计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(4):131−137 [doi: 10.15888/j.cnki.csa.006287]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

粗糙集规则匹配算法及其在文本分类中的应用①

朱敏玲[1], 吴海艋[1], 石　磊[2]

[1](北京信息科技大学 计算机学院, 北京 100101)
[2](中国科学院 软件研究所, 北京 100190)
通讯作者: 朱敏玲, E-mail: zhuminling@bistu.edu.cn

摘　要: 为提高中文文本分类的效果, 提出了一种基于粗糙集理论的规则匹配方法. 在对文本特征的提取过程中, 对 CHI 统计方法进行了适当的改进, 并对特征项的权值进行了缩放和离散化. 结合区分矩阵实现关于粗糙集理论的属性约简和规则提取, 并采用规则预检验的方法对规则匹配的决策参数进行优化, 以提高中文文本分类的效果. 实验结果表明改进后的规则匹配方法分类准确率更高, 同时在训练数据较少的情况下也可以取得不错的效果.

关键词: 粗糙集; 中文文本分类; 属性约简; 规则提取; 规则匹配

引用格式: 朱敏玲,吴海艋,石磊.粗糙集规则匹配算法及其在文本分类中的应用.计算机系统应用,2018,27(4):131–137. http://www.c-s-a.org.cn/1003-3254/6287.html

Rough Set Rule Matching Method and its Application in Text Categorization

ZHU Min-Ling[1], WU Hai-Meng[1], SHI Lei[2]

[1](Computer School, Beijing Information Science and Technology University, Beijing 100101, China)
[2](Institute of Software, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: To improve the performance of Chinese text classification, a rule matching method based on rough set theory is proposed in this study. In the extracting process of textual features, the CHI statistical method is improved and the weight of the feature is scaled and discretized. It combines the discriminant matrix to achieve the attribute reduction and rule extraction for rough set theory, and uses rule pre-test method to optimize the decision parameters of rule matching to improve the effect of Chinese text categorization. The experimental results demonstrate that the categorization accuracy of the improved matching method is higher, and in the case of less training data, it can also achieve decent results

Key words: rough set; Chinese text classification; attribute reduction; rule extraction; rule matching

随着信息技术的飞速发展, 万维网的文本信息量急剧增长[1]. 2008 年 7 月 26 日, 谷歌在官方微博中称, 其索引的网页数量已经突破 1 万亿幅, 截止至 2014 年 12 月底, 这一数值更是突破了 30 万亿幅大关, 并以每日 50 亿的增长速度持续递增[2]. 可见, 如何从庞大的网页数据中获得有用信息成为人们的迫切需求, 而自动文本分类是获取相关信息的一种方法[3].

目前, 文本分类领域常用的方法有支持向量机 (SVM), 朴素贝叶斯 (Naïve Bayes), K 近邻 (KNN), 决策树方法 (Decision Tree) 等[4], 与这些传统的分类方法相比, 粗糙集理论用于分类的优点在于其能够通过属性约简在不影响分类精度条件下降低特征向量的维数, 从而获得分类所需的最小特征子集, 并配合值约简得到最简的显式分类规则[5], 最后根据粗糙集的规则匹配方法对待分类文本进行有效的分类.

本文首先对粗糙集理论和中文文本分类的相关知识进行介绍与分析, 及如何将中文文本转化为粗糙集所能处理的知识库系统, 和如何通过粗糙集的属性约

① 基金项目: 国家自然科学基金 (11401031); 北京信息科技大学 2016-2017 学年度“实培计划”项目
收稿时间: 2017-07-16; 修改时间: 2017-07-28; 采用时间: 2017-08-09; csa 在线出版时间: 2018-03-31

简和值约简来实现规则的提取; 然后, 分析本研究中提出的粗糙集规则匹配的改进算法; 再次, 对原始方法和改进算法进行对比实验, 并对实验数据进行对比和分析; 最后, 对本研究工作进行了总结.

1 相关知识

1.1 粗糙集理论

粗糙集 (Rough Set, RS) 理论是由波兰华沙理工大学的 Pawlak 教授在 1982 年提出的一种新的数学工具, 它能有效地处理和分析不精确、不协调和不完备的信息, 并从中发现隐含的知识和潜在的规律[6]. 本文通过粗糙集理论中的知识约简对文本进行分类规则提取, 并通过改进的粗糙集匹配方法对新的待分类文本进行规则匹配和文本分类[7].

