文章编号:1001 - 9081(2005)08 - 1821 - 03

基于差异—相似矩阵的文本降维方法

黄晓春,晏蒲柳,夏德麟,陈 健
(武汉大学 电子信息学院,湖北 武汉 430079)
(xiaochun. huang@ gmail. com)

摘 要:由于文本文档数量多、词量大,形成的文档空间维度高,很多自动文本分类算法不能直接有效地发挥作用。基于差异—相似矩阵(DSM)的方法在很大程度上降低了文档空间的维度。已经分好类的文集经过预处理后被表示成特征项—文档矩阵,再转化为差异—相似矩阵,其中同类文档采用相似项描述,而异类文档则采用差异项描述。通过对差异—相似矩阵的处理,最终得到维度较低的文本特征集,并同时生成分类规则。实验说明,对于大规模文集,DSM 方法能在保持良好的分类质量的同时,获得较高的属性降维率和样本降维率。

关键词:文本分类;维度消减;差异—相似矩阵
中图分类号: TP391.3 **文献标识码**:A

Dimensionality reduction for text document using difference-similitude matrix

HUANG Xiao-chun, YAN Pu-liu, XIA De-lin, CHEN Jian
(*School of Electronic Information, Wuhan University, Wuhan Hubei* 430079, *China*)

Abstract: Due to the huge amount of text documents and their vocabulary, document spaces are commonly of high dimensionality, and many automatical text categorization algorithms can not get their best performences directly. Difference-similitude Matrix-based (DSM) method reduces dimensionality to a great extend. Pre-classified collection is represented as a item-document matrix after preprocessing, then transmitted into a DSM, in which documents in the same classes are depicted with similitude while documents in different classes with difference. The method generates an item set of low dimensionality and a set of classification rules after dealing with the DSM. Results of experiments suggest that DSM-based method could achieve high attribute reduction degree and sample reduction degree with good classification quality.

Key words: text categorization; dimensionality reduction; DSM(Difference-Similitude Matrix)

0 引言

自动文本分类有助于提高信息检索的速度、挖掘潜在于字里行间的深层知识,帮助人们快速、正确地从文本中获得所需的知识。向量空间模型(Vector Space Model)是一种常见的文本文档表示方式,然而对于类似神经网络分类器和 K 最近邻分类器的绝大多数传统分类算法而言,由这种方式直接形成的文档空间维度太高,不便于直接处理[1]。此外,文档集合通常包含了成千上万个词条,实际上却只有部分真正有利于分类。维度消减技术利用这个特点,很好地解决了文档空间维度过高的问题。

从文本分类的角度而言,维度消减就是选择尽量少的关键词来准确描述文档内容和分类规则,同时要求生成的规则数尽可能少,即实现属性消减和样本消减。文档降维技术大致分为两类:特征提取和特征重构[2]。常见的特征提取方法有 TFIDF 系列方法、信息熵、信息增益、主成分分析、基于粗集理论的约简方法等。前三种方法因为要过滤掉部分不满足阈值要求的特征项,因而会导致一定程度的信息损失;此外,这些方法没有利用特征项和特征项、特征项和文档以及特征项和类别之间的关系。主成分分析考虑了前两种关系,却没有体现特征项和类别之间的关系,所以也会错误地放弃一些不是主成分却具有很强区别能力的特征项[3]。粗集理论充分利用了特征项、文档和类别之间的关系,但只利用上近似和下近似描述对象间的差异性和不可辨别关系[4],而忽略了对象间的相似性。为了充分利用数据集提供的知识和降低计算复杂度,我们采用了差异—相似矩阵(Difference-Similitude Matrix,DSM)。基于 DSM 的方法同时利用了对象之间的差异性和相似性,因而能够更有效地降低文本特征空间的维度,同时生成数量少而准确率高的分类规则。研究表明,这种理论比粗集理论方法更简单、快速,描述能力更强[5]。

