第 35 卷 第 14 期 计 算 机 工 程 2009 年 7 月
Vol.35 No.14 Computer Engineering July 2009
·人工智能及识别技术· 文章编号：1000—3428(2009)14—0227—03 文献标识码：A 中图分类号：TP391

基于正交分解的文本分类模型

熊忠阳，李智星，张玉芳，江 帆
(重庆大学计算机学院，重庆 400030)

摘 要：针对文本分类领域中向量空间模型维数过高和空间扭曲的问题，提出一种基于正交分解的新模型。借用物理学中力的正交分解，将高维的文本向量映射到低维的以类别为坐标轴的空间中，解决了高维的向量和扭曲的空间这 2 个问题。实验表明，与向量空间模型相比，新模型下分类速度有较大提高，精度也有所增加。

关键词：文本分类；正交分解；向量空间模型

Text Classification Model Based on Orthogonal Decomposition

XIONG Zhong-yang, LI Zhi-xing, ZHANG Yu-fang, JIANG Fan
(School of Computer, Chongqing University, Chongqing 400030)

【Abstract】 In text classification area, Vector Space Model(VSM) is the most widely used model while it has two drawbacks: high dimensions and warped space. This paper presents a new model based on orthogonal decomposition. In this model, higher dimensional vectors of texts are mapped in a lower dimensional space which uses categories as its coordinate axes to solve these two drawbacks. Experiment shows that under the new model, the classification process is speeded up to a considerable degree and the precision is increased.

【Key words】 text classification; orthogonal decomposition; Vector Space Model(VSM)

文本是获取和传播信息最快捷和最有效的手段。面对海量的文本信息，自动文本分类大大提高了分类的效率。向量空间模型(Vector Space Model, VSM)[1]是研究较多的基于统计的文本分类模型。但向量空间模型有 2 个固有的缺陷：高维的向量和扭曲的空间[2]。目前针对向量空间模型的研究主要集中在分类算法以及特征提取算法的设计和改进上，而针对这 2 个缺陷对向量空间模型本身进行的改进很少。本文针对向量空间模型的这 2 种缺陷，提出了一种新型的基于正交分解的模型。

1 向量空间模型

VSM 由 Salton[1]于 20 世纪 70 年代中期提出。向量空间模型的基本原理是：对文本进行预处理，抽取代表其特征的元数据，将这些特征用结构化的形式保存作为文档的中间表示形式。在文本分类中，向量空间的维度为所有文档包含的所有特征项，每一个特征项代表向量空间中的一维。而文本被表示成这个空间中的一个向量。文本向量某一维的值由该维代表的特征项在这个文本中的权重确定。在向量空间模型中，特征项之间的关系并没有被考虑，也就是说所有的维都被认为是正交的[3]，这就不可避免地会产生以下 2 个问题：

(1)高维的向量。在向量空间模型中，空间的维度是由特征项的数目决定的。在实际的文本分类过程中，特征项的数目往往十分庞大。

(2)扭曲的空间。在向量空间模型中，特征项被看成是相互独立的。也就是说，某个特征项的出现的几率并不影响其他特征项出现的几率，也不受其他特征项是否出现所影响。向量空间模型的所有维都被认为是正交的，而这显然不符合实际情况。例如：同一篇文档中，如果出现了“战争”这个词，那么这篇文档中出现“武器”的概率显然比出现“乐器”的概率要大。实际上，如果不正交的维在向量空间模型中被当成正交的处理，那么空间就会产生扭曲。可以通过一个简单的例子来分析。

图 1~图 3 分别表示直线 $x+y=1$ 在直角坐标系、非直角坐标系、扭曲的坐标系中的图形。

图 1 表示直线 $x+y=1$ 在直角坐标系中的图形(仅限第一象限)。在这个坐标系中，X 轴和 Y 轴正交，从图形上看就是 X 轴和 Y 轴垂直。

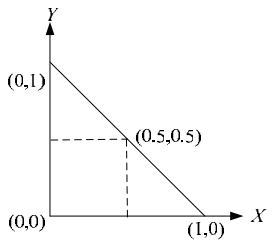

图 1 直角坐标系中 $x+y=1$ 的图形(第一象限)

图 2 表示直线 $x+y=1$ 在非直角坐标系中的图形(仅限第一象限)。在这个坐标系中，X 轴和 Y 轴夹角余弦为 0.6，也就是说在 X 轴上，X 值每增加 1，Y 值增加 0.6；在 Y 轴上，Y 值每增加 1，X 值增加 0.6。可以很容易算出直线在此坐标系中与 X，Y 坐标轴的交点分别为：(0.625,0.375)，(0.375,0.625)。而且还可以算出点(0.5,0.5)也是 $x+y=1$ 上的一点。