定义 1. 设$S=(U,C\cup D)$是一个决策表, C 是条件属性集, D 是决策属性, $C\cap D\neq\phi$. 若$P\subseteq C$为满足$Pos_c(D)=Pos_p(D)$的极小属性子集, 则称 P 为决策表 S 的一个约简$Red(S)$[8].

定义 2. 设有信息系统 S, $a(x)$是记载 x 在属性 a 上的值, C_{ij}表示矩阵中的第 i 行, 第 j 列元素, C_{ij}被定义为:

$$C_{ij}=\begin{cases}\{a\in A|a(x_i)\neq a(x_j)\} & D(x_i)\neq D(x_j)\\ 0 & D(x_i)=D(x_j)\end{cases}$$

定义 3. (区分函数) 区分函数是从分辨矩阵中创造的. 约简算法是先求C_{ij}的每个属性的析取, 再求其合取[9].

1.2 中文文本分类

相比于英文文本分类, 中文文本分类的一个主要差别在于预处理阶段, 因为中文文本的词与词之间没有明显的切分标志, 不像英文文本的单词那样有空格来区分. 首先, 通过现有的分词技术来对中文文本进行分词处理, 并在此基础上提取一些重要的文本特征来将文本表示在向量空间. 本文的重点在于如何通过向量空间模型 (VSM) 和特定的特征选择函数, 将文本分出的字、词、词组或概念转化为粗糙集理论所能处理的知识库或信息系统, 关键词集即为信息系统中的条件属性集, 文本类别集即为决策属性集. 通过 Skowron 提出的区分矩阵进行属性约简和规则提取[10], 生成决策规则表, 最后采用改进的规则匹配方法确定每条规则的规则支持度, 最终作用于新文本的分类匹配中.

2 规则提取

2.1 文本预处理

文本预处理的过程主要包括:分词处理、停用词过滤、文本特征提取等[11]. 本文采用 IKAnalyzer 分词工具, 它是一款以开源项目 Lucene 为应用主体, 结合词典分词和文法分析算法的中文分词组件, 其采用了特有的“正向迭代最细粒度切分算法”, 有词性标注、命名实体识别、分词排歧义处理和数量词合并输出等功能, 并支持个人词条的优化的词典存储, 如“北京奥运会”, “1949 年”, “反装甲狙击车”被纳入用户词典后, 可被正确分为一个词条, 而不会拆分为“北京”、“奥运会”, “1949”, “年”, “反”, “装甲”, “狙击”, “车”, 同时停用词过滤可以将文本使用频率较大但对文本分类没有实际作用的字、词和词组, 例如: “的”, “和”, “同时”等, 以及网络文本中的格式标签进行去除, 例如: “@123456”, “本文来源”, “相关新闻”, “组图”等, 该分词工具可在不影响文本原信息表达的情况下进行中文分词, 在文本分词预处理中具有比较好的效果[12].

2.2 特征选择

在文本分类中, 常用的特征选择函数有信息增益 IG (Information Gain), 期望交叉熵 ECE (Expected Cross Entropy), 互信息 MI (Mutual Information) 等[13]. 但是它们并不按类别计算统计值, 所以选出的特征词往往都是全局意义上的, 而实际情况中, 往往很多极具类别区分度的词, 如“剧组”, “直升机”, “导弹”, “演员”, “电子书”等, 根据函数计算出的值不是很大, 很可能被除掉, 为了避免以上情的发生, 本文采用 CHI 统计方法进行特征词的选择[14], 选出的特征词往往更具备类别区分度, 其定义如公式 (1) 所示.

$$\chi 2(w,D_j)=\frac{N(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)} \quad (1)$$

其中, w 代表特定词汇, D_j 代表文本类别, N 为文本总篇数; A 为词汇 w 与类别 D_j 共现的文本篇数; B 为词汇 w 出现类别 D_j 不出现的文本篇数 C 为类别 D_j 出现而词而汇 w 不出现的文本篇数; D 为词汇 w 和类别 D_j 均不出现的文本篇数.