在本文中我们采用基于 DSM 的降维方法,对描述已分类文集的特征项—文档矩阵进行维度消减,并生成用于文本分类的规则集合。

1 差异—相似矩阵理论

为了更好地理解 DSM 在文本分类器中的应用,首先简单地介绍 DSM 的基本知识。假设 IS 是一个信息系统,并且有 $IS = < U, C, D, V, f >$,其中 U 表示系统的对象集合;C 表示条件属性集;D 表示决策属性集;$V = \cup (V_a : a \in (C \cup D))$ 表示属性值的集合;$f: U \times (C \cup D) \rightarrow V$ 是决定属性值的函数。对于信息系统内的对象,可以定义一个 $m \times m$ 维的差异—相似矩阵 M_{DS} 来描述它们的属性和值。这个矩阵由两种类型的元素构成:相似项 m_{ij}^s 和差异项 m_{ij}^d,根据公式(1)可以确定元素的值。其中,m 是条件属性的个数,n 是数据集中样本的个

收稿日期:2005 - 01 - 14;修订日期:2005 - 03 - 22 基金项目:国家自然科学基金资助项目(90204008)

作者简介:黄晓春(1976 -),女,湖北人,博士研究生,主要研究方向:数据挖掘、自然语言理解; 晏蒲柳(1962 -),女,湖北人,教授,博士生导师,主要研究方向:人工智能、网络管理; 夏德麟(1938 -),男,湖北人,教授,主要研究方向:人工智能、网络管理; 陈健(1972 -),男,广西人,讲师,博士,主要研究方向:传感器网络、数据挖掘.

数。我们对 m_{ij}^{s} 的定义稍微作了修改，即用 m_{ii}^{s} 表示对角线上的元素，而不是如文献[5]那样用 0 来表示，这样有利于简化算

$$m_{ij}=\begin{cases} m_{ij}^{s}=\begin{cases}\{q\in C:f(q,x_i)=f(q,x_j)\},D(x_i)=D(x_j)\\ \{\phi:\forall(f(q,x_i)\neq f(q,x_j))\},D(x_i)=D(x_j)\end{cases}\\ m_{ij}^{d}=\begin{cases}\{q\in C:f(q,x_i)\neq f(q,x_j)\},D(x_i)\neq D(x_j)\\ \{\phi:\forall(f(q,x_i)=f(q,x_j))\},D(x_i)\neq D(x_j)\end{cases}\end{cases}$$

$$i=1,2,\cdots,m;\ j=1,2,\cdots,m;\ D(x)=1,2,\cdots,n \quad (1)$$

后文中用到的属性重要度 $sig_{a\in c}(D)$、C_i^b（核条件属性集）和 B_i^{op}（最佳条件属性集）等概念的定义，以及求解补充属性集的具体方法，详细描述请参见文献[6]。基于 DSM 的维度消减遵循的原则是：在不损失原有系统信息的前提下，使得描述规则所需的属性个数最少，并且使得生成的分类规则个数最少。相较于文本分类维度消减的最终目标，这条原则也十分适用。

2 基于 DSM 的文本特征降维和分类

文本分类的目的是将文本文档划分到预定义好的类别中去。为此，常用的方法是先从已划分到不同类别的文档中抽取出关键词，计算这些关键词在待分类文档中的词频，并将结果与已分类文档的词频比较，根据比较结果将文档分到最近的类别中。由关键词和文档组成的空间是特征项 — 文档空间，简称特征空间。令 $C=\{c_1,\cdots,c_j,\cdots,c_m\}$ 表示某个文集的预定义类别集合，$D=\{d_1,\cdots,d_i,\cdots,d_n\}$ 表示划分到这些类别中的文档集合，$T=\{t_1,\cdots,t_k,\cdots,t_p\}$ 表示特征项集合，$F=\{f_1,\cdots f_k,\cdots f_p\}$ 表示特征项 t_k 的词频。其中 m 是类别数，n 是文集中已分类文档的数目，p 是文集中不同特征项的总数。则可以用 $n\times(p+1)$ 维矩阵 A 表示该文集的特征空间，A 中的元素 a_{ij} 表示特征项 t_j 在文档 d_i 中出现的词频。

文集通常包括数以万计的文档和词条，所以特征空间的维度非常高。为了提高自动分类的效率，在运用各种分类器之前，通常需要使用降维技术消减特征空间的维度。我们提出了一种基于 DSM 的文本分类器，其中差异—相似矩阵发挥了两个重要作用：维度消减和规则生成。由于文本文档的内容一般都是无结构数据，所以在分类之前应该将文档进行预处理，使其满足文档表述的需要。

