

文章编号: 1003-0077(2018)08-0091-12

基于包含度和频繁模式的文本特征选择方法

池云仙^{1,2}, 赵书良², 李仁杰¹

- (1. 河北师范大学 资源与环境科学学院, 河北 石家庄 050024;
2. 河北师范大学 数学与信息科学学院, 河北 石家庄 050024)

摘要: 大数据时代, 文本数据量的爆炸式增长使得特征选择成为文本挖掘领域最关键的 task 之一。文档中的词语和模式规模庞大, 故需保证所挖掘特征的质量充满挑战。“基于模式”特征选择方法具有传统“基于词语”方法所没有的优越特性, 可以进行有效地信息去噪, 提升文本挖掘性能。该文提出基于包含度和频繁模式的文本特征选择方法: 首先, 定义基于包含度的相似性度量原理; 然后, 提出基于包含度的冗余文本频繁模式过滤方法。基于包含度度量文本频繁模式间相似性, 以此去除了子模式及相似度较高的交叉模式。再通过冗余模式去噪, 提升文本频繁模式挖掘性能; 提出基于关联度的文本特征选择方法。以经过过滤处理后的非冗余文本频繁模式为基础, 进行文本特征选择, 并利用词语与文档的关联度进行词语类别划分及权重分配。使所选特征与文档关联度更加清晰, 分类效果更好。通过在数据集 Reuters-21578 上的实验得知, 基于包含度和频繁模式的文本特征选择算法性能, 优于当前普遍应用的传统文本特征选择方法和新的特征选择及特征抽取方法。

关键词: 大数据; 文本挖掘; 文本频繁模式; 包含度; 文本特征选择

中图分类号: TP391

文献标识码: A

Text Feature Selection Based on Inclusion Degree and Frequent Pattern

CHI Yunxian^{1,2}, ZHAO Shuliang², LI Renjie¹

- (1. College of Resources and Environment Science, Hebei Normal University, Shijiazhuang, Hebei 050024, China;
2. College of Mathematic and Information Science, Hebei Normal University, Shijiazhuang, Hebei 050024, China)

Abstract: In big data era, the growth rate of text information is too fast to deal with. Finding text features is one of the key issues in field of text mining. It is a great challenge to ensure the quality of features, which are mined from texts, due to the presence of large-scale words and patterns. Pattern-based methods have many superior characters while term-based methods have not. Pattern-based methods can remove noises efficaciously and promote performance of text mining. Algorithm Text Feature Selection Based on Inclusion Degree and Frequent Pattern (TFSIDFP) is proposed. First of all, standard of similarity measure for frequent patterns based on inclusion degree is defined. Secondly, algorithm Filtration of Redundancy for Frequent Patterns based on Inclusion Degree Theory (FRFPIDT) is put forward, algorithm FRFPIDT measures similarity of frequent patterns based on inclusion degree, and removes subpatterns and cross-patterns with high similarity degree. Performance of frequent patterns mining is increased by cutting out redundancy patterns. At last, feature weighting model is put forward. In this model, features are selected based on non-redundant frequent patterns that are disposed through algorithm TFSIDFP. Correlation between features and documents is taken into account in feature weighting, thus correlation degree between them is higher and effect of classification is better. Experimental results on data sets from Reuters-21578 indicate algorithm TFSIDFP is superior to the widely used feature selection and feature extraction methods.

Key words: big data; text mining; text frequent pattern; inclusion degree; text feature selection

收稿日期: 2017-09-19 定稿日期: 2017-11-13

基金项目: 国家自然科学基金(71271067); 国家社科基金(13&ZD091); 河北省高等学校科学技术研究项目(QN2014196)

0 引言

文本数据维度在大数据时代下呈迅猛增长趋势。在影响数据挖掘性能的各因素中,特征选择成为其中至关重要的环节之一。特征选择通过提取特征子集来有效缩小高维特征空间,可有效提高数据挖掘性能,故各领域学者均致力于特征选择方法的研究。Zhao等基于特征选择算法“保留样本相似性”的共同点,提出一种通用的相似性保留特征选择框架^[1]。Zhuang等提出基于主题模型进行特征选择以提高模型预测性能的ssLDA模型^[2]。Song等提出基于图论聚类模型的高维数据子类划分与关联特征子集速选方法^[3]。Li等提出基于文档与特征关联性的关联特征选择方法^[4]。张延祥等针对数据不平衡问题提出基于类别区分力的文本特征选择方法DA^[5]。

“基于词语”的特征选择方法,因其被“同义词、一词多义及噪声词语”等问题所困扰,特征提取效率大打折扣。相比之下,“基于模式”方法凭借“保留词语间关联性”的优势,很好地克服了以上问题。它可以在从“数据”中高效挖掘“知识”的同时,有效减轻“数据爆炸”问题带给大数据时代的困扰。作为数据挖掘领域的重点与热点,“基于模式”研究已扩展至诸多领域。Gao等提出基于主题最大匹配模式的文档过滤模型MPBTM,依托用户所需信息与模式间的关联度去除不相关文档^[6]。Zhao等提出基于未确定数据库的潜在频繁序列模式挖掘方法^[7]。Kessl提出基于概率性平衡负载的并行频繁序列模式挖掘方法^[8]。Pumjun等提出基于动态数据库调整支持度阈值的多级关联规则挖掘模型MLUPCS^[9]。Zhang等提出基于马尔科夫性质的DNA序列模式挖掘模型^[10]。Turdi等结合维吾尔文间关联规则进行频繁模式挖掘,进而实现语义串快速抽取^[11]。