一般特征项的 CHI 值选取为对所有类别的 CHI 平均值或最大值, 但是 CHI 统计方法由于考虑了特征项与类别的负相关性. 所以, 在实际情况中, 选词结果往往偏向于类别区分度更高的那一类或那几类文

本, 而对于文本内容比较相似、区分度较低的文本, 选出的词函数值普遍偏低, 从而只有较少的类别区分词被选中, 对后续的粗糙集知识库的知识约简造成影响. 故本文对 CHI 特征选择算法进行了改进, 规定选取时特征项的 CHI 值为其对所有类别的 CHI 最大值, 并加入新的选择公式对每类文本的特征词数量进行重新分配, 使选择出的特征词更偏向于类别区分度较低的几类文本. 假设从 K 类文本中选取出 N 个特征项, 改进后的公式 (2).

$$N\left(D_{\mathrm{j}}\right)=\frac{N}{k}\frac{AVG(N)}{AVG\left(N,D_{\mathrm{j}}\right)} \tag{2}$$

即在原方法中, 每类文本平均分到的特征词数量为 N/k, 由于原 CHI 方法在特征选择上对类别区分度较高的文本的偏袒, 因此类别区分度较低的那几类文本实际分到的特征词数量将小于 N/k, 改进后的公式在 N/k 的基础上乘以类别因子$AVG(N)/AVG\left(N,D_{\mathrm{j}}\right)$消除后者在特征词数量上的劣势, 其中$AVG(N)$为全部文本的前 N 个特征项的 CHI 平均值, $AVG\left(N,D_{\mathrm{j}}\right)$为类别 D_j 中前 N 个特征项的 CHI 值平均值. 从式 (2) 可以看出, 类别区分度较小的类别, 其$AVG(N)/AVG\left(N,D_{\mathrm{j}}\right)$更大, 故实际分到的特征词数也更多. 这也更有利于接下来的粗糙集属性约简, 因为在类别区分度较大的类别中, 过多的特征词必定造成条件属性的冗余, 加大属性约简的负担, 甚至影响属性约简的结果.

2.3 生成文本决策表

根据改进后的 CHI 特征选择方法选出前 N 个特征词组成了决策表的条件属性集, 文本类别集合组成了决策属性集. 特征词的权重根据 TF-IDF 公式计算, 如公式 (3).

$$Weight_{ik}=tf_{ik}\times idf_{ik} \tag{3}$$

其中, tf_{ik}为特征项t_k在文本 d_i 中出现的频率, idf_{ik}为特征项 t_k 的逆向文档频率.

考虑到 TF-IDF 公式计算出的权值为连续值, 因此还需要对连续值进行离散化, 如公式 (4).

$$W_{ik}=[(Weight_{ik}-W_{\min})/(W_{\max}-W_{\min})]*(b-a)+a \tag{4}$$

其中, $Weight_{ik}$表示该特征词 i 在文本 k 中的权值, $W_{\min}$和$W_{\max}$分别表示特征词 i 在所有决策表中的最小值和最大值. a 和 b 表示缩放范围$[a, b]$.本文中对 $Weight_{ik}$为 0 的项, W_{ik}取 0, 其余项根据缩放范围取[1, 3]进行权值离散化.并对最终结果取整 (如 1.123 取值为 1) 作为离散化后的权值. 经过离散化后的决策表 1 所示.

表 1 文本分类决策表

K	C_1	C_2		C_n	D
1	1	0		0	1
2	0	3	…	0	2
3	0	0		2	3
4	0	0		1	3
		…			

注: K表示文本编号, C_n表示特征词, D表示文本K的类别.

2.4 决策表的规则提取

在规则提取上分两步走, 首先进行特征词的属性约简, 随后再进行属性值约简.

2.4.1 属性约简

为删除对文本分类决策没有影响的特征词, 利用粗糙集的属性约简能力在保证决策表分类能力不变的前提之下, 删除其中不相关、对决策结果不会造成影响的条件属性, 即文本特征词, 从而达到属性约简和降低特征维数的目的[15].

Skowron 教授提出的区分矩阵和区分函数可以通过区分函数中的极小析取范式进行合取, 获得知识系统中的所有属性约简的集合, 但是对于最优约简子集的选择一直都是一个 NP 问题[16], 因此不在本文的讨论范围之内. 本文直接选取所有属性约简集合中条件属性最少的约简子集生成新的约简决策表, 并通过从约简决策表中减少条件属性的方法, 依次计算每个条件属性的重要程度, 作为后续规则匹配中的一个重要参数, 如公式 (5).

$$SIG(a,R,D)=\gamma_{R+C}(D)/\gamma_R(D) \tag{5}$$

其中$\gamma_{R+C}(D)$为约简决策表$DT=\langle U,C\cup D,V,f\rangle$论域中相容决策项的比率, $\gamma_R(D)$为去除某一条件属性 C 后其相容决策项的比率. 此时的$SIG(a,R,D)$恒大于 1, 因为 $SIG(a,R,D)$=1 的条件属性已经在属性约简阶段被去除. $\gamma_{R+C}(D)/\gamma_R(D)$的比值越大, 说明该条件属性的重要程度越高. 在分类匹配的过程中, 与规则匹配前件数及规则支持度等指标配合使用来计算每条分类规则的规则强度.

