浙江理工大学学报,第 28 卷,第 4 期,2011 年 7 月
Journal of Zhejiang Sci-Tech University
Vol. 28, No. 4, Jul. 2011

文章编号: 1673-3851 (2011) 04-0544-05

基于粗糙集特征加权的文本分类

徐　欣, 黄理灿, 赵玉虹

(浙江理工大学信息电子学院, 杭州 310018)

摘　要: 文本分类是当今信息检索和数据挖掘等领域的研究热点,而特征加权是文本分类过程中的重要步骤。为了提高分类质量,文章通过深入分析粗糙集理论和逆文本频率加权的思想,提出了一种基于粗糙集的特征加权方法,从近似分类精度和近似分类质量两个方面考虑特征词对分类的全局作用,将文本的类别属性信息引入到权重中。通过文本分类实验证明,该加权方法有助于提高分类系统的分类效果。

关键词: 粗糙集理论; 特征加权; 文本分类; 近似分类精度; 近似分类质量

中图分类号: TP301　　**文献标识码**: A

0　引　言

近年来,随着网络与信息技术的发展信息量急剧扩增,人们在享受日益丰富的信息的同时也被其所淹没。面对浩如烟海的信息,想要搜索自己感兴趣的信息或者管理这些信息都变得越来越困难,这是一个迫切需要解决的问题。自动文本分类技术就可以很好地帮助人们解决这些问题。

文本分类是当今信息检索和数据挖掘等领域的研究热点,其主要任务是在预先给定类别标记集合下,根据文本的内容来判定其所属类别[1]。由于文本信息的复杂性,不可能直接对其进行分类,在进行分类之前必须进行预处理。为了更准确地描述特征词在文本中的重要性,要为文本中的特征词加权。这主要通过分析特征词对文本分类的贡献程度得到的,采用合理的加权方式可以增大特征词之间的差异、凸显文本的特性从而提高分类精度。

粗糙集理论是波兰华沙理工大学的坡那克(Pawlak)[2]教授在 1982 年提出,这是一种分析不确定知识的强有力的数学工具。粗糙集理论是建立在分类机制的基础上的,将分类理解为在特定空间上的等价关系,而等价关系构成了对该空间的划分[3],粗糙集理论用上下近似来描述这种划分。本文的主要工作就是结合粗糙集理论的优势提出一种新的加权方法。

1　相关研究工作

目前很多学者对特征加权做了研究,其中广泛用到的加权公式是逆文本频率加权公式(TFIDF)[4],其主要思想是:如果某个词或短语在一篇文章中出现的频率高,并且在其他文章中很少出现,则认为该词或者短语具有很好的类别区分能力,适合用来分类。利用该公式进行加权,特征词的重要性与文本中的词频成正比,与文本中出现该特征词的文本频率成反比,这种加权方法比较简单有效。

在利用粗糙集进行加权这方面也有学者做了一些工作。胡清华等[5]提出了采用可变精度粗糙集模型中的近似分类质量构造特征词加权计算公式,这个方法思路比较创新,但通过分析 TFIDF 方法,笔者认为该加

收稿日期: 2010－09－23

基金项目: 浙江省"钱江人才计划"项目(2007R10013)

作者简介: 徐　欣(1986－),男,山东临沂人,硕士研究生,主要从事智能信息处理、人工智能方面的研究。

权公式只是考虑了特征词的频率以及该特征词在整个样本中的分布情况,并没有将文本的类别信息引入到权重中,存在着分类精度难以提高的局限性。该加权公式由特征词频和分类质量的乘积来构成。通过分类试验验证了可以提高样本的可分性,在支持向量机和 KNN 分类器上对文本进行分类,结果均比 TFIDF 加权的分类效果好。不过为了消除文本长度对于权值的影响,有必要对于加权公式进行归一化处理。