基金项目：教育部留学回国人员启动基金资助项目(教外司留[2007]1108-10)

作者简介：熊忠阳(1964－)，男，教授、博士生导师，主研方向：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；李智星，硕士研究生；张玉芳，副教授；江 帆，硕士研究生

收稿日期：2009-01-12 **E-mail：**adam0730@126.com

在此坐标系中 $x+y=1$ 仍然是一条直线。

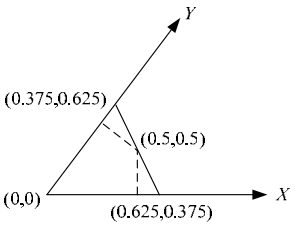

图 2　非直角坐标系中 $x+y=1$ 的图形(第一象限)

图 3 表示图 2 所示坐标系在强行转换为直角坐标系之后 $x+y=1$ 的图形(仅限第一象限)。在这个坐标系中，X 轴和 Y 轴原本不正交，但被当成正交处理。原本是一条直线的 $x+y=1$ 的图形在这个坐标系中被扭曲成了一根曲线。很明显，空间发生了扭曲。在这个空间中，点与点之间的距离计算必然出现误差。

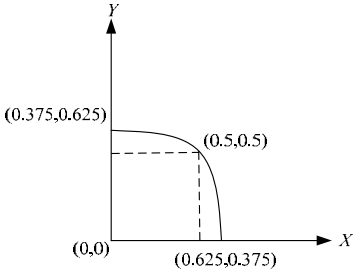

图 3　扭曲的坐标系中 $x+y=1$ 的图形(第一象限)

2　正交分解模型

2.1　力的正交分解

在物理学中，经常要分析物体的受力情况，而一个物体受到的力一般都是大小不一、方向各不相同的。要考虑一个物体的总体受力情况，一般采用力的正交分解的方法。力的正交分解步骤如下：

(1)以物体质心为原点建立直角坐标系(二维或三维)。

(2)将各个分力以坐标轴为标准进行分解，分力在坐标轴上的投影即为分力在该坐标轴的分量。

(3)统计物体在各个坐标轴的分量之和，即物体的整体受力情况。

这种分力分解到坐标轴，再从坐标轴统一到合力的模型即为正交分解模型，如图 4 所示。

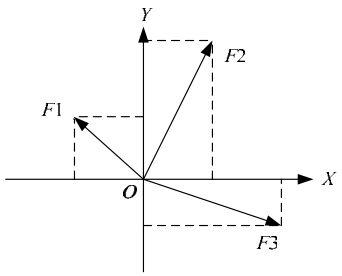

图 4　力的正交分解

这种模型避免了分力间夹角不尽相同的问题。它并没有直接计算分力之间的关系，而是将其转化到易于计算并具有实际意义的坐标上。因为在实际问题中，通常需要的是物体垂直和水平方向的受力情况，而分力之间的关系通过它们在不同坐标上的分量得到了体现。

2.2　正交分解模型到文本分类的应用

由上面的分析可以看出，正交分解实际上是一个从部分决定整体的一个计算过程。这与文本分类的过程非常相似。文本分类是通过一个个的特征项来决定文本的类别。在中文语义里，词是表达语义最基本的单位，因此，在“词袋[4]”模式中，文档被看成是许多词的集合，如同力正交分解，合力被看成是众多分力的集合。合力和分力都有方向，图形上看是倾向于某个坐标轴，同样文档和词都有意义上的倾向，倾向于某个类。在力的正交分解中，需要的是物体在垂直和水平方向的受力情况，同样在文本分类中，需要的只是文档在各个类别上的分量的大小。在将正交分解模型应用到文本分类时，不难得出表 1 所示的对应关系。