2.4.2 值约简

与粗糙集理论的属性约简相比, 值约简再次用到了区分矩阵获取每一项中的极小析取范式, 但两者的

不同之处在于, 在对结果进行合取转化时, 属性约简是从全局出发, 对所有的极小析取范式进行统一的合取化, 其结果为所有属性约简结果的集合[17]. 而值约简中是对区分矩阵的每一行进行合取化, 每一条完整的规则最终被约简为了多个能区分其他不同类别的最小规则集合.

区分矩阵$M(S)=\{M[i,j]\}$定义如公式 (6).

$$M[i,j]=\begin{cases}\{a\in C: f(x_i,a)\neq f(x_j,a)\}, & f(x,D)\neq f(x,D)\\ \phi(\text{空集}), & \text{其他}\end{cases} \tag{6}$$

例如表 2 所示的一张决策表, 其中共有 5 个文本 ($k\in[1,5]$) 和 3 个关键词组成的条件属性集 $C=\{C1,C2,C3\}$, D 为决策属性[18].

表 2 决策表

K	C_1	C_2	C_3	D
1	1	0	0	1
2	0	3	0	2
3	0	0	2	3
4	0	0	1	3
5	1	1	0	1

根据表 2 构造的区分矩阵如表 3 所示.

表 3 区分矩阵

K	1	2	3	4	5
1		$C1C2$	$C1C3$	$C1C3$	ϕ
2	$C1C2$		$C2C3$	$C1C2C3$	$C1C2$
3	$C1C3$	$C2C3$		ϕ	$C1C2C3$
4	$C1C3$	$C1C2C3$	ϕ		$C1C2C3$
5	ϕ	$C1C2$	$C1C2C3$	$C1C2C3$	

以表 3 的第 2 行为例, 根据区分矩阵获取第 i 行完全规则的约简规则的步骤如下.

步骤 1. 把每一行的空项和重复项去除, 获得互不重复的最小析取范式集.处理后的第 2 行, 第 2 项和第 5 项被去除, 剩下$C1\wedge C2, C2\wedge C3, C1\wedge C2\wedge C3$.

步骤 2. 把每一行的最小析取范式进行合取化, 获得约简规则集的条件属性下标集合. 第 2 行提取出的规则集合表示为:

$$\begin{aligned}ReductRule(2)&=(C1\wedge C2)\vee(C2\wedge C3)\vee(C1\wedge C2\wedge C3)\\&=(C1\vee C2)\wedge(C1\vee C2\vee C3)\wedge\\&\quad(C1\vee C3)\wedge C2\wedge(C2\vee C3)\end{aligned}$$

步骤 3. 根据离散定律中的吸收律和幂等律删除冗余和包含关系, 获得每一行的最简规则集合. 第 2 行的最简规则集合表示为:

$$ReductRule(2)=C2\wedge(C1\vee C3) \tag{7}$$

经过值约简后导出的约简规则如表 4 所示, *代表约简掉的属性权值.

表 4 决策规则表

K	$C1$	$C2$	$C3$	D
1	1	*	*	1
1	*	0	0	1
2	*	3	*	2
2	0	*	0	2
3	0	0	*	3
3	*	*	2	3
4	*	*	1	3
4	0	*	*	3
5	1	*	*	1
5	*	1	*	1

然后, 对约简后的决策表中的重复规则和冗余规则进行合并, 可得出表 5 的决策规则表.

