即假定词条与词条之间相互独立,彼此没有关系)作为特征所带来的数据稀疏问题。

3 类别特征域及其构建方法

本文在解决数据稀疏的问题上引入了一种新的方法,我们称之为基于"类别特征域"的方法。首先,我们解释一下什么叫"类别特征域"、为什么要构建类别特征域以及如何构建类别特征域。

3.1 类别特征域

如前所述,数据稀疏的根本原因在于原始特征空间的维数过于巨大,语言内容丰富、使用灵活。然而我们应当看到,这种特征空间的高维度和语言使用的灵活多样对于分类来讲实际上是利弊参半的。一方面,对于那些对分类比较重要的特征词条来讲,上述特性确实带来了数据稀疏问题;但另一方面,对于那些对分类没有帮助的"噪音"词条,上述特性又可以起到弱化

和“屏蔽”噪音的作用。因此,我们很自然地想到,在特征选择时应尽可能地突显主要特征,而保持次要特征和噪音的稀疏性。在实际操作中,就是要从巨大的原始特征空间中提取出一个包含主要特征的子集(我们称之为“主要特征集”),然后仅对这个子集中的特征进行某种“强化”,以使其突显。值得注意的是,这种“有选择地进行特征强化”的策略与潜在语义标引和词聚类的方法是有着明显区别的。

问题是如何从巨大的原始特征空间中准确地挑选出主要特征集。显然,总的主要特征集应该是分类体系中各个类别的主要特征集的并集。这里,我们将每个类别的主要特征集称作它的类别特征域。这样,问题便由构建总的主要特征集转化为构建每个类的类别特征域。

3.2 利用训练语料构建类别特征域

从类别特征域的定义中容易看出,每个类的类别特征域中包含了与这个类别关联紧密的若干主要特征,而互信息恰好可用于衡量词条与类别的关联程度。因此,我们通过计算训练语料中每个词条与某个类别的互信息来确定这个词条是否属于该类的类别特征域。

对于词条 t 和类别 c,如果用 A 表示包含词条 t 且属于类别 c 的文档数,B 表示包含 t 但不属于 c 的文档数,C 表示属于 c 但不包含 t 的文档数,N 表示训练语料中的文档总数,则 t 与 c 的互信息可以利用下面的公式计算:

$$MI(t,c) \approx \log \frac{A \times N}{(A + C) \times (A + B)} \tag{1}$$

我们事先设定一个阈值,当 $MI(t,c)$ 高于该阈值时,便可认为词条 t 属于类别 c 的类别特征域。利用这一方法,就可以为每个类别构建类别特征域。需要说明的是,某些词条可能成为多个类别的特征(例如“击败”一词,可能既是“军事”类的特征,又是“体育”类的特征),这样的词条将属于多个类的类别特征域。当然,在类别特征域构建完成后可通过人工校对来进一步提高其精确性。

事实上,可将上述过程看作是在分类之前首先建立起分类体系中各个类别的领域知识,而这些领域知识具体表现为每一类的主要特征的集合,即每一类的类别特征域。

4 基于类别特征域的特征选择方法

本章将详细介绍如何将类别特征域的方法用于文本分类的特征选择当中,以解决数据稀疏问题。另外,正像上面提到的,文档频率、互信息等方法在去噪音方面效果明显。因此,在特征选择时,我们将文档频率及互信息的方法与类别特征域的方法结合了起来,这样既去除了噪音,又避免了数据稀疏。概括地讲,我们的特征选择包括以下三个主要步骤:a) 预处理;b) 候选特征的抽取;c) 利用类别特征域进行特征合并。

4.1 预处理

预处理步骤主要是对训练文档进行分词,并去掉其中的停用词。这里的停用词既包括那些对于分类没有意义的各种符号、虚词以及部分代词、副词,又包括诸如“网站”、“新闻”、“下载”等当前网络文档中的常用词汇。

4.2 候选特征的抽取

候选特征的抽取是指从原始特征空间(去掉了停用词)中初步筛选出那些可能作为特征的词条。该步骤的主要目的在于降维和去噪音。

由于信息增益、互信息等方法对低频词有不同程度的倚重,势必会影响候选特征的抽取效果,因此有必要先去掉原始特征空间中的低频词。基于上述考虑,本方法分两步来抽取候选特

征。首先,利用文档频率 DF 移除原始特征空间中 DF 值低于 5 的所有词条。在此基础上,利用互信息 MI 再去掉一部分的原始特征(互信息的计算公式见公式(1))。需要说明的是,由于分类体系中含有多个类别,因此需要分别计算词条 t 与每个类别 c_i 的互信息,并取其中的最大值作为这个词条的互信息,用公式表示为:

$$MI_{\max}(t) = \max_{i=1}^{n} MI(t, c_i) \quad (n \text{ 为类别数}) \tag{2}$$

在计算得到每个词条的最大互信息之后,我们将按照从大到小的顺序对其进行排序,并将排在前面的一定比例的词条作为候选特征保留下来。应该说,我们这里所采用的是一种"组合特征抽取"的方法($DF + MI$),已有实验证明,在中文特征选择问题中,这种组合特征抽取的方法要优于单独使用 DF、MI、IG 以及 CHI 等方法[1]。