刘赫等[6]提出了一种基于特征重要度的特征加权方法,该方法基于实数粗糙集理论,通过计算特征重要度即计算各个类别的重合度进行加权,重合度小的重要度就大,重合度大的重要度就小。在粗糙集理论中这可以理解为两个相邻集合的边界粗糙程度,该加权方法体现了特征词对于分类的精确度,但是特征词对于文本分类的正确性却没有考虑进去。

在本文中,笔者将从近似分类精度与近似分类质量两个方面考虑特征对于分类的重要性,同时借鉴 TFIDF 加权的思想,引入文本中特征值的频度信息,从而可以更好地体现特征词对分类的重要性。

2　基于粗糙集的特征加权

2.1　粗糙集理论模型

在 Pawlak 教授提出的经典粗糙集模型中,研究对象是一个知识表示系统 $S=(U,A,V,f)$,其中,$U=\{U_1,U_2,\cdots,U_n\}$为非空有限集合,称为论域对象空间;$A=\{a_1,a_2,\cdots,a_m\}$是非空有限的属性集合;V 是 A 中属性的值域;f 为信息函数。粗糙集的核心假设认为知识体现为分类的能力,基于属性或者属性集合形成的等价关系对论域进行划分形成等价类,在粗糙集理论中等价类体现了知识的粒度,是知识的基本单元。根据已有的知识判断,一个对象 $x\in U$ 是否属于集合 X(其中 $X\subseteq U$)可分为 3 种情况,即对象 x 肯定属于集合 X,对象 x 肯定不属于集合 X,对象 x 有可能属于也有可能不属于集合 X,这主要由粗糙集的上近似$\overline{R}(X)$和下近似$\underline{R}(X)$来表示:

$$\overline{R}(X)=\{x\in U:[x]_R\cap X\neq\varnothing\}$$

$$\underline{R}(X)=\{x\in U:[x]_R\subseteq X\}$$

其中$[x]_R$是按照等价关系 R 对论域进行分类得到的包含对象 x 的等价类。在文本分类中,$[x]_R$ 可以理解为基于某个特征词对整个文本集进行划分,下近似就是确定属于某个类别子集的文本,而上近似即是可能属于某个类别子集的文本。给定一个集合 $F=\{D_1,D_2,\cdots,D_n\}$是论域 U 的一个分类,这个分类独立于等价关系 R。则该集合的下近似和上近似分别可以定义为:

$$\underline{R_a}U=\{\underline{R_a}D_1,\underline{R_a}D_2,\cdots,\underline{R_a}D_n\}$$

$$\overline{R_a}U=\{\overline{R_a}D_1,\overline{R_a}D_2,\cdots,\overline{R_a}D_n\}$$

2.2　粗糙集的近似分类质量和近似分类精度

为了度量知识的依赖性,引入粗糙集理论中的近似分类质量和近似分类精度。给定一个论域 U 和该论域上的一个等价关系 R,$X=\{X_1,X_2,\cdots,X_n\}$为论域中的一个划分,则等价关系 R 定义的集合 X 的近似分类质量 $\gamma_R(X)$为:

$$\gamma_R(X)=\frac{\sum_{k=1}^{n}|\underline{R}X_k|}{|U|}$$

近似分类质量是基于某个属性子集的关系 R 对对象进行分类时,能够确定的决策对象在整个论域中所占的比例。这个比例越大说明正域所包含的对象越多,也体现出条件属性与决策属性对论域划分的一致性程度。在文本分类中,近似分类质量可以理解为基于某个特征词形成的关系 R 对文本集进行分类时,被正确分类的文本在文本集中所占的比例。这个比例越大说明基于这个特征词的文本分类效果越好,包含的分类信息就越多;反之则说明这个特征词对于分类的作用越小。近似分类质量体现了形成关系 R 的特征词的分类能力。