2.1 预处理

对于一个已分类的训练文集，生成分类器之前，常见的预处理工作包括清除网页等各种格式的平面文档中的标签、分词、去除停用词和计算余下词条的词频。本文采用了一种常见的词频统计方法 TFIDF[7] 来计算词频，但不删除任何高频或低频特征项：

$$f_k=f_k'\cdot\log\left(\frac{N}{N_k}\right)\Big/\sqrt{\sum_{i=1}^{M}\left[\log\left(\frac{N}{N_i}\right)\right]^2} \quad (2)$$

其中 f_k' 表示文档 d_i 中出现关键词 t_k 的次数；N 是文集中包含 t_k 的文档总数；N_k 是所有文档中出现 t_k 的总次数；M 是文集中至少出现一次的词语总数。这样，文集就可以表示为一个特征空间矩阵，矩阵中的样本 i 就是文档 d_i，形式如下：

$$d_i=[F\quad\overline{C}]=[f_1\quad\cdots\quad f_k\quad\cdots\quad f_p\quad\overline{C}] \quad (3)$$

这里 $\overline{C}$ 表示 d_i 所属的类别。

由于 DSM 算法本身只能处理离散数据，必须在将特征空间矩阵转换为差异—相似矩阵之前，把连续的词频进行离散化处理。离散化后，原本大量的不同词频值可以被映射到小规模的区间中，也能间接地降低分类规则的数目。

2.2 DSM 维度消减和规则生成

经过上述预处理后，文集被转换成一个能被 DSM 算法处理的特征项—文档矩阵。基于 DSM 的降维方法选择最能代表所属类别的特征项，从而实现在不损失文集信息的前提下降低维度，与此同时，我们还能够得到数目最少的分类规则。下面是 DSM 降维方法的算法步骤：

DSM_reduction：

输入：特征项 — 文档矩阵 $A(m\times(p+1)$ 维，这里 m 表示文集中的文档总数，p 表示特征项总数）

输出：维度消减后的条件属性（特征项）集合 RA，分类规则集合 CR

步骤：

(1) 将 A 中的文档按类别排序；

(2) 根据 A 构造差异 — 相似矩阵 M_{DS}；令 $ON(k)$ 表示对应与类别 k 的子矩阵，$k=1,j=1,s=\phi$；

(3) for $k=1\sim m$：

记录类别 k 的开始位置的下标 $js(k)$，以及结束位置的下标 $je(k)$；

for $i=js(k)+1\sim je(k)$：

a) if $i\in s$

then $r(i)=r(i-1);i=i+1$；

b) 求 C_i^b 和 $sig_{a\in C}(\overline{C})$（特征项的重要度），令 $a=\phi$；

c) If $C_i^b\cap(\forall m_{ij}^d)\neq\varnothing$

then 记录 s 中满足 $C_i^b\cap m_{ij}^s\neq\varnothing$ 的文档的下标 j；

else $seek\_C_i^a$；

d) $C_i^b\cup C_i^a\rightarrow B_i^{op};RA=RA\cup B_i^{op}$；

e) $r_k(i):B_i^{op}(v_i)\rightarrow\overline{C}_k;CR=CR\cup\{r_k(i)\}$；

(4) 去除 CR 中冗余的规则。

$seek\_C_i^a$ 是计算补充条件属性集的函数（详见文献[5]）。

最后生成的分类规则形式如下：

$$r_{ik}:=(h_1=a_1)\wedge(h_2=a_2)\wedge\cdots\wedge(h_n=a_n)\Rightarrow d_j\rightarrow C_i \quad (4)$$

其中 r_{ik} 是第 k 个判断 d_j 是否属于类别 c_i 的规则，h_n 是经过 DSM 降维后保留下来的关键词 t_n 的词频，a_n 是 h_n 的值。

