基于差异频度的类别空间模型的二值分类①

蒋 华[1,2] 刘同来[1] 张万桢[1] (1.桂林电子科技大学 计算机与控制学院 广西 桂林 541004;
2.国家软件与电子电路公用平台广西分中心 广西 桂林 541004)

摘 要： 针对目前文本分类中对向量空间模型的依赖以及文档频率(DF)特征提取方法在二值分类方面的不足，提出了基于差异频度的类别空间模型的二值分类方法，该方法突破了向量空间模型的限制，采用改进DF 的差异频度方法进行特征提取，实现了二值分类功能。实验结果表明，改进的方法是有效的，其分类结果中精确率、召回率、F1 测试值均有改善，提高了分类的准确率。并且本文的方法在其他领域的二值分类中同样值得借鉴。

关键词： 文本分类；差异频度；类别空间模型；向量空间模型；二值分类

Binary Classification Based on Class Space Model of Difference Frequency

JIANG Hua[1,2], LIU Tong-Lai[1], ZHANG Wan-Zhen[1]
(1.School of Computer and Control, Guilin University of Electronic Technology, Guilin 541004, China;
2.Guangxi Centre, National Software and Integrated Circuit Public Service Platform, Guilin 541004, China)

Abstract: As current text classification depends on vector space model and document frequency lacks binary classification, a method based on class space model of difference frequency is presented in this paper. The method breaks the constraint on vector space model, and selects feature with difference frequency improved on document frequency, thus realizes the function of binary Classification. The experiment shows that improved method is effective. Three evaluation parameters, including Precision, Recall and F1, are improved in classification result, and classification precision is better. In addition, the method is worth learning in binary Classification of other areas.

Keywords: text classification; difference frequency; class space model; vector space model; binary classification

信息过滤、信息内容安全管理、舆情褒贬分析等过程基本上就是二值分类问题，二值分类就是指分类结果只有两种可能的情况[1]。也就是说二值分类问题只有两个可选的类别，而且分类结果必居其一，较多值分类问题求解相对简单。

目前的文本分类系统大多是基于向量空间模型[2]，广泛使用文档频率(DF)、互信息量(MI)、信息增益(IG)、x2 统计(CHI)、词汇强度(TS)、TF*IDF 等技术进行特征提取[3]，结合 Bayesian、KNN、SVM 等算法构造分类器进行文本分类。DF 是最常用的特征项提取方法,该方法计算复杂度低，得到广泛使用.但其基于 DF 值小的特征项对分类结果影响较小的假设在二值分类中往往效果较差，其原因是 DF 值较小的特征项有可能具有对分类影响较大的信息量。文献[2]提出的基于类别空间模型的文本分类系统中，采用类别空间概念来描述词语与类别之间的关系。实验证明，该系统突破了向量空间模型的局限，达到了较高的分类效果。

本文在认真分析文献[2]的方法后，针对目前文本表示模型对向量空间模型的依赖性以及文档频率特征提取方法在二值分类方面的不足，提出了基于差异频

① 基金项目:广西自然科学基金 (0991071)
收稿时间:2009-08-06;收到修改稿时间:2009-09-15

度的类别空间模型的二值分类方法，该方法突破了向量空间模型的限制，采用改进 DF 的差异频度方法进行特征提取，实现二值分类功能.通过实验证明，该方法是有效的(分类结果中 F1 值分别由 88.12%到 91.96%、88.31%到 91.77%的改进),并且提高了分类的准确率。

1 二值分类中的类别空间模型

类别空间模型就是由类别构成的空间模型[4]。在类别空间模型中，每一个类别 C_j 就是该空间的一坐标轴 X_j,每一个词语 w_i 作为该空间中的一个点，文献[2]指出词语 w_i 在类别 C_j 中出现的频率 f_{ij} 作为其在坐标轴 X_j 上的分量.假设要把文档分成 m 类,则 m 类便构成了一个 m 维的空间 $X(X_0,X_1,\cdots,X_m)$,词语 w_i 的映射点坐标即为$(f_{i1},f_{i2},\cdots,f_{im})$。

类别空间模型中以类别为坐标轴，刻画的是考虑词对类别的代表能力。词语 wi 在某类别中出现频率越高，距离相应类坐标轴越近，对所在类的代表力就越强，具有越高的区分类别能力[5]。如图 1 所示，二维类别空间图中，词语 wi 距离坐标轴的距离可由词到坐标轴的余弦计算，如公式(1)所示。

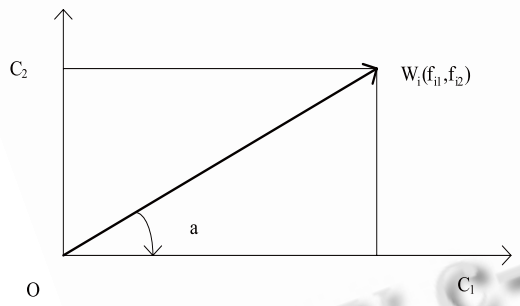

图 1　二维类别空间图

$$\cos(a)=\frac{f_{i1}}{\sqrt{{f_{i1}}^2+{f_{i2}}^2}} \tag{1}$$

二值分类问题中,我们只需构造如图 1 所示的二维类别空间坐标轴，考察词语 w_i 对类别 C_j 的代表能力，即词语在类别中的权重可用代表性系数表示，其计算公式[4]为：

$$w_{ij}=\frac{f_{ij}*\exp(N_{w_{ij}}/N_{w_j})}{\sqrt{{f_{i1}}^2+{f_{i2}}^2}};\ j=1,2 \tag{2}$$