4.3 利用类别特征域进行特征合并

前面介绍了类别特征域的概念以及构建方法,本节里,我们将看到类别特征域是如何被用于特征选择的。

首先,我们为分类体系中的每一类都规定一个类别码(设类别 c_i 的类别码为 d_i)。接下来,对于训练文档 D 的每一个候选特征 t,将分下列三种情况对其进行处理:

1) 若 t 属于且仅属于类别 c_i 的类别特征域,则 c_i 的类别码 d_i 将被作为训练文档 D 的一个特征;

2) 若 t 同时属于 m 个类别 $c_{i1}, c_{i2}, \cdots\cdots, c_{im}$ 的类别特征域,则上述 m 个类别的类别码 $d_{i1}, d_{i2}, \cdots\cdots, d_{im}$ 都将被作为文档 D 的特征;

3) 若 t 不属于任何一类的类别特征域,则 t 本身将被作为该训练文本的一个特征。

我们可以通过一个例子来理解上述过程:假设某篇文本 D 所含有的候选特征及其相应权值为"战场:w1;双方:w2;防守:w3;进攻:w4;战士:w5;两方:w6;轰炸:w7",且"战场"、"战士"、"进攻"、"防守"以及"轰炸"等候选特征属于"军事类(设类别码为'A')"的类别特征域;"进攻"和"防守"这两个候选特征同时又属于"体育类(设类别码为'B')"的类别特征域;"双方"和"两方"这两个候选特征不属于任何一个类别特征域。则经过合并后得到的文本 D 的特征及权值将是"A:w1 + w3 + w4 + w5 + w7;双方:w2;B:w3 + w4;两方:w6"。从这个结果很容易看出:

1) 经过合并之后,文本 D 的主要特征"A"被突显出来;

2) 这种方法也兼顾了"进攻"和"防守"作为体育类特征的可能;

3)"双方"和"两方"这两个候选特征虽然几乎同义,但因为它们不属于任一类别特征域,所以未对其进行合并,即保持了这两个候选特征的稀疏性。

从上述过程中可以看出:训练文档中那些包含于各类的类别特征域中的候选特征将被替换为其对应类的类别码。而在一篇文档中,当有多个候选特征对应同一类别码时,它们的权重将被累加起来作为这个类别码在这篇训练文档中的权重(关于特征权重的问题我们将在后面的实验部分作详细介绍),这无疑起到了突显主要特征的作用。另外,本方法只将各类的类别特征域中的候选特征替换为类别码,其余候选特征保持不变,这也正反映了前面提到的"有选择地进行特征强化"的想法。

5 实验结果及分析

本章将通过实验来证明上述基于类别特征域的特征选择方法的有效性。

5.1 实验设置

1. 分类体系与数据集

本文采用了“中图分类法”[6]的分类体系。中图分类法包含 38 个类别,我们使用了其中除“T 工业技术”和“Z 综合性图书”之外的 36 个类别。在实验数据方面,我们收集了 7333 篇文档,用其中的 5533 篇作为训练语料,其余的 1800 篇作为测试语料。

2. 分类器

本实验使用支持向量机(Support Vector Machines(SVM))[7]作为文本分类器。SVM 是公认的效果最好的文本分类器之一。在训练阶段,SVM 从训练文档中寻找区分能力最强的样本点集(支持向量);在分类阶段,SVM 利用这些支持向量对未知类别的测试文档进行分类。实验中核函数采用了 RBF 函数,并采用交叉检验(cross validation)的方法确定其中的参数 C 取 16,γ 取 0.25。另外,在具体实现多值分类时,我们采用了“一对一”(one-against-one)的模式。即对于本分类体系的 36 个类别,共训练 $36\times(36-1)/2$ 个分类器,其中每个分类器针对两个类别的数据进行训练。