3 实验评估

Reuters-21578 文本分类集合[8] 内包含的新闻稿件都经过人工分类和标注，我们采用了其中的三个子类别作为 DSM 分类器的训练集和测试集。这三个子类别分别是 acq、Money-fx和 grain，对于每一类随机抽取 3/4 的文档作为训练，随机抽取 1/4 作为测试集。为了评估 DSM 方法的性能，我们选择了精确率和召回率作为分类性能指标，属性降维度 R_a 和样本降维度 R_s 作为维度消减性能指标[9]。其中：

$$R_a=\frac{\text{原有属性数}-\text{最佳属性集中的属性}}{\text{原有属性数}}$$

$$R_s=\frac{\text{原有样本数}-\text{生成的规则数}}{\text{原有样本数}}$$

图 1 是三个子类别分类精确率和召回率的实验结果。可以看出，这两个指标是相互制约的，当精确率很高时，召回率就低，反之亦然。通常，如果文集中同类文档内容比较一致或者明确，则精确率和召回率接近时的值就高，反之则低。

类别 grain 包括了 Reuters-21578 中所有出现了 grain 的文档，我们计算了这个类别的属性降维度 R_a 和样本降维度 R_s。表 1 中的结果表明，使用差异—相似矩阵方法可以显著降低样本和规则的维度，同时还能保证良好的分类质量。此外，从

结果中可以看到,选取的特征项太多也会对属性降维和规则降维产生不好的影响,我们认为这是由于某些特征项引入了一些干扰信息。

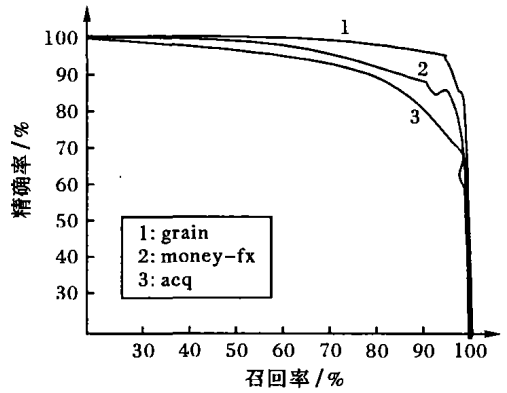

图 1　类别 acq,Money-fx,grain 的分类精确率与召回率

表 1　类别 grain 在不同特征项维数下的属性降维度和样本降维度

	100	500	1000	2000
R_a	0.785	0.681	0.632	0.574
R_s	0.55	0.582	0.617	0.630

4　结语

本文提出的方法总体而言进行了三次维度消减:词频离散化降维、条件属性(特征项)降维和分类规则降维。实验结果说明 DSM 方法可以应用于文本分类,并能获得相当好的结果。它能够同时消减文集中文档特征属性的维度以及分类规则所需的样本维度,从而提高分类的效率。然而,由于我们只针对 Reuter-21578 的三个子类别进行了实验,有关维度的选择阈值如何确定、相似度和重要度如何折中等很多问题尚需要深入探讨。另外,我们还将在以后的工作中引入基于DSM 的增量式学习方法、动态修订和产生规则,进一步改善分类器的性能。

参考文献:

[1] YANG Y, PEDERSEN JP. A Comparative Study on Feature Selection in Text Categorization[A]. Proceedings of the Fourteenth International Conference on Machine Learning[C]. Tennessee, USA: Vanderbilt University, 1997.

[2] 孙建军, 成颖, 丁芹, 等. 信息检索技术[M]. 北京: 科学出版社, 2004. 168 - 170.

[3] DUIN RPW, LOOG M, HAEB - UMBACH R. Multi - class Linear Feature Extraction by Nonlinear PCA[A]. Proceedings of 15th International Conference on Pattern Recognition[C]. Barcelona, Spain: IEEE Computer Science Press, 2000. 398 - 401.

[4] NGUYEN, SON H. Scalable classification method based on rough sets[A]. Rough Sets and Current Trends in Computing, Lecture Notes in Computer Science[C]. PA, USA: Springer, 2002. 433 - 440.

[5] 夏德麟, 晏蒲柳. 一种新的信息系统知识约简方法——DSM 方法[D]. 武汉: 武汉大学大学电子信息学院, 2001.