表 1　力的正交分解与文本分类的对应关系

力的正交分解	文本分类
坐标轴	类别
分力	特征项(词)
合力	文档

按照这个对应关系，基于正交分解的文本分类模型的分类步骤如下：

(1)以类别为坐标轴建立坐标系。

(2)求出从训练集获得的所有特征项与各个类别之间的关系。每个词用一个 m 维向量表示，其中，m 是类别的数目。定义 $T=\{t_1,t_2,\cdots,t_n\}$ 是特征项的集合，$C=\{c_1,c_2,\cdots,c_m\}$ 是类别的集合，$TCF(t_i,c_j)$ 表示特征项 t_i 在类 c_j 中出现的频次，$DF(t_i)$ 表示特征项 t_i 出现的文档数，$CF(c_j)$ 表示类 c_j 所包含的文档数。那么可以用以下公式求出特征项 t_i 在类 c_j 上的分量 $W(t_i,c_j)$。

$$W(t_i,c_j)=TCF(t_i,c_j)/(DF(t_i)\times CF(c_j)) \tag{1}$$

其中，$CF(c_j)$ 起到了防止由于类文档数目不平均造成的分类偏差问题的作用。

(3)对测试集中的每一篇文档，将其包含的所有特征项的向量相加，得出该文档的 m 维向量表示。每一维的数值表示该文档与对应类的相关度。相关度最高的那一维对应的类就是该文档被分到的类。定义 $D=\{d_1,d_2,\cdots,d_l\}$ 为文档的集合，$TF(t_i,d_k)$ 为特征项 t_i 在文档 d_k 中出现的频次，$WD(d_k,c_j)$ 为文档 d_k 与类 c_j 的相关度，可得 $WD(d_k,c_j)$ 计算公式如下：

$$WD(d_k,c_j)=\sum_{i=1}^{n}W(t_i,c_j)TF(t_i,d_k) \tag{2}$$

3　实验结果及分析

3.1　实验目的及实验设计

本实验的目的是检验正交分解模型的有效性和分类速度，并通过对实验结果的分析找出正交分解模型的缺点，以期在下一步的工作中改进。

本文采用复旦大学李荣陆博士的中文分类语料的一个小型语料库。共 2 816 篇文档，其中，训练集 1 882 篇，测试集 934 篇。共分电脑、艺术、教育、交通、环境、经济、医疗、军事、政治、体育 10 个类。采用中科院 ictclas1 1.0 分词程序进行文档分词处理以及开源的 lucene 进行倒排索引的建立，在 eclipse3.2 上进行的程序设计。同时将结果与向量空间模型下 K 近邻算法(以下简称 KNN)[5]以及支持向量机算法(以下简称 SVM)[5]比较。特征提取采用信息增益值。采用宏平均查准率和微平均查准率以及分类时间作为评价指标[6]对 2 种模型分类性能进行比较。

3.2 实验结果及分析

表 2 是正交分解模型下不采取任何特征提取算法分类的查准率和用时数据。表 3 和表 4 分别表示采用信息增益值作为特征提取方法时，正交分解模型与向量空间模型下 KNN 算法与 SVM 算法查准率和消耗时间的比较。

表 2 不采用特征提取算法的分类结果

宏平均查准率	微平均查准率	特征数	训练用时/ms	分类用时/ms
0.957 88	0.961 38	38 858	4 906	2 047

表 3 查准率比较

特征数	正交分解模型		KNN		SVM	
	宏平均查准率	微平均查准率	宏平均查准率	微平均查准率	宏平均查准率	微平均查准率
2 000	0.912 53	0.922 61	0.906 20	0.865 10	0.965 16	0.959 31
4 000	0.946 00	0.950 64	0.882 99	0.821 20	0.963 07	0.953 96
6 000	0.952 48	0.956 63	0.878 88	0.799 79	0.962 53	0.969 57
10 000	0.955 72	0.958 66	0.880 79	0.762 31	0.955 03	0.963 77
20 000	0.955 72	0.958 55	0.874 02	0.721 63	0.953 96	0.963 32
30 000	0.957 88	0.961 38	0.856 94	0.721 63	0.956 10	0.966 11

表 4 消耗时间比较 s

特征数	正交分解模型		KNN		SVM	
	训练用时	分类用时	训练用时	分类用时	训练用时	分类用时
2 000	5	0.3	12	9	40	26
4 000	5	0.5	13	10	63	38
6 000	5	0.6	15	12	71	45
10 000	5	0.8	23	13	84	54
20 000	5	1.3	37	18	103	64
30 000	5	1.7	38	22	119	68

基于以上数据，做出以下分析：

(1)本文提出的分类模型效果明显好于向量空间模型下 KNN 算法的效果。在低维特征下表现不如 SVM 算法，但在高维特征下表现与 SVM 算法相当。

(2)正交分解模型下分类算法复杂度明显小于向量空间模型下的 KNN 算法与 SVM 算法。这是因为通过正交分解，文档向量维度从 n 维降低到了 m 维，以非常小的代价获得了文档在各个类别上的分量。可以通过分析式(2)，得到正交分解模型下分类算法复杂度为 $O(mnl)$，其中，m 为类别数；n 为特征数；l 为待分类文档数。

