

采用类别相似度聚合的关联文本分类方法

田丰^{1,2}, 桂小林^{1,2}, 杨攀^{1,2}, 王刚^{1,2,3}, 郭岳龙^{1,2}

(1. 西安交通大学电子与信息工程学院, 710049, 西安; 2. 西安交通大学陕西省计算机网络重点实验室, 710049, 西安;
3. 西安财经学院信息学院, 710100, 西安)

摘要: 针对基于关联规则的分类方法在分类时仅考虑规则的置信度并使用规则修剪技术, 导致分类器的分类精度难以进一步提高的问题, 提出了一种基于类别相似度聚合的关联文本分类方法. 该方法采用修改的 χ^2 统计技术提取各类别的特征词; 为保证规则匹配的精度和速度, 使用 CR-tree 存储分类规则, 并给出了 CR-tree 的构建与匹配算法; 采用向量内积来计算文本类别分量与类别标志向量的相似度, 进而使用规则置信度和类别相似度的聚合值作为文本分类的依据. 基于实际网络文本的实验表明, 该方法仅需提取 30 个特征词, 分类结果的微平均值即可达到 92.42%, 优于未经剪枝的 ARC-BC 分类器及 KNN、Bayes 分类器; 在分类耗时方面, 该方法与未经剪枝的 ARC-BC 分类器持平, 表明该方法引入的相似度与聚合值的计算开销在可接受的范围内.

关键词: 文本分类; 关联规则; 类别相似度; 聚合

中图分类号: TP391 **文献标志码:** A **文章编号:** 0253-987X(2012)12-0006-06

Associative Rule-Based Text Categorization Method
Using Category Similarity

TIAN Feng^{1,2}, GUI Xiaolin^{1,2}, YANG Pan^{1,2}, WANG Gang^{1,2,3}, GUO Yuelong^{1,2}
(1. School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China;
2. Shaanxi Province Key Laboratory of Computer Network, Xi'an Jiaotong University, Xi'an 710049, China;
3. School of Information, Xi'an University of Finance and Economics, Xi'an 710100, China)

Abstract: Conventional association rule-based categorization methods have bottleneck in improving classifier's accuracy, since these methods only consider the rule confidence degree and use the pruning technique. A novel method to solve this problem is proposed, and is called associative rule-based classifier aggregating with category similarity (AACS). The method adopts the modified chi-square statistical technique to extract feature terms from each category, and employs the CR-tree to store classification rules. Algorithms to construct and to match CR-tree are proposed. Inner-product is used to calculate the similarity between the category sub vector of the text and the category feature vector, and then is aggregated with the rules' confidence degree to serve as the foundation of text categorization. Experimental results show that the method presented achieves a micro-average value of categorization 92.42% with extracting only 30 feature terms, which is better than the results of AWOPR, KNN, and Bayes classifiers. And the time complexity of the method is the same as that of AWOPR, indicating that the cost to calculate both the similarity and the aggregation is acceptable.

Keywords: text categorization; association rule; category similarity; aggregation

随着 Internet 在线文本资源的爆炸式增长,如何对这些信息进行分类组织、获取有价值的内容,是目前研究的热点问题.海量文本信息处理需要特别考虑计算及存储资源的消耗,传统的基于统计理论的分类方法,例如 KNN、Bayes、Nnet、SVM、LLSF 等,必须提取大量的特征项才可获得较好的分类效果,为了能够在低维特征空间处理文本信息, Liu 等^[1]提出关联分类方法 CBA,该方法易于解释,取得了比决策树分类算法 C4.5 更好的分类效果.研究者针对 CBA 的不足提出各种改进的分类方法,典型的有 ARC-BC^[2]和 CMAR^[3],其中 ARC-BC 是目前分类效果较好的关联规则分类方法.为了提高关联规则的分类性能,研究者又提出了 TRARC^[4]、WARC^[5]、FWARC^[6]、ISARC^[7] 等分类方法. TRARC 对关联规则挖掘过程中的特征项加入了词频属性,并采用分类规则树作为关联分类规则的存储结构,从而在保证分类精度的前提下加快了分类速度. WARC 则利用对规则加权的方法来改善分类器的稳定性. FWARC 在 ARC-BC 的基础上进行了改进,利用特征项权重定义了规则和文本的匹配度,以此作为关联分类器的分类标准,提高了分类器的分类能力. ISARC 利用特征项权重定义了 k 项集重要度,通过挖掘重要项集来产生关联规则,并考虑提升度对待分类文本的影响,提高了分类的准确率.为了减少挖掘出的无关规则的数量, Baralis 等^[8]提出了一种特征项修剪方法,使用兴趣度对特征项进行度量,一旦发现兴趣度低于阈值的特征项则将其去除,从而大大减少了关联规则挖掘的时间.