给定一个论域 U 和该论域上的一个等价关系 R,$\forall X\subseteq U$ 论域中的一个划分,基于关系 R 对 X 的近似分类精度 $d_R(X)$公式如下:

$$d_R(X)=\frac{|\underline{RX}|}{|\overline{RX}|}$$

近似分类质量是基于某个条件属性子集的关系 R 对论域的对象进行分类时，确定属于某个决策的对象与可能属于某个决策的对象的比值，可以理解为由关系 R 对论域进行划分的边界的粗糙程度。近似分类质量越高，则说明边界越窄，说明了该条件属性形成的关系对于论域的分类效果好；反之则说明大部分对象处于几个类别的交叉重合部分。在文本分类中，近似分类质量就可以描述为基于某个特征词形成的关系对文本集进行分类，可以确定类别属性的文本数和不确定类别属性的文本数的比值，这可以理解为基于该特征词对文本集分类的精确程度，这个比值越高，说明基于这个特征词分类时，文本被错误划分类别的概率就越小。

2.3 基于粗糙集的特征加权

通过以上对粗糙集理论的分析，近似分类质量和近似分类精度可以在全局的角度去分析特征词对文本分类的作用，可以利于粗糙集的这些方法对特征词进行加权[7]。不过这两个度量只是体现了全局的重要性，加权过程是对每个文档中的特征词进行加权，所以在处理过程中还要考虑到单个文本的特性。笔者认为如果某个特征词在一类文本中文本频次比较高，而在整个训练集中的其他类中出现的频率又比较低，则认为该特征词含有的分类信息比较多，应该赋予较大的权重。在此分析结合了逆文本频率加权方法的合理性，将词频、近似分类质量和近似分类精度结合起来构建了加权公式。给定 n 类文本的论域按照类别划分为 $U=\{D_1,D_2,\cdots,D_n\}$，R 是基于某个特征词形成的等价关系，加权公式定义如下：

$$w_{ij}=\frac{Tf_{ij}}{\sqrt{\sum_{t=1}^{s}(Tf_{it})^2}}\cdot\frac{Cdf_k}{Df_j}\cdot\gamma_j(U)\cdot d_j(D_k)=\frac{Tf_{ij}}{\sqrt{\sum_{t=1}^{s}(Tf_{it})^2}}\cdot\frac{Cdf_k}{Df_j}\cdot\frac{\sum_{k=1}^{n}|\underline{R_a}D_k|}{|U|}\cdot\frac{|\underline{R_a}D_k|}{|\overline{R_a}D_k|}$$

其中 w_{ij} 表示第 j 个特征词在第 i 篇文本中的权重，Tf_{ij} 表示第 j 个特征词在第 i 篇文本中的出现频率，Df_j 为特征词的训练集中的文本频率，Cdf_k 为特征词在所属类子集 D_k 中的文本频率，$\gamma_j(U)$ 为第 j 个特征词对训练集的近似分类质量，$d_j(D_k)$ 为第 j 个特征词在所属类子集 D_k 中的近似分类精度，近似分类质量和近似分类精度两者共同体现了该特征词在整个训练集中对分类的重要程度。

在构建决策表的时候要包含训练集中的所有特征词，但是某些特征词在单个文本中的出现频率是零，这就出现了相当多的冗余信息并干扰了分类信息的提取，比如对于某些低频的特征词只出现了很少的次数，则在计算某类的上近似时，得到的结果几乎就是整个训练集，所以在计算 $d_j(D_k)$ 的时候只使用词频为非零值的特征词来划分等价类，然后对于每个类计算出近似分类精度。

3 实验与分析

为了验证基于粗糙集的加权方法在文本分类中的效果，在试验中将采用 KNN 算法设计文本分类器，然后分别采用粗糙集加权方法和 TFIDF 方法对特征进行加权，通过实验结果的对比来检验该加权方法的有效性。

3.1 实验环境

硬件环境：DELL(OPTIPLEX 755)，Intel(R) Core(TM)2 Quad CPU，内存 2 G；

软件环境：Windows Vista™ Home Basic，MyEclipse7.5，JDK1.6.0_17。

3.2 系统框架与实验数据集

KNN 是一种简单而且分类效果很好的算法[8]，其基本思想是：给定一个测试文本，通过计算其与训练集中的各个文本之间的相似度，选取与之最相似的 K 个文本，然后通过一定策略根据这 K 篇文本来确定测试文本的类别。基于 KNN 算法的文本分类系统的框架如图 1。