表 5 决策规则表

K	$C1$	$C2$	$C3$	D
1	1	*	*	1
2	*	0	0	1
3	*	3	*	2
4	0	*	0	2
5	*	*	2	3
6	*	*	1	3
7	0	*	*	3
8	*	1	*	1

则, 化简后的规则如下:

$$\begin{aligned}&C1(1)\rightarrow D(1)\\&C2(0)\wedge C3(0)\rightarrow D(1)\\&C2(3)\rightarrow D(2)\\&C1(0)\wedge C3(0)\rightarrow D(2)\\&C3(2)\rightarrow D(3)\\&C3(1)\rightarrow D(3)\\&C1(0)\rightarrow D(3)\\&C2(1)\rightarrow D(1)\end{aligned}$$

对表 2 和表 5 分析可知, 经过值约简后的决策规则表, 每条规则的条件前件长度得到了进一步的缩减, 同时每两条规则之间互不冲突, 并且与原决策表的完整规则一一对应. 约简后的规则集更加清晰明了, 也具可解释性.

3 规则匹配

决策规则生成之后, 就可以运用规则对新数据项

或文本进行预测和分类. 基于粗糙集的规则匹配分为完全匹配和部分匹配两个阶段.

3.1 完全匹配

1) 完全匹配的基本步骤

步骤 1. 在分类器中对新数据项进行规则化处理, 抽取出与完全规则条件属性一一对应的表达式.

步骤 2. 在决策规则集中进行规则查找, 如果有且只有一条规则与之完全对应, 则新数据项的类别归至该决策规则所属的类别; 如果遍历完所有规则后, 没有任何规则与之相匹配, 则把该数据项归入待定项进入部分匹配阶段.

步骤 3. 如果出现多个规则的前件与该数据项相匹配, 则根据规则支持度的排序, 把支持度最高的规则的决策类别定义为新对象的类别, 如公式 (7).

$$\omega(R)=\sum Strength(R)\times Specificity(R) \tag{8}$$

其中, $Strength(R)$是规则强度 (*Strength*), 即训练集中与之匹配的训练项个数; $Specificity(R)$是规则专指度 (*Specificity*), 即规则中条件属性前件的个数; $\omega(R)$是规则支持度[19].

但是, 由于规则专指度会对规则中属性条件较长的规则有所偏袒, 导致完全匹配的规则结果往往选出条件属性数较多的规则作为分类的依据, 这与粗糙集理论的本意有所矛盾. 故本文对完全匹配的算法进行了改进, 在完全匹配阶段之前, 对约简规则进行规则预检验.

2) 规则预检验

规则预检验的过程分为如下几个步骤.

步骤 1. 选取一份新的验证集, 并进行规则化.

步骤 2. 将约简规则与验证集进行比较, 依次求出规则强度和规则置信度 (*Confidence*).

此时的规则支持度可表示为公式 (8).

$$\omega(R)=\sum Strength(R)\times Confidence(R) \tag{9}$$

其中, $Confidence(R)$是规则置信度, 即约简规则与验证集的规则条件匹配且类别标签相同的比率. $\omega(R)$值越大, 表示根据该规则推导出的类别标签的可信赖程度越大, 在多个规则同时满足匹配条件的情况下选择$\omega(R)$值最高的规则的类别进行匹配, 其结果的准确率往往更高. 同时, 如果某一新数据项完全匹配出的规则的$\omega(R)=0$或没有任何规则与之相匹配, 则把该数据项归入待定项进入部分匹配阶段.

从以上步骤得知, 规则预检验的方法是基于规则支持度$\omega(R)$而展开的, 其也存在一些缺憾. 若选取的支持度过高, 则某些有价值的规则模式不能被获取; 反之, 过低时会产生很多无实际意义的规则模式, 分类系统性能下降. 本文通过实际训练来选取合适的特征词数来弥补其缺憾.

3.2 部分匹配

部分匹配的基本过程是逐一减少新数据项的条件属性个数, 直到出现一条或多条规则能与之匹配为止. 其匹配思路与完全匹配基本相同. 因此, 部分匹配的规则支持度$\omega(R)$可以表示为公式 (9).

$$\begin{aligned}&\omega(R)=\sum Strength(R-C)\times Confidence(R)\\&Strength(R-C)=Matching(R)\times\frac{N-Nc}{N}\end{aligned} \tag{10}$$

其中, N 为表示新对象的总条件属性个数, Nc 表示部分匹配过程中去掉的条件属性个数.

同时, 关于对新数据项条件属性的去除次序的确定方法, 本文规定, 条件属性去除的先后次序与文章之前的属性约简过程中计算的$SIG(a,R,D)$的升序次序保持一致, 即属性重要程度低的属性会在部分匹配的过程中优先被去除.