现有的关联分类方法主要以规则的置信度作

为文本分类的依据,为了减少规则匹配所需的时间,使用了规则修剪技术对规则进行修剪,但这会导致分类精度的下降.为此,本文使用规则置信度和类别相似度的聚合值作为文本分类的依据,对关联规则集不进行修剪,为保证分类效率,使用 CR-tree^[3]存储分类规则,并给出了 CR-tree 的构建与匹配算法.

1 基于类别相似度聚合的关联文本分类方法

为了保证文本分类的精度和效率,本文提出了基于类别相似度聚合的关联文本分类方法(AACS, associative rule-based classifier aggregating with category similarity),文本分类过程如图 1 所示.

AACS 的算法步骤如下.

- 步骤 1 使用修改的 χ^2 统计方法获得各个类别的标志向量,各类标志向量的特征词集合构成关联规则挖掘中的项.
- 步骤 2 对训练文本进行向量化表示.
- 步骤 3 使用 Apriori 算法对向量化表示的训练文本集进行关联规则挖掘,获得规则集.
- 步骤 4 使用 CR-tree 构建算法对规则集进行处理,得到 CR-tree.
- 步骤 5 对待分类文本进行向量化表示.
- 步骤 6 采用 CR-tree 匹配算法获得该文本所匹配的所有规则,对每一条规则,计算规则所指向类别与该文本的相似度,然后将规则置信度和类别相似度使用聚合函数求得其聚合值,将聚合值按类别求和,将文本判定为聚合值之和最大的类别.