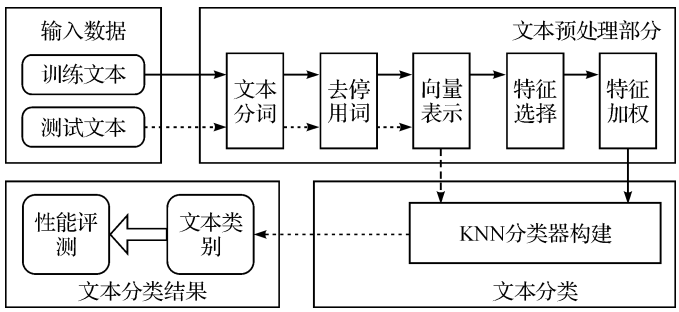

图 1 基于粗糙集理论的文本分类模型

在实验中使用的数据集是复旦大学整理提供的分类语料库，从该语料库中选教育、环境、农业、

经济、医药、政治六大类共计 480 篇文献作为实验语料,其中训练集为 300 篇,测试集为 180 篇。

3.3　实验结果与分析

通过对训练集文本进行分词、过滤停用词、词频统计等的预处理,可以得到每个文本的词频、文本频率等信息。利用特征词作为条件属性,文本类别作为决策属性构建决策表,将文本的词频作为条件属性的值,可以计算出基于该特征词在整个训练集中形成的等价关系的上下近似,从而利用加权公式计算出权值。对于训练集中的农业类的一个文本特征权重计算结果如表 1(tf 表示该特征词在这个文本中出现的次数,df 表示该特征词在多少文本中出现,cdf 表示在该类文本中有多少文本包含该特征词,$TFIDF$ 表示逆文本频率加权的权重,CW 表示基于粗糙集加权的权重)。

表 1　对于关于农业的一篇文本的 tf、df、r、$TFIDF$ 和粗糙集加权的计算结果

特征词	tf	df	cdf	$\gamma_i(U)$	$TFIDF$	$CW(10^{-4})$
渠道	2	53	21	0.005 0	4.161 4	0.338 3
滩涂	1	5	4	0.012 5	4.382 2	0.422 9
产量	1	46	30	0.027 5	2.163 9	0.930 3
农户	19	38	37	0.080 0	44.741 7	51.420 8
入手	1	31	5	0.000 0	2.624 9	0.000 0
全省	11	16	6	0.007 5	35.412 0	2.790 9
流通	8	43	23	0.012 5	18.387 6	3.382 9
……						

通过表 1 对比 $TFIDF$ 和 CW 的值可以发现,虽然不同的加权得到的权值是不同的,但是体现特征词的重要度的趋势还是一致的,不过对于有些特征词赋予了更突出的权重,这也就更好地体现了特征词的重要程度。下面就分别用 TFIDF 加权算法和基于粗糙集的加权方法对特征词进行加权,构建 KNN 分类器进行训练和测试(其中 K 取值 20)。分别统计各个分类的准确率、召回率和 $F1$ 值,分类结果见表 2。

表 2　基于 TFIDF 加权的分类结果,基于粗糙集加权的分类结果对比

类别	文本数/篇	基于 TFIDF 加权的分类			基于粗糙集加权的分类		
		准确率/%	召回率/%	$F1$/%	准确率/%	召回率/%	$F1$/%
教育	30	80.95	56.67	66.67	89.47	56.67	69.39
环境	30	71.43	66.67	68.97	94.74	60.00	73.47
农业	30	51.11	76.67	61.33	43.40	76.67	55.43
经济	30	47.62	33.33	39.21	50.00	40.00	44.44
医药	30	82.76	80.00	81.36	80.65	83.33	81.97
政治	30	61.90	43.33	50.98	78.26	60.00	67.92