经过完全匹配和部分匹配之后, 如果出现没有与现有规则相匹配的数据项, 则将验证集中规则支持度最高的结果赋给该项. 到此为止, 规则匹配完全结束.

4 实验结果及分析

为验证该分类器的效果, 进行了如下的实验验证. 首先, 选取合理的训练集是非常必要的. 因为训练集的文本数、类别数及特征项数对于分类器的执行效果都有重大影响[20]. 在此, 选取了 UCI(University of California Irvine) 数据库中的 iris 和 diabetes 数据集和 Statlog 中的 australian 和 heart 数据集作为训练样本, 在每个数据集中任意选择了 3 类数据. 同时, 考虑到分类器默认情况下假设的样本数是大致均匀分布的, 如果一类比其他类数据量大得过分, 分类器会把其他类的数据判为大的类别上, 从而换取平均误差最小. 为了避免该情况的发生, 采取不同的样本比例进行训练的方法.

然后, 采用第 2 节中提到的方法对原始数据进行处理, 并把数据样本分别按 1:1:1 和 5:2:3 的比例随机打乱, 各生成 10 份不同的训练集, 并记录平均的分类

准确率情况, 实验结果如表 6 所示. 改进后的匹配方法在 4 组数据集上的准确率相比于原方法, 分类效果均有不同程度的提升. 同时, 改进后的匹配方法在训练集数据较少的情况下仍获得不错的分类效果.

表 6 4 种训练集的训练结果

训练集	准确率(%)			
	比例1:1:1		比例5:2:3	
	原方法	改进方法	原方法	改进方法
iris	86.89	93.62	92.44	95.83
diabetes	68.73	76.63	73.62	79.81
australian	70.21	81.73	75.62	85.26
heart	68.25	76.16	72.83	78.62

表 7 为特征词数相同而取不同训练样本数量时, 2 种匹配方法的训练结果, 数据集采用 UCI 的 iris 数据集. 对表 7 进行对比分析, 可以看出改进后的匹配方法在训练数据取不同数量的情况下, 均获得不错的分类效果; 同时, 在训练数据小于测试数据的情况下分类效果的提升更加明显. 因此, 在对训练数据量有限的数据进行分类的时候, 改进后的匹配方法更加实用.

表 7 iris 训练集的训练结果 (特征词数=125)

训练文本数	准确率(%)	
	原匹配方法	改进后的方法
200	82.68	92.96
240	86.89	93.62
280	88.24	94.93
320	90.12	95.36
360	92.44	95.83
400	93.01	94.44

表 8 为训练样本数相同, 而特征词数不同的情况下, 原方法与本文改进后方法的执行结果. 可以看出, 并不是特征词数量越多准确率越高. 当训练文本数都取 360 时, 特征词数量较少的情况下, 改进方法的分类效果更佳; 特征词数大于 125 后, 两种匹配方法的分类效果相差不大.

表 8 特征词数对分类效果的影响 (训练文本数=360)

特征词数	准确率(%)	
	原方法	改进方法
50	79.13	86.85
75	85.65	91.26
100	90.83	93.32
125	92.63	94.25
150	93.16	94.37
200	92.78	94.18

经过训练集的训练后, 不仅验证了粗糙集约简的效率, 也验证了本文规则提取方法的合理性, 同时得到比较好的特征词参数范围.

最后, 运用一般数据进行测试, 验证其泛化能力等. 从网上下载和收集了来自腾讯新闻、凤凰新闻、新浪新闻及网易新闻的新闻报道组成的语料库, 从中选用了军事、娱乐、阅读和法制四个类别共 600 篇文章作为实验语料. 从特征词数量和训练文本数量两个方面对改进前后的匹配方法进行分析, 实验结果如表 9 所示. 由表 9 可知, 当特征词数都取 125 时, 测试文本取不同数量的情况下, 改进方法的分类效果均有不同程度的提高; 同时, 在测试文本数较少时, 改进方法对分类效果的提升更加明显.