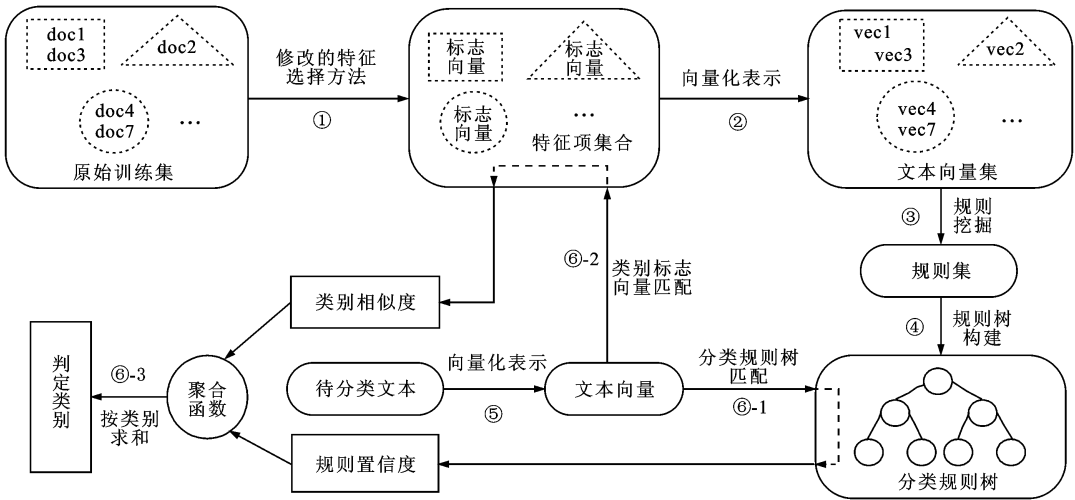


图 1 AACS 的总体流程

1.1 修改的文本特征选择方法

向量空间模型广泛用于文本表示,该模型将文本 d 映射为一个特征向量 $V(d) = (t_1, w_1(d); t_2, w_2(d); \dots; t_n, w_n(d))$, 其中 $t_i (i=1, 2, \dots, n)$ 为一列互不相同的词条项, $w_i(d)$ 为 t_i 在 d 中的权值. 由于该模型需要文本集中的所有词条构成特征空间, 当文本集规模较大时会出现特征空间维数过高的问题, 因此需要对文本特征进行选择, 使用降维后的特征子集进行文本分类.

本文为了获得测试文本与关联规则判定类别的相似度, 需要分别确定各类别的特征词集合, 为便于描述, 引入如下定义.

定义 1 类别 C_i 的标志向量 $F(C_i)$ 表示为

$$F(C_i) = (t_1(C_i), w_1(C_i); t_2(C_i), w_2(C_i); \dots; t_n(C_i), w_n(C_i))$$

式中: $t_k(C_i) (k=1, 2, \dots, n)$ 为类别 C_i 的第 k 个特征词; $w_k(C_i)$ 为 $t_k(C_i)$ 的权重.

定义 2 给定文本 $d_i \in D$, 其特征向量为

$$V(d_i) = \{ \{ (t_{11}, f_{11}), (t_{12}, f_{12}), \dots, (t_{1n}, f_{1n}) \}, \{ (t_{21}, f_{21}), (t_{22}, f_{22}), \dots, (t_{2n}, f_{2n}) \}, \dots, \{ (t_{m1}, f_{m1}), (t_{m2}, f_{m2}), \dots, (t_{mn}, f_{mn}) \} \}$$

则文本 d_i 属于类别 C_k 的分量 $S(d_i, C_k)$ 为

$$S(d_i, C_k) = \{ (t_{k1}, f_{k1}), (t_{k2}, f_{k2}), \dots, (t_{kn}, f_{kn}) \}$$

式中: t_{kn} 为类别 C_k 中的第 n 个特征项; f_{kn} 为文本 d_i 中 t_{kn} 的出现频率. 为防止某些高频词影响分类效果, 经过参数选择, 本文将特征项的最高词频限定为 6.

由于目前常用的特征选择方法, 如文档频率 (DF)、信息增益 (IG)、互信息 (MI)、 χ^2 统计 (CHI) 等^[9], 均是全局特征选择方法, 无法获得每个类别的特征项, 因此本文基于 χ^2 统计方法, 获得特征词在每个类别中的 χ^2 值, 然后根据类别对特征词分组, 对组内的特征词根据其 χ^2 值降序排列, 取前 n 个特征词作为该类的标志向量中的特征词 $t_k(C_i)$, 与该特征词对应的权重则由如下公式确定

$$w_k(C_i) = \frac{\chi^2(t_k(C_i), C_i)}{\sum_{j=1}^n \chi^2(t_j(C_i), C_i)} \quad (1)$$

1.2 规则树的构建与匹配算法

在关联规则挖掘算法中, 以 Apriori 算法^[10] 和 FP 增长算法^[11] 最具代表性. 由于通过关联规则挖掘生成的规则集规模可能非常大, 搜索与文本匹配的规则非常耗时, 常用的关联分类方法是对规则进

行修剪, 但这样会造成分类的准确性降低, 因此, 本文采用 CR-tree 来存储分类规则, 在保证分类准确性的情况下, 大大提高规则匹配的速度.