“通过最优化特征排序标准来进行特征排序与选择”的思想是大多特征选择方法的共同特点,但由此产生的“相关特征排序相近”的特征冗余问题严重影响文本挖掘效率。因此,将冗余特征进行去噪处理将明显提升文本挖掘性能。Ding等提出基于贪心算法的连续特征选择冗余最小化方法mRMR^[12]。Wang等提出基于全局冗余最小化的整体局部差异化特征选择方法^[13]。

事物间的“差异性”和“不确定性”是普遍存在

的,而这种“相似程度”和“不定关系”通常用包含度原理进行描述。Gong等提出基于模糊集包含度的非参数统计模型^[14]。Ma等在模糊粗糙集的基础上提出包含度与相似度计算的通用模型^[15]。Liu等提出基于最大包含度原理的样本决策表分类方法^[16]。李阳等基于知识图谱提出一种通用的实体相似性度量方法^[17]。

为扩充基于频繁模式的文本特征选择方法在文本挖掘领域的应用,提出基于包含度和频繁模式的文本特征选择方法TFSIDFP。TFSIDFP方法利用频繁模式词语间的关联,有效避免了“基于词语”方法的噪声问题影响;同时,利用包含度原理可以对文本中的冗余频繁模式进行过滤,有效提高了模式提取效率及特征选择性能。

后续内容为:第一节介绍基于包含度和频繁模式进行文本特征选择的模型框架;第二节详细介绍基于包含度和频繁模式的文本特征选择方法;第三节为实验;第四节为全文总结。

1 模型框架

基于包含度和频繁模式进行文本特征选择,旨在基于包含度原理过滤掉文本中的冗余频繁模式,并在经过优化处理后的非冗余文本频繁模式基础上进行文本特征选择。该框架主要分为以下几部分:

(1) 文本频繁模式挖掘:利用FP-Growth算法挖掘文本中所有频繁模式;

(2) 冗余文本频繁模式过滤:基于包含度原理,度量文本频繁模式间的相似性,将子模式和相似度高与阈值的交叉模式进行去冗余操作;

(3) 非冗余文本频繁模式特征选择:基于过滤后的非冗余频繁模式,进行文本特征选择,并利用特征与文档的关联度进行词语类别划分及权重分配;

(4) 文本分类:利用所选择的特征词语进行文本分类。

基于包含度和频繁模式的文本特征选择流程图如图1所示。

2 基于包含度和频繁模式的文本特征选择方法

文本频繁模式挖掘过程中会不可避免地产生大量冗余模式。例如,较长文本频繁模式所蕴含的子模式集合以及与该文本频繁模式相似的交叉模式集

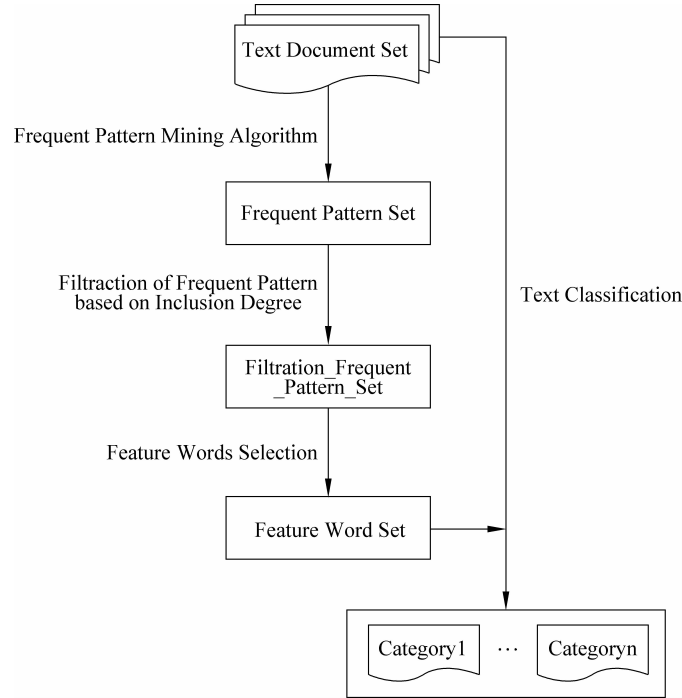


图1 基于包含度和频繁模式的文本特征选择流程图

合,对于同一类别主题而言,往往是冗余的。冗余模式会严重制约文本挖掘性能。因此,为提高文本分类运行效率,本文提出基于包含度和频繁模式的文本特征选择算法 TFSIDFP。首先,提出基于包含度的相似性度量原理;然后,提出基于包含度的冗余文本频繁模式过滤方法;最后,提出基于关联度的文本特征选择方法。