其中：$N_{w_{ij}}$ 是每 j 类中包含特征词 w_i 的文本数，N_{w_j} 是训练语料中出现特征词 w_i 的文本数.为了消除文档长度及每一类训练文档数目对词频的影响，词语 w_i 在第 j 类中的频率 f_{ij} 的计算公式[2]为：

$$f_{ij}=\sum_{k=1}^{N_j}\frac{count(i,k)}{count(k)}\Big/N_j \tag{3}$$

其中：N_j 是第 j 类训练文档总数，count(i, k)是词语 w_i 在第 k 篇文档中出现的次数，count(k)是第 k 篇文档的总词数。

计算训练语料中词语 w_i 在各个类别中的代表性系数 w_{ij} 便可得到一个代表性系数矩阵[2],这个矩阵刻画出每个词语对类别的代表性，即每个词语对类别的倾向性。

向量空间模型是使用最多的文本表示模型,它以特征项(通常情况以词为单位比较合理[6])作为向量空间的轴，文档作为空间中一个点，特征项的权重作为文档的坐标。中文环境下，特征数达到上万,甚至几十万[7]往往会造成“ 维灾难 ”，降低向量空间维数势在必行，但信息损失在所难免。类别空间中以类别为轴，词语为点，通常维数也只有一位或两位数，因此空间维数已不是主要问题。

2 特征提取方法

2.1 文档频率(Document Frequency,DF)

文档频率(DF)是指训练语料中包含词语 w_i 的文档数。这种衡量特征项重要程度的方法基于这样一个假设：DF 值小于某阀值的词是低频词，它们不含或含有较小的类别信息。在训练数据集中，统计所有词语的文档频率，将低于某个阈值的词语从特征集中移除，或选择 DF 值最高的若干个特征作为特征集。

DF 是最简单的特征项提取方法,该方法计算复杂度低,能够胜任大规模的分类任务。Y. Yang 的实验证明：在英文环境中，当 IG 和 CHI 等统计方法的计算“ 费用 ”太高而变得不可用时，DF 可以安全的代替它们被使用[7]。文档频率方法假设低频词是噪声，没有信息量，应该将其从特征集中移除，从而达到向量空间模型中降维的目的。如果被移除的低频词正好是噪声，则可以提高分类的正确率，但并不能保证所有的低频词都是噪声。在二值分类中，往往一些低频词具有很高的信息量,而且在类别空间模型中，空间维数已不会造成问题，因此文档频率方法有可能将一些低频但具有高信息量的词语漏选。这也正是文档频率

方法在二值分类中的不足之处。因此尽量减少“ 贡献 ”词的损失也是我们需要考虑的问题之一。

2.2 改进 DF 的 “ 差异频度 ” 特征提取方法

通常特征提取的目的是降维或去除干扰。本文基于类别空间模型的二值分类系统中，降维已不必考虑,因此 DF 低频 ” 贡献 ” 词的损失得以减少。针对去除干扰，一方面通过设定 ” 完美 ” 停用词表,剔除对分类无意义的词;另一方面,由于文档通常讨论的是同一主题，因此不同的文档使用相同词语较多，这些词语会在两类间分布比较均匀、分布差异不大而影响分类的正确率。本文提出的 ” 差异频度 ” 特征提取方法便能很好的解决这一问题，保留大量出现在某一类中，而在另一类中出现次数比较少的特征，去除了类别代表能力[5]差的词语对分类的干扰。本文提出的特征词 w_i 的差异频度 d_i 的计算公式为：

$$d_i = \exp(|\frac{DF(w_i,C_1)}{N_1} - \frac{DF(w_i,C_2)}{N_2}|*\lambda) \quad (4)$$

其中：N_1 为类别 1 中的文本数，N_2 为类别 2 中的文本数，为倍数调节因子，目的是加大区分度。

在训练数据集中，统计所有词语的 d_i 值，选择高于某阀值的若干个特征作为特征集，将低于某个阈值的词从特征集中移除。同时为避免区分类别能力差的词语的干扰，需要将低于某个阈值的词语从训练语料中剔除。