在构建 CR-tree 之前, 需要对获得的频繁项集计算其置信度, 再添加其类别信息, 由此构成规则集. 接下来需要对每一条规则前件中的项按其在规则集中的出现频率进行降序排序, 由处理过的规则集构建 CR-tree. 算法描述如下:

输入 规则集 R , 初始根节点 $root$

输出 规则集的树形表示, 返回其根节点

1. For each 规则集 R 中的每个规则 r
2. 获取规则 r 的类别和置信度信息
3. 当前节点 $p \leftarrow root$
4. For each 规则 r 中的每个项 t
5. If 项 t 是规则 r 的最后一项
6. If 当前节点 p 的子节点 c 包含项 t
7. 添加 r 的类别和置信度到子节点 c
8. Else
9. 创建树节点 n , 添加项 t 与规则 r 的类别和置信度, 并将 n 作为 p 的子节点
10. End If
11. Else
12. If 当前节点 p 的子节点 c 包含项 t
13. $p \leftarrow c$
14. Else
15. 创建树节点 m , 添加项 t , 将 m 作为 p 的子节点, 并执行 $p \leftarrow m$
16. End If
17. End If
18. End For
19. End For

通过对训练文本集进行关联规则挖掘, 构成 CR-tree 后, 就可以利用该 CR-tree 对测试文本进行规则匹配. 在对测试文本进行分词、词频统计后, 需要根据训练阶段提取的特征词对文本向量进行处理, 获得该文本的项集表示, 即文本中出现了哪些特征项, 并以规则集中各项的出现频率降序排列, 从而进行 CR-tree 的匹配, 进而获得该文本所匹配的规则. 匹配算法描述如下:

输入 CR-tree 树节点 p , 文本项集 T

输出 文本项集 T 所匹配的规则集 R

MatchRule(p, T)

1. 创建规则集 R
2. If 文本项集 T 仅包含一个项 t

```

3.   If 节点  $p$  的子节点  $c$  包含项  $t$ 
4.       将  $c$  对应的规则添加到  $R$ 
5.   End If
6. Else
7.   For each 文本项集  $T$  中的每个项  $t$ 
8.       If 节点  $p$  的子节点  $c$  包含项  $t$ 
9.           将  $c$  对应的规则添加到  $R$ 
10.          创建临时树节点  $n$ , 执行  $n \leftarrow c$ 
11.          创建文本项集  $M$ , 执行  $M \leftarrow T - t$ 
12.          添加 MatchRule( $n, M$ ) 所获得的规则到  $R$ 
13.      End If
14.  End For
15. End If
16. Return  $R$ 

```

该算法对待分类文本的项集进行处理,其中文本项集 T 已经过排序,其各项根据规则集中出现的频率降序排列。

MatchRule 是一个递归函数,算法 2~5 行是终止条件,若当前项集 T 仅包含一个项,且该项与树节点 p 的子节点匹配,则返回匹配子节点所包含的规则(规则列表);否则就对当前项集 T 中的每一项 t 执行 8~12 行的操作,其中第 12 行的 n 和 M 分别为已匹配树节点和从 T 中去除项 t 后的项集,递归调用 MatchRule 函数,直至获得所有匹配规则。

1.3 类别相似度与规则置信度的聚合方法

通过 CR-tree 匹配可以获得测试文本所匹配的所有规则,常用的根据关联分类规则判定文本所属类别的方法包括两种:将文本判定为置信度最大的规则所指向的类别;将文本所匹配的规则按类别分组,对各个类别中的规则置信度求和,把文本判定为置信度之和最大的类别。

经典的关联规则分类方法仅通过规则的置信度信息来对文本的类别进行判定,忽略了其他因素的影响。文献[6]提出了基于特征权重改进的关联分类方法,获得了比 ARC-BC 方法更好的分类效果。

本文采用向量内积来计算文本的类别分量 $S(d_i, C_k)$ 与类别标志向量 $F(C_k)$ 的相似度 γ , 即

$$\gamma(S(d_i, C_k), F(C_k)) = \sum_{j=1}^n w_j(C_k) f_{kj} \quad (2)$$

在文本分类的过程中,测试文本 d_i 经过 CR-tree 匹配可以获得多条匹配规则,记 RS_i^k 表示文本 d_i 所匹配的指向类别 C_k ($k=1, 2, \dots, m$) 的所有规则的集合,假定有一条规则 $* \rightarrow C_k \in RS_i^k$, 其置信度