表 9 训练文本数对分类效果的影响 (特征词数=125)

测试文本数	准确率(%)	
	原方法	改进方法
60	81.68	88.57
75	84.82	90.04
90	87.36	92.58
105	90.62	93.96
120	92.63	94.25

5 结束语

本文把粗糙集理论应用于中文文本分类的规则提取和规则匹配中, 并对基于 CHI 方法的类别特征词选取方法进行了相应的改进, 使其更加适用于粗糙集的知识约简. 在训练阶段使用区分矩阵对完整决策规则进行属性约简和规则提取, 并通过规则预验证的方法对规则支持度进行优化; 同时, 通过调整特征词的数量来弥补规则预检验方法所带来的信息损失而影响有效规则提取的问题. 实验结果表明, 改进后的规则匹配方法在实际的文本分类中分类准确率更高, 同时在一定程度上克服了原匹配方法容易选出条件前件数较多的规则的缺点, 也使得匹配出的规则更加简单明了, 更具可解释性.

参考文献

1 Fan W, Bifet A. Mining big data: Current status, and forecast to the future. ACM SIGKDD Explorations Newsletter, 2012, 14(2): 1–5.

2 朱基钗, 高亢, 刘硕. 互联网络发展状况统计. 党政论坛·干部文摘, 2016, (9): 19. [doi: 10.3969/j.issn.1006-1754.2017.01.016]

3 Shen YD, Eiter T. Evaluating epistemic negation in answer

set programming. Artificial Intelligence, 2016, 237: 115–135. [doi: 10.1016/j.artint.2016.04.004]
4 吴德, 刘三阳, 梁锦锦. 多类文本分类算法 GS-SVDD. 计算机科学, 2016, 43(8): 190–193. [doi: 10.11896/j.issn.1002-137X.2016.08.038]
5 程学旗, 兰艳艳. 网络大数据的文本内容分析. 大数据, 2015, (3): 62–71.
6 朱敏玲. 属性序下的粗糙集与 KNN 相结合的英文文本分类研究. 黑龙江大学自然科学学报, 2015, 32(3): 404–408.
7 Mitra S, Pal SK, Mitra P. Data mining in soft computing framework: A survey. IEEE Transactions on Neural Networks, 2002, 13(1): 3–14. [doi: 10.1109/72.977258]
8 Miao DQ, Duan QG, Zhang HY, *et al*. Rough set based hybrid algorithm for text classification. Expert Systems with Applications, 2009, 36(5): 9168–9174. [doi: 10.1016/j.eswa.2008.12.026]
9 Grzymala-Busse WJ. Rough set theory with applications to data mining. In: Negoita M, Reusch B, eds. Real World Applications of Computational Intelligence. Berlin, Heidelberg, Germany: Springer, 2005.
10 Pawlak Z, Skowron A. Rudiments of rough sets. Information Sciences, 2007, 177(1): 3–27. [doi: 10.1016/j.ins.2006.06.003]
11 朱敏玲. 基于粗糙集与向量机的文本分类算法研究. 北京信息科技大学学报, 2015, 30(4): 31–34.
12 马晓玲, 金碧漪, 范并思. 中文文本情感倾向分析研究. 情报资料工作, 2013, 34(1): 52–56.
13 李扬, 潘泉, 杨涛. 基于短文本情感分析的敏感信息识别. 西安交通大学学报, 2016, 50(9): 80–84. [doi: 10.7652/xjtuxb201609013]
14 黄章树, 叶志龙. 基于改进的 CHI 统计方法在文本分类中的应用. 计算机系统应用, 2016, 25(11): 136–140.
15 梁海龙. 基于邻域粗糙集的属性约简和样本约减算法研究及在文本分类中的应用[硕士学位论文]. 太原: 太原理工大学, 2015.
16 杨传健, 葛浩, 汪志圣. 基于粗糙集的属性约简方法研究综述. 计算机应用研究, 2012, 29(1): 16–20.
17 胡清华, 于达仁, 谢宗霞. 基于邻域粒化和粗糙逼近的数值属性约简. 软件学报, 2008, 19(3): 640–649.
18 段洁, 胡清华, 张灵均, 等. 基于邻域粗糙集的多标记分类特征选择算法. 计算机研究与发展, 2015, 52(1): 56–65. [doi: 10.7544/issn1000-1239.2015.20140544]
19 时希杰, 沈睿芳, 吴育华. 基于粗糙集的两阶段规则提取算法与有效性度量. 计算机工程, 2006, 32(3): 60–62.
20 李湘东, 曹环, 黄莉. 文本分类中训练集相关数量指标的影响研究. 计算机应用研究, 2014, 31(11): 3324–3327. [doi: 10.3969/j.issn.1001-3695.2014.11.028]