试验选取了 300 篇文献作为训练集,在不同的类中选取相同数量的文档进行训练,这就排除了不平衡类别因素对分类影响。在预处理过程中,虽然通过了停用词过滤,但是特征词的数量是非常庞大的,构建决策表的条件属性甚至达到 26 356 个,这大大增加了计算粗糙集的上下近似的时间,所以在分类过程中非常有必要对特征进行降维。

从试验数据中可以看出,在相同的文本分类器上使用 TFIDF 加权算法和基于粗糙集的加权方法对相同的训练集和测试集进行实验,除了在农业类中分类效果没有提高,其他类别的文本分类效果均有提高,这也说明基于粗糙集的加权方法是可行有效的。不过对于某些类别,分类效果不是很理想。这主要有两方面原因,一是这两类文本主题比较相近,同时也体现出本实验设计的分类器分类精度需要提高;另一方面原因就是训练不够,训练集的规模比较小,各类的信息没有充分体现。总体说来基于粗糙集加权的分类效果比使用 TFIDF 加权方法的分类效果有所提高。

4　结　语

粗糙集在文本分类中的主要应用是属性约简,而在本文中通过分析其原理,将粗糙集理论创新性地应用到加权过程中。由于逆文本加权方法是考虑了特征词在整个文本中的分布情况和在文本中的出现频率,对于文本的类别信息并没有体现在权重中。在本文提出的基于粗糙集的加权方法中将文本在训练集中所属类的子集的分布信息引入权重,更合理地体现了特征词的重要性。通过在 KNN 分类系统中的测试结果表明,本文提出的加权方法比逆文本频率加权方法取得了更好的分类效果。但是在试验中出现了有的特征值的权值为零的情况,这主要是在计算近似分类质量的时候产生的,这些信息并非完全没有分类作用,在后续工作中将对加权中出现的零值进行处理。

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)：1848-1855.

[2] Zdzislaw Pawlak，Andrzej Skowron. Rudiments of rough sets[J]. Information Sciences，2007，177：3-27.

[3] 王国胤. Rough集理论与知识获取[M]. 西安：西安交通大学出版社，2001.

[4] Salton G，Mc Gill M J. Introduction to Modern Information Retrieval[M]. New York：Mc Graw-Hill Book Co，1983.

[5] 胡清华，谢宗霞，于达仁. 基于粗糙集加权的文本分类方法研究[J]. 情报学报，2005，24(1)：59-63.

[6] 刘 赫，刘大有，裴志利，等. 一种基于特征重要度的文本分类特征加权方法[J]. 计算机研究与发展，2009，46(10)：1693-1703.

[7] Huang Li-can，Xu Xin，Zhao Yu-hong，et al. Feature weighting scheme for text categorization based on rough Set[C]//The First International Conference on Networking and Distributed Computing(ICNDC2010)，2010：186-188.

[8] Yang Yi-ming，Liu Xin. A re-examination of text categorization methods[C]//Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR'99)，1999：42-49.

Text Categorization by Feature Weighting Scheme Based on Rough Set

XU Xin，*HUANG Li-can*，*ZHAO Yu-hong*

(School of Informatics and Electronics，Zhejiang Sci-Tech University，Hangzhou 310018，China)

Abstract：Text Categorization is the focus of many areas like Information Retrieval，Data Mining and so on. Feature weighting is an important problem in text categorization. For computing feature weights，this paper presents a feature weighting scheme for text categorization based on rough set theory. The authors analyze the characteristics of rough set theory and TF-IDF，and consider the overall influence which the keywords establish over the classification from the aspects of approximation accuracy and approximation quality. The decision information of a feature for categorization is introduced into the weight of this feature，and the importance of the feature will be fully reflected. The experimental results indicate the effectiveness of approach.

Key words：rough set；feature weighting；text categorization；approximation accuracy；approximation quality

(责任编辑：陈和榜)