为 μ , 文本 d_i 与类别 C_k 的标志向量 $F(C_k)$ 的相似度为 γ , 由于规则置信度和类别相似度均是从单一角度对类别进行判定,两者均可能受到噪声数据的影响,产生错误的判断,因此需要对这两个指标进行聚合,得到统一的判定指标,本文给出的聚合函数为

$$\phi(\mu, \gamma) = \mu e^\gamma \quad (3)$$

该聚合函数使用置信度 μ 与 e 的 γ 次方的乘积作为聚合值,避免了单一角度分类方式的片面性。由于 $y=\mu$ 与 $y=e^\gamma$ 均为单调增函数,这两个函数的叠加可以增强分类过程中的抗干扰能力,只有当 μ 与 e^γ 两个指标均为较低值,即无论从规则的置信度还是文本与类别标志向量的相似度考虑,文本 d_i 属于类别 C_k 的可能性都较低时,聚合值才会较低,否则聚合值会取定一个中和的结果,屏蔽噪声数据产生的影响。

对属于 RS_i^k 的每一条规则 r_s^k ($s=1, 2, \dots, |RS_i^k|$), 计算其置信度 μ_s^k 与类别相似度 λ_k 的聚合值 ϕ_s^k , 然后对聚合值求和,得到

$$\phi^k = \sum_s \phi_s^k = \sum_s \mu_s^k e^{\lambda_k} \quad (4)$$

最后将文本判定为聚合值总和最大的类别,即

$$v = \arg \max_k (\phi^k) \quad (5)$$

文本 d_i 最终被判定属于类别 C_v , 这种方法借鉴了多方表决的策略,即指向某一类别的聚合值总和越大,则文本属于该类别的概率就越大。

2 实验与结果分析

我们在 Intel Core2 Quad Q8300, 2 GB 内存计算机上实现了本文的训练与分类算法。实验所需的中文数据集(D1)为课题组从新浪网上抓取的网页,分为地产、教育、军事、旅游、女性、汽车、体育、音乐、游戏 9 类,其中训练集每个类包含 200 篇文档,测试集总共包含 2 343 篇文档。英文数据集(D2)为 Reuters-21578,根据 topic 进行分类,选出文档数最多的 10 个类,其中训练集每个类包含 100 篇文档,测试集总共包含 842 篇文档,采用开放测试的方法进行实验。

参照文献[12]的评价标准,准确率 π_i 指一个文档被正确分类到类别 C_i 的概率,召回率 ρ_i 指属于类别 C_i 的文档被分到该类的概率,宏平均准确率 π^M 为各类别准确率 π_i 的均值,宏平均召回率 ρ^M 为各类别召回率 ρ_i 的均值;微平均准确率 π^μ 为分类器对数据集中每一个文档正确分类的概率,微平均召回率 ρ^μ 为分类器对数据集中每一类别的文档分

到各自所属类别的概率, π^M 与 ρ^M 合并为度量宏平均 F_1^M

$$F_1^M = 2\pi^M\rho^M/(\pi^M + \rho^M)$$

(6)

π^μ 与 ρ^μ 合并为度量微平均 F_1^μ

$$F_1^\mu = 2\pi^\mu\rho^\mu/(\pi^\mu + \rho^\mu)$$

(7)

式(6)、(7)即为本文的主要评价标准.