3. 文本的表示

本实验对训练文档及待分类文档的表示采用的是向量空间模型(VSM)[8~10]。在计算文档的特征权重时,运用的是 TF-IDF 公式。目前存在多种 TF-IDF 公式,本文所使用的是一种较为普遍的 TF-IDF 公式:

$$W(t,\vec{b}) = \frac{tf(t,\vec{b})\times\log(N/n_t+0.01)}{\sqrt{\sum_{t\in\vec{b}}[tf(t,\vec{b})\times\log(N/n_t+0.01)]^2}} \tag{3}$$

其中,$W(t,\vec{d})$ 为词条 t 在文本档 $\vec{d}$ 中的权重,$W(t,\vec{d})$ 为词条 t 在文本 $\vec{d}$ 中的词频,N 为训练文档总数,n_t 为训练文档集中出现词条 t 的文档数,分母为归一化因子。

5.2 评价方法

由于特征选择效果的好坏直接决定着最终分类的结果,因此可以通过评价分类的效果来测试特征抽取方法的有效性。这里,我们采用的是准确率 P(Precision)、召回率 R(Recall)以及 $F1$ 值等通用的性能评价方法,其数学公式表示如下:

$$\text{第 } j \text{ 类的准确率}: P_j = (l_j/m_j)\times 100\% \tag{4}$$

其中,l_j 为第 j 类分类正确的文本数,m_j 为分类系统实际分类为 j 的文本数。

$$\text{第 } j \text{ 类的召回率}: R_j = (l_j/n_j)\times 100\% \tag{5}$$

其中,l_j 为第 j 类分类正确的文本数,n_j 为第 j 类实际包含的文本数。

$$\text{第 } j \text{ 类的 } F1 \text{ 值}: F1_j = \frac{P_j\times R_j\times 2}{P_j+R_j} \tag{6}$$

由于分类体系中包含多个类别,因此采用微平均和宏平均两种计算准确率、召回率和 $F1$ 值的方法。具体定义如下:

$$\text{宏平均准确率}: MacroP = \frac{1}{n}\sum_{j=1}^{n}P_j \tag{7}$$

$$\text{宏平均召回率}: MacroR = \frac{1}{n}\sum_{j=1}^{n}R_j \tag{8}$$

$$\text{宏平均 } F1 \text{ 值}: MacroF1 = \frac{MacroP\times MacroR\times 2}{MacroP+MacroR} \tag{9}$$

$$\text{微平均准确率}: MicroP = \sum_{j=1}^{n} l_j \Big/ \sum_{j=1}^{n} m_j \tag{10}$$

$$\text{微平均召回率}: MicroR = \sum_{j=1}^{n} l_j \Big/ \sum_{j=1}^{n} n_j \tag{11}$$

$$\text{微平均 } F1 \text{ 值}: MacroF1 = \frac{MicroP \times MicroR \times 2}{MacroP + MicroR} \tag{12}$$

5.3 实验结果

文档频率 DF、互信息 MI、信息增益 IG 以及 CHI 统计是在英文文本分类中公认的效果较好的特征选择方法,而文章[1]认为:对于中文而言,由于原始特征空间维数比英文更高,可能出现的低频词更多,因此可先利用文档频率 DF 去掉部分低频词,然后再使用 MI、IG、CHI 等方法,这样会使特征选择的效果进一步提高。进而,该文章通过实验证明了这种"组合特征抽取"的有效性。

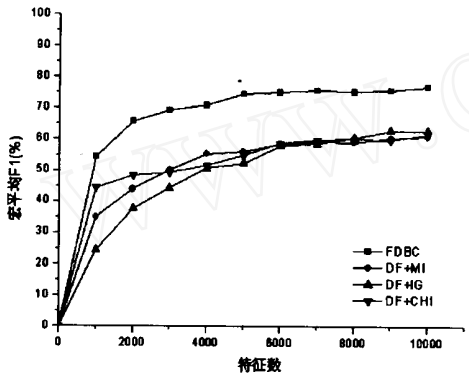

图 1　四种方法宏平均 F1 值的比较

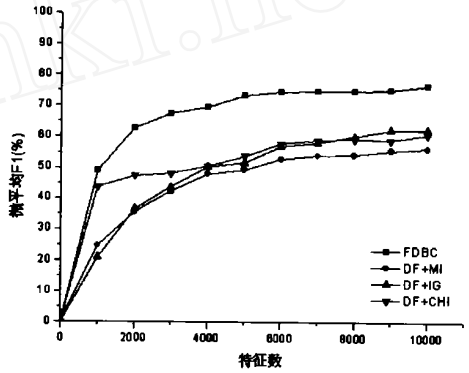

图 2　四种方法微平均 F1 值的比较

基于上述结论,我们在实验中对 FDBC(即本文介绍的基于类别特征域的特征选择方法)、DF + MI、DF + IG 以及 DF + CHI 等四种方法进行了比较。实验结果如图 1、图 2 所示。