(3)利用信息增益值对正交分解模型进行特征提取效果并不明显。当特征为 30 000 维时，分类效果达到最佳，同时这个效果与不进行特征提取的效果相同。这说明利用信息增益值作为正交分解模型的特征提取值并不会显著影响分类效果，但可以减少分类算法的时间复杂度，同时说明正交分解模型对于干扰词有很大的耐受程度。

(4)正交分解模型只是部分地解决了文本向量空间扭曲的问题，因为作为坐标轴的类别本身也不是正交的，但与向量空间模型相比，扭曲程度已经大为降低。

4 其他分类模型

文献[7]提出基于潜在语义的文本分类模型(MPLS)，思路是从原始文本空间中得出一个潜在语义空间再进行偏最小二乘分析。与 SVM 相比，该模型优点在于可以进行多类分类。在本文提出的模型中，文本被表示成一个以类别为坐标的向量，每一维代表该文本与对应类的相关度，因此，也可很好地处理多类和兼类的问题。MPLS 的时间复杂度为 $O(mnsl)$，其中，m 为文档数；n 为特征数；s 为决定潜在语义变量对数目的循环次数；l 计算潜在语义变量对的循环次数。显然，相对 MPLS，本文提出的模型具有时间复杂度上的优势。

5 结束语

本文提出了基于正交分解的文本分类模型，通过实验证实了模型的有效性，并且对正交分解模型的优缺点进行了分析。实验结果表明，正交分解模型的分类效果与向量空间模型下 SVM 算法分类效果相当，明显好于 KNN 算法，复杂度较 KNN 算法与 SVM 算法都要小很多。但实验中发现，基于信息增益值的特征提取算法对于正交分解模型效果不明显，有必要找到一种适合正交分解模型的特征提取算法。同时正交分解模型并未完全解决空间扭曲的问题，需要在以后的工作中进一步的研究。

参考文献

[1] 焦玉英, 宋晓晴. 基于 VSM 的文档信息检索改进[J]. 情报理论与实践, 2007, 30(1): 97-104.

[2] 王 煜. 基于决策数和 K 最近邻算法的文本分类研究[D]. 天津: 天津大学, 2006.

[3] 廖 玲, 文敦伟. 基于改进向量空间模型的邮件分类[J]. 计算机数字与工程, 2007, 35(4): 190-193.

[4] 彭时名. 中文文本分类中特征提取算法研究[D]. 重庆: 重庆大学, 2005.

[5] 李荣陆, 王建会, 陈晓云, 等. 使用最大熵模型进行中文文本分类[J]. 计算机研究与发展, 2005, 42(1): 94-101.

[6] 程泽凯, 林士敏. 文本分类器准确性评估方法[J]. 情报学报, 2004, 23(5): 631-636.

[7] 叶 浩, 王明文, 曾雪强. 基于潜在语义的多类文本分类模型研究[J]. 清华大学学报: 自然科学版, 2005, 45(S1): 1818-1822.

编辑 索书志

(上接第 205 页)

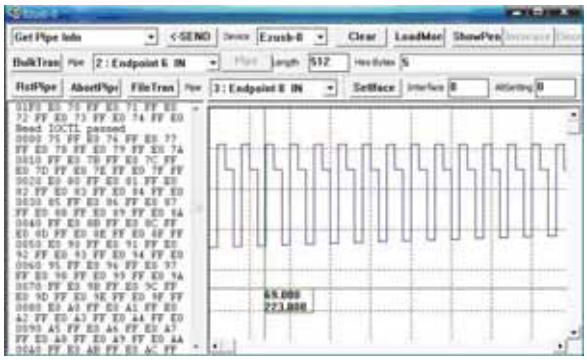

图 6 USB 采集到的 CCD 模拟信号

参考文献

[1] Cypress Semiconductor Corporation. EZ-USB FX2 Technical Reference Manual(Version2.1)[Z]. (2006-05-25). http://www.cypress.com/.

[2] 钱 峰. EZ-USB FX2 单片机原理、编程及应用[M]. 北京: 北京航空航天大学出版社, 2006.

[3] 侯俊杰. 深入浅出 MFC[M]. 2 版. 武汉: 华中科技大学出版社, 2002.

[4] 柴东岩, 侯紫峰. 引导程序中 USB 下载功能的设计与实现[J]. 计算机工程, 2008, 34(6): 268-269.

编辑 任吉慧