2.1 实验参数设定

由于本文使用了 CR-tree 来存储规则,使得文本与规则集的匹配速度大大提高,因此我们使用未剪枝的 ARC-BC(ARC-BC without pruning rules, 简称 AWOPR)分类方法、KNN、Bayes 与本文的类别相似度聚合方法进行比较. 在实验中,对 AWOPR、KNN、Bayes 方法使用 χ^2 统计方法提取特征词,对本文方法使用修改的 χ^2 统计方法提取每个类别的特征向量,经过参数选择,对 KNN 和 Bayes 方法选择 500 个特征词,在数据集 D1 上对 AWOPR 和 AACS 方法选择 30 个特征词,在数据集 D2 上对 AWOPR 和 AACS 方法选择 50 个特征词,设定支持度阈值为 0.1;对 KNN 分类, k 表示距离最近的 k 个数据本样数量,取 k 值为 15,文本向量间的相似度使用余弦相似度公式计算得到;在 Bayes 分类中使用多项式模型,分别在中文数据集和英文数据集上进行测试.

2.2 AACS 与其他分类方法的比较

通过与其他文本分类方法的比较,由表 1、表 2 可得出以下结论:

(1)AACS 方法在两个数据集上的分类准确性均优于 AWOPR、KNN 和 Bayes 方法.

表 1 不同分类方法在中文数据集上的比较

参数	AACS	AWOPR	KNN	Bayes
训练时间/ms	28 201	27 755	28 566	28 809
分类时间/ms	19 134	18 591	32 172	18 381
$F_1^M/\%$	91.88	89.27	85.51	86.60
$F_1^\mu/\%$	92.42	89.89	85.28	86.38

表 2 不同分类方法在英文数据集上的比较

参数	AACS	AWOPR	KNN	Bayes
训练时间/ms	5 751	5 629	4 936	4 941
分类时间/ms	2 912	2 881	4 928	2 652
$F_1^M/\%$	86.90	84.13	85.27	84.15
$F_1^\mu/\%$	86.32	83.92	85.04	84.32

(2)AACS 方法在分类阶段需要将规则置信度与类别相似度进行聚合,因此其分类所需时间要略高于直接使用置信度分类的 AWOPR 方法.

(3)由于在训练阶段对 Bayes 分类方法的条件概率值进行了存储,因此 Bayes 分类方法的分类时间与 AACS 和 AWOPR 持平.

2.3 特征词数量对关联规则分类方法的影响

为了研究关联规则分类方法在不同特征词数量下的分类效果,本文以 10 为步长,在[30,90]特征词数量区间上分别测试 AACS 和 AWOPR 方法的分类准确率,结果见图 2、图 3. 由图 2、图 3 可以得出以下结论.

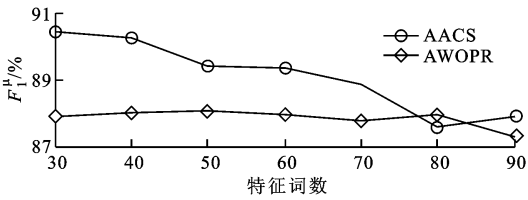


图 2 D1 上特征词数量对微平均值的影响

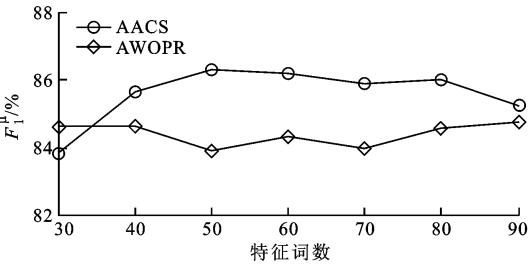


图 3 D2 上特征词数量对微平均值的影响

(1)对于中文数据集(D1),随着特征词数量的增加,类别相似度聚合方法(AACS)的 F_1^μ 值呈下降趋势,这是由于在文本分类中,每个类别的有效特征词并不多,过多的特征词会干扰类别特征向量的相似度聚合效果,进而影响分类的准确率.

(2)在中文数据集(D1)上,当特征词数取值为 30 时,AACS 方法会取得最好的分类效果,且在大多数情况下, AACS 方法的分类效果均优于 AWOPR 方法.

(3)在英文数据集(D2)上,当特征词数从 30 增至 50 时,AACS 的 F_1^μ 值呈上升趋势,而特征词数从 50 增至 90 时,AACS 的 F_1^μ 值则开始下降,这与结论 1 的推测相符;AWOPR 方法的 F_1^μ 值则在一定范围内(84%~85%)波动.