我们在实验中分别比较了四种方法在保留 1000,2000,……,10000 个特征时,分类系统的宏平均 F1 值和微平均 F1 值。实验结果显示:对于各不同数量的特征集合,FDBC 的效果始终明显好于其余三种方法。下表展示了在保留 10000 个特征时,应用上述四种方法的分类系统的宏平均 F1 值和微平均 F1 值:

表 1　保留 10000 个特征时,四种方法的宏平均 F1 值和微平均 F1 值

	FDBC	DF + MI	DF + IG	DF + CHI
宏平均 F1	76.99 %	60.93 %	62.81 %	61.51 %
微平均 F1	76.61 %	56.22 %	62.28 %	60.61 %

从实验中我们发现:当保留 10000 个特征时,FDBC 的宏平均 F1 值和微平均 F1 值均达到 76 %以上,而其余三种方法却仅能达到 60 %左右。之所以出现怎么大的差距,主要是由于以下几点原因:

第一,正如前面讲到的,DF + MI,DF + IG 及 DF + CHI 等方法在去除噪音方面表现出色,但

不能很好地解决数据稀疏问题。加之本实验中所使用的训练语料仅有 5500 余篇,规模相对较小,这使得数据稀疏问题变得尤其突出。由于这些原因,DF + MI,DF + IG 及 DF + CHI 等三种方法在本实验中效果较差。

第二,FDBC 在特征选择中的主要作用即在于通过合并和强化主要特征来解决数据稀疏问题。对于本实验,以保留 10000 个特征为例,特征合并前共有 24000 余候选特征,而合并后仅剩下 10000 个特征。也就是说,有 14000 余候选特征被合并成了 36 个类别的类别码所表征的特征。这显然在很大程度上解决了数据稀疏问题。对数据稀疏问题的有效解决是 FDBC 效果出众的一个主要原因。

第三,FDBC 的特征合并过程同时也是降维的过程,但这种降维不是简单地移除一部分候选特征,因此在这个过程中并没有信息的丢失。仍以保留 10000 个特征为例,合并后剩下的 10000 个特征实际上保持了合并前 24000 余个候选特征所含有的信息量。也就是说,在特征数相同的情况下,FDBC 所选择出来的特征所含有的分类信息要远多于其余几种方法,分类效果也就自然要好于其余方法。

6 结论

本文针对文本分类中常见的噪音和数据稀疏问题提出了一种新的解决方法——基于类别特征域的特征选择方法。该方法通过为分类体系中的每一类构建一个类别特征域来实现重要特征的合并和强化,从而解决数据稀疏问题。这种方法可以看作是从训练语料中自动统计出每一类的领域知识,并将其用于特征选择。同时,该方法还体现出了一种"有选择地进行特征强化"的想法。

实验表明:将传统的文档频率和互信息的方法与这种基于类别特征域的方法结合起来可以同时解决噪音和数据稀疏的问题,从而显著改善文本分类的效果。

参 考 文 献:

[1] 代六玲,黄河燕,等. 中文文本分类中特征抽取方法的比较研究[J]. 中文信息学报,2003,18(1):26 - 32.

[2] 周茜,赵明生,等. 中文文本分类中的特征选择研究[J]. 中文信息学报,2003,18(3):17 - 23.

[3] 孙丽华,张积东,等. 一种改进的 KNN 方法及其在文本分类中的应用[J]. 应用科技,2002,29(2):25 - 27.

[4] Wang Qiang, Wang Xiaolong, Guan Yi. A Study of Semi - Discrete Matrix Decomposition for LSI in Automated Text Categorization[A]. In: First International Joint Conference on Natural Language Processing[C]. 2003: 302 - 309.

[5] Chen Wenliang, Chang xingzhi et al. Automatic Word Clustering for Text Categorization Using Global Information [A]. In: Asia Information Retrieval Symposium[C]. 2004: 1 - 6.

[6] Board, C. L. C. E.. China Library Categorization (The 4th ed.)[M]. Beijing Library Press, Beijing, 1999.

[7] Vladimir N. Vapnik. The Nature of Statistical Learning Theary[M]. Springer, New York, 1998.

[8] 庞剑锋,卜东波,等. 基于向量空间模型的文本自动分类系统的研究与实现[J]. 计算机应用研究,2001,18(9):23 - 26.

[9] 朱华宇,孙正兴,等. 一个基于向量空间模型的中文文本自动分类系统[J]. 计算机工程,2001,27(2):15 - 17.

[10] 李雪蕾,张冬茉. 一种基于向量空间模型的文本分类方法[J]. 计算机工程,2003,29(17):90 - 92.