3 基于差异频度的类别空间模型的分类算法描述

本文二值分类系统主要包括两大模块：训练模块和测试模块。

训练模块的算法主要步骤如下：

1) 文本预处理,主要进行网页净化、中文分词、去除停用词；

2) 根据第三节提出的特征提取方法进行特征提取，得到特征词表，将低于某阀值的词语从训练语料去除；

3) 按本文中的公式(2)对特征词表中的每一个词语计算其对各类别的代表性系数，最后得到一个代表性系数矩阵。

测试模块的算法主要步骤如下：

1) 对待分类文档进行净化、分词、去除停用词等预处理；

2) 根据训练模块中生成的代表性系数矩阵，计算待分类文档相对于每一类别的权值 S_j,计算公式[4]如公式(5)：

$$S_j = \sum_{i=1}^{S} w_{ij}; j = 1,2 \quad (5)$$

其中：j 是类别号，S 是特征词表中的总词数，S_j 是待分类文档相对于 j 类别的权值；

3) 比较 S_j，最大值对应的类别即为待分类文档的类别。

4 实验结果及分析

本文分类系统实验环境如下：

硬件配置 :Pentium(R) 2 CPU 2.4GHz 2.4GHz，1.00G 的内存。

操作系统及应用软件：Microsoft XP，SQL Server2000，VS2005(C#开发语言) 。

本文实验数据集是从网上下载以及手工编撰的共 2000 篇文档。其中健康类文档 C1、不健康类文档 C2 各 1000 篇。将每类文档平均分成五份，每次取其中的四份作为训练库，另一份作为测试库，确保训练库与测试库无重复文档。分三次测试取平均值。

本文二值分类系统性能评估采用精确率(Precision)、召回率(Recall)、F1 测试值 3 种常用指标[8]：

精确率：

$$P = \frac{\text{正确分为某类的文本数}}{\text{测试集中分为该类型的文本总数}} \times 100\% \quad (6)$$

召回率：

$$R = \frac{\text{正确分为某类的文本数}}{\text{测试集中属于该类型的文本总数}} \times 100\% \quad (7)$$

F1 测试值：

$$F_1 = \frac{2 \times P \times R}{P+R} \quad (8)$$

本实验分别采用两种方法进行二值分类测试：

法一是基于类别空间模型和文档频数特征提取方法进行二值分类；系统判断类别文档数目结果如下：被判为健康类别文档数共 215 篇，其中 185 篇为健

康文档，30 篇为不健康文档。被判为不健康类别文档数共 185 篇，其中 15 篇为健康文档，170 篇为不健康文档；

法二是基于类别空间模型和差异频度特征提取方法进行二值分类。系统判断类别文档数目结果如下：被判为健康类别文档数共 211 篇，其中 189 篇为健康文档，22 篇为不健康文档。被判为不健康类别文档数共 189 篇，其中 11 篇为健康文档，178 篇为不健康文档。根据通用指标评估结果如表 1 所示：

表 1　实验结果评估表

类别	指标	法一	法二
C_1	P	86.05%	89.57%
	R	92.50%	94.50%
	F1	88.12%	91.96%
C_2	P	91.89%	94.17%
	R	85.00%	89.50%
	F1	88.31%	91.77%

从上述实验结果可以看出,考察精确率、召回率、F1 测试值 3 种通用指标，F1 值分别由 88.12%到 91.96%、88.31%到 91.77%的改进，表明本文提出的基于类别空间模型和差异频度的二值分类方法是有效的，并且分类效果比基于类别空间模型和文档频数的方法效果要好。

5　结语

目前很多研究工作从分类模型选择、特征降维技术和训练语料构建方法等方面来改善分类器的性能，取得了很好的效果[9]。本文针对目前文本表示模型对向量空间模型的依赖性以及文档频率(DF)特征提取方法在二值分类方面的不足，提出了基于差异频度的类别空间模型的二值分类方法，采用改进 DF 的差异频度方法进行特征提取，实现二值分类功能。通过实验证明，本文改进 DF 的差异频度特征提取方法能很好的去除干扰,分类结果中 F1 值分别由 88.12%到 91.96%、88.31%到 91.77%的改进,明显提高了分类效果。然而，进一步提高分词准确率、结合语义分析、提高分类效果是我们下一步需要研究的问题。

参考文献

1 闫鹏,郑雪峰,李明祥,等.二值文本分类中基于 Bayes 推理的特征选择方法.计算机科学, 2008,35(7):173－176.
2 黄冉,郭嵩山.基于类别空间模型的文本分类系统的设计与实现.计算机应用研究, 2005,22(8):60－63.
3 Yang YM, Pedersen JO. A comparative study on feature selection in text categorization. Nashville, Ten- nessee, USA: Proc. of the 14th International Confer- ence on Machine Learning(ICML 97), 1997. 4l2－420.
4 李艳玲,戴冠中,朱烨行.基于类别空间模型的文本倾向性分类方法.计算机应用, 2007,27(9):2194－2196.
5 徐燕李,锦涛,王斌,孙春明.基于区分类别能力的高性能特征选择方法.软件学报, 2008,19(1):82－89.
6 陈治纲,何丕廉,孙越恒,等.基于向量空间模型的文本分类系统的研究与实现.中文信息学报, 2005,19(1):36－41.
7 代六玲,黄河燕,陈肇雄.中文文本分类中特征抽取方法的比较研究.中文信息学报, 2004,18(1):26－28.
8 樊兴华,孙茂松.一种高性能的两类中文文本分类方法.计算机学报, 2006,29(1):124－131.
9 朱靖波,王会珍,张希娟.面向文本分类的混淆类判别技术.软件学报, 2008,19(3):630－639.