(4)AACS 方法在英文数据集(D2)上的最优特征词数为 50,且在特征词数区间[40,90]上,AACS 方法的分类效果均优于 AWOPR 方法.

图 4~7 显示了特征词数的变化对中、英文数据集的训练时间和分类时间的影响. 可以看出,在中文数据集上,特征词数的增加对训练时间影响不大,而对于英文数据集,分类器的训练时间随特征词数的增加而增加,这是由于在英文数据集的训练阶段,随着特征词数的增加,挖掘到了更多的分类规则,进而需要消耗更多的时间来构建 CR-tree. 特征词数在一定范围内的变化对分类时间影响不大,这也与关联规则分类的基本特性相符,其中,由于 AACS 方法需要进行置信度与相似度的聚合操作,因此会比 AWOPR 方法消耗更多的时间,但从图 5、图 7 可以看出,这种影响几乎可以忽略.

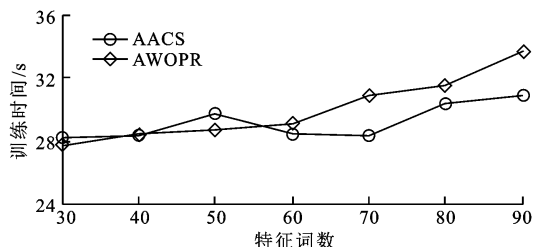


图 4 D1 上特征词数量对训练时间的影响

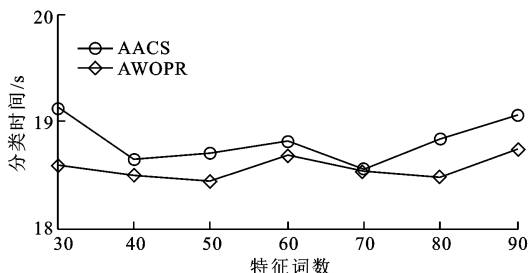


图 5 D1 上特征词数量对分类时间的影响

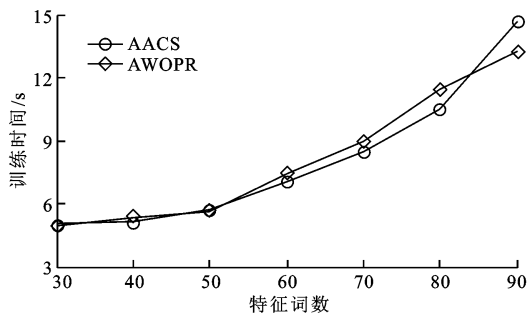


图 6 D2 上特征词数量对训练时间的影响

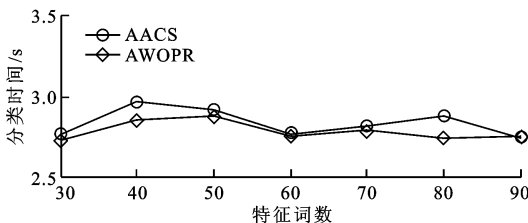


图 7 D2 上特征词数量对分类时间的影响

3 结束语

海量文本信息处理需要高效率、易解释的文本分类方法,经典的基于统计理论的分类方法需要提取较多的特征项才能够获得理想的分类效果,其系统开销太大,而现有的关联规则分类方法在分类时仅考虑规则的置信度,这将会忽略影响分类效果的其他因素. 本文在进行特征选择时,对每个类分别提取其特征向量,并由此提出了基于类别相似度聚合的关联规则文本分类方法,提高了文本分类的精度. 后续的工作主要包括以下 3 个方面. ①文本特征选择方法的改进. 本文对 χ^2 统计方法进行了修改,但该方法在其他场合不具有可用性,因此,研究更加通用的文本特征选择方法十分必要. ②进一步研究高效的分类规则树的构建与匹配算法. ③规则置信度与类别相似度的聚合方法对提高分类器的精度有重要作用,需要深入研究其作用机理,以期获得更优的分类器.