[6] ZHOU JG, XIA DL, YAN PL. Incremental Machine Learning Theorem and Algorithm Based on DSM Method[A]. Proceedings of the Third International Conference on Machine Learning and Cybernetics [C]. Shanghai: IEEE, 2004. 2202 - 2207.

[7] AIZAWA A. The Feature Quantity: An Information Theoretic Perspective of Tfidf-like Measures[A]. Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval [C]. Tarrytown, NY, USA: Pergamon Press, Inc, 2000.104 - 111.

[8] Reuters - 21578 Text Categorization Collection [DB/OL]. http://kdd. ics. uci. edu/databases/reuters21578/reuters21578. html, 2004.

[9] 江昊, 晏蒲柳. 基于 DSM 的数据约简[J]. 武汉大学学报(理学版), 2003, 49(3): 378 - 382.

(上接第 1820 页)

DCellO 算法不能根据密度识别出孤立点及通过孤立点搭桥将若干个聚类聚为一类的不足,提出了改进算法 DCellO1:沿用了 DCellO 算法邻接网格向四周弥散及使用 Bresenham 及最短路径算法计算带障碍距离的思想,引入核心网格、聚类半径、下限域值及密度相连的概念,真正考虑了密度对聚类的影响,能够处理任意形状的对象、任意形状的障碍,可方便处理光栅图像,易于并行化,在对象分布不均匀时聚类效果尤其好,其时间复杂度与空间复杂度在数量级上与 DCellO 算法相同。

参考文献:

[1] MACQUEEN J. Some methods for classification and analysis of multivariate observations[A]. 5th Berkeley symposium on mathematics, statistics and probability[C], 1967,1. 281 - 296.

[2] KAUFMAN L, ROUSSEEUW P. Finding groups in data: an introduction to cluster analysis[M]. New York: John Wiley & Sons, 1990.

[3] KARYPIS G, HAN E - H, KUMAR V. Chameleon: a hierarchical clustering algorithm using dynamic modeling[J]. Computer, 1999, 32(1):32 - 68.

[4] ESTER M, KRIEGEL H-P, SANDER J, *et al*. A density-based algorithm for discovering clusters in large spatial databases with noise [A]. Second International Conference on Knowledge Discovery and Data Mining [C]. Portland: AAAI Press, 1996. 226 - 231.

[5] TUNG AKH, HOU J, HAN J. Spatial Clustering in the Presence of Obstacles[A]. Proceedings of 2001 International Conference on Data Engineering[C], 2001.

[6] 陈克平, 周丽华, 王丽珍, 等. DCellO——网格弥散聚类算法[J]. 计算机研究与发展, 2004,41(增刊):205 - 212.

[7] HAN J, KAMBER M. 数据挖掘——概念与技术(影印版)[M]. 范明, 孟小峰, 译. 北京: 机械工业出版社, 2001.

[8] NG R, HAN J. Efficient and effective clustering methods for spatial data mining[A]. Twentieth International Conference on Very Large Databases[C]. Morgan Kaufmann, 1994.144 - 155.

[9] NG R, HAN J. Clarans: A method for clustering objects for spatial data mining[J]. IEEE Transactions on Knowledge and Data Engineering, 2002,14(5): 1003 - 1016.

[10] ZHANG T, RAMAKRISHNAN R, LIVNY M. Birch: An efficient clustering method for very large databases[A]. ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery[C]. Montreal, 1996.103 - 114.

[11] GUHA S, RASTOGI R, SHIM K. Cure: an efficient clustering algorithm for large databases[A]. ACM SIGMOD International Conference on the Management of Data[C]. Seattle, WA, USA, 1998.73 - 84.

[12] KAUFMAN L, ROUSSEEUW P. Finding Groups in Data: an Introduction to Cluster Analysis[M]. John Wiley& Sons, 1990.

[13] WANG W, YANG J, MUNTZ R. STING: A statistical information grid approach to spatial data mining[A]. Proceedings of International Conference on Very Large Data Bases (VLDB'97)[C]. Athens, Greece, 1997. 186 - 195.

[14] BRESENHAM JE. Algorithm for computer control of a digital plotter[J]. IBM Systems Journal, 1965,4(1): 25 - 30.