参考文献:

- [1] LIU Bing, HSU W, MA Yiming. Integrating classification and association rule mining [C]//Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 1998: 80-86.
- [2] ZAÏANE O R, ANTONIE M L. Classifying text documents by associating terms with text categories [C]//Proceedings of the 13th Australasian Database Conference. New York, USA: ACM, 2002: 215-222.
- [3] LI Wenmin, HAN Jiawei, PEI Jian. CMAR: Accurate and efficient classification based on multiple classification rules [C]//Proceedings of the 2001 IEEE International Conference on Data Mining. Piscataway, NJ, USA: IEEE, 2001: 369-376.
- [4] 陈晓云,陈伟,王雷,等. 基于分类规则树的频繁模式文本分类[J]. 软件学报,2006,17(5):1017-1025.
CHEN Xiaoyun, CHEN Yi, WANG Lei, et al. Text categorization based on classification rules tree by frequent patterns [J]. Journal of Software, 2006, 17(5): 1017-1025.
- [5] 陈晓云,胡运发. 基于自适应加权的文本关联分类[J]. 小型微型计算机系统,2007,28(1):116-121.
CHEN Xiaoyun, HU Yunfa. Text association categorization based on self-adaptive weighting [J]. Journal of Chinese Computer Systems, 2007, 28(1):116-121.
- [6] 商炳章,白清源. 基于特征项权重改进的关联文本分

- 类[J]. 计算机研究与发展, 2008, 45(S0): 252-256.
- SHANG Bingzhang, BAI Qingyuan. Improved association text classification based on feature weight [J]. Journal of Computer Research and Development, 2008, 45(S0): 252-256.
- [7] 蔡金凤, 白清源. 挖掘重要项集的关联文本分类[J]. 南京大学学报, 2011, 47(5): 544-550.
- CAI Jinfeng, BAI Qingyuan. Association text classification of mining ItemSet significance [J]. Journal of Nanjing University, 2011, 47(5): 544-550.
- [8] BARALIS E, GARZA P. I-prune: item selection for associative classification [J]. International Journal of Intelligent Systems, 2012, 27(3): 279-299.
- [9] YANG Yiming, PEDERSON J O. A comparative study on feature selection in text categorization [C]// Proceedings of the 14th International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann, 1997: 412-420.
- [10] AGRAWAL R, SRIKANT R. Fast algorithms for mining association rules [C]// Proceedings of the 20th VLDB Conference. San Francisco, CA, USA: Morgan Kaufmann, 1994: 487-499.
- [11] HAN Jiawei, PEI Jian, YIN Yiwen, et al. Mining frequent patterns without candidate generation: a frequent-pattern tree approach [J]. Data Mining and Knowledge Discovery, 2004, 8(1): 53-87.
- [12] SEBASTIANI F. Machine learning in automated text categorization [J]. ACM Computing Surveys, 2002, 34(1): 1-47.

[本刊相关文献链接]

- 杨攀, 桂小林, 田丰, 等. 一种高效的用于话题检测的关键词元聚类方法. 2012, 46(10): 24-28.
- 霍战鹏, 魏正英, 张梦, 等. 手机短信远程控制灌溉系统. 2012, 46(10): 36-41.
- 豆增发, 高琳. 利用膜粒子群优化和信息熵的医学文本特征选择. 2012, 46(4): 45-51.
- 薛峰, 周亚东, 高峰, 等. 一种突发性热点话题在线发现与跟踪方法. 2011, 45(12): 64-69.
- 豆增发, 高琳. 应用粒子群优化-条件随机域的文本生物实体识别. 2010, 44(12): 38-42.
- 赵煜, 蔡皖东, 樊娜, 等. 采用并行遗传算法的文本分割研究. 2009, 43(12): 40-44.
- 姜庆民, 吴宁, 刘伟华. 面向入侵检测系统的模式匹配算法研究. 2009, 43(2): 58-62.

(编辑 武红江)