2.1 基于包含度的相似性度量原理

“包含度”概念源于真实世界中信息的“不完整性”。这种“不完整性”无法用经典逻辑问题的两个绝对标准(“相等”和“不相等”)度量,由此便衍生出包含度理论。

冗余模式产生问题在挖掘文本频繁模式的过程中无法规避。这不仅影响模式挖掘效率,还会间接制约文本特征选择性能。基于包含度理论对待度量的文本频繁模式进行评估,先过滤掉相似度超过预定阈值的冗余模式,可有效地缩减文本频繁模式集合的规模,进而提高文本频繁模式的挖掘性能。首先,定义“包含度”和“相似度”的概念;然后,提出并证明基于包含度的相似性度量原理的性质。

定义1 包含度(Inclusion Degree) 设论域 Dom_Dis 有两个子集 Dom_Sub_A 和 Dom_Sub_B , 即 $Dom_Sub_A, Dom_Sub_B \subseteq Dom_Dis$ 。若存在 $ID(Dom_Sub_B/Dom_Sub_A)$ 满足下述三个性质:

(I) 非负性: $0 \leq ID(Dom_Sub_B/Dom_Sub_A) \leq 1$;

(II) 规范性: 当 $Dom_Sub_A \subseteq Dom_Sub_B$ 时, $ID(Dom_Sub_B/Dom_Sub_A) = 1$;

(III) 传递性: 当 $Dom_Sub_A \subseteq Dom_Sub_B \subseteq Dom_Sub_C$ 时, 有 $ID(Dom_Sub_A/Dom_Sub_C) \leq ID(Dom_Sub_A/Dom_Sub_B)$ 。

则称 $ID(Dom_Sub_B/Dom_Sub_A)$ 为 Dom_Sub_B 包含 Dom_Sub_A (或 Dom_Sub_A 包含于 Dom_Sub_B) 的包含度。

定义2 相似度(Similarity Degree) 设论域 Dom_Dis 有两个子集 Dom_Sub_A 和 Dom_Sub_B , 即 $Dom_Sub_A, Dom_Sub_B \subseteq Dom_Dis$ 。若存在 $SD(Dom_Sub_A, Dom_Sub_B)$ 满足下述四个性质:

(I) 非负性: $0 \leq SD(Dom_Sub_A, Dom_Sub_B) \leq 1$;

(II) 自反性: $SD(Dom_Sub_A, Dom_Sub_A) = 1$;

(III) 对称性: $SD(Dom_Sub_A, Dom_Sub_B) = SD(Dom_Sub_B, Dom_Sub_A)$;

(IV) 传递性: 当 $Dom_Sub_A \subseteq Dom_Sub_B \subseteq Dom_Sub_C$ 时, 有 $SD(Dom_Sub_A, Dom_Sub_B) \geq SD(Dom_Sub_A, Dom_Sub_C)$ 。

则称 $SD(Dom_Sub_A, Dom_Sub_B)$ 为 Dom_Sub_A 和 Dom_Sub_B 之间的相似度。

性质1 设论域 Dom_Dis 有两个子集 $Dom_$

Sub_A 和 Dom_Sub_B , 即 $Dom_Sub_A, Dom_Sub_B \subseteq Dom_Dis$ 。那么, Dom_Sub_A 和 Dom_Sub_B 之间基于包含度的相似性度量公式如式(1)所示。

$$\begin{aligned} SD(Dom_Sub_A, Dom_Sub_B) &= ID\left(\frac{Dom_Sub_A \cap Dom_Sub_B}{Dom_Sub_A \cup Dom_Sub_B}\right) \\ &= \frac{Num(Dom_Sub_A \cap Dom_Sub_B)}{Num(Dom_Sub_A \cup Dom_Sub_B)} \quad (1) \end{aligned}$$

其中, $Num(Dom_Sub_A \cap Dom_Sub_B)$ 为集合 Dom_Sub_A 和 Dom_Sub_B 公共元素数目, $Num(Dom_Sub_A \cup Dom_Sub_B)$ 为集合 Dom_Sub_A 和 Dom_Sub_B 中互异元素总数。

证明:

(一) 相似性证明:

(I) 非负性:

$$0 \leq SD\left(\frac{Dom_Sub_A \cap Dom_Sub_B}{Dom_Sub_A \cup Dom_Sub_B}\right) \leq 1;$$

(II) 自反性:

$$\begin{aligned} SD(Dom_Sub_A, Dom_Sub_A) &= ID\left(\frac{Dom_Sub_A \cap Dom_Sub_A}{Dom_Sub_A \cup Dom_Sub_A}\right) \\ &= ID(Dom_Sub_A / Dom_Sub_A) = 1; \end{aligned}$$

(III) 对称性:

$$\begin{aligned} SD(Dom_Sub_A, Dom_Sub_B) &= ID\left(\frac{Dom_Sub_A \cap Dom_Sub_B}{Dom_Sub_A \cup Dom_Sub_B}\right) \\ &= SD(Dom_Sub_B, Dom_Sub_A); \end{aligned}$$

(IV) 传递性: 当 $Dom_Sub_A \subseteq Dom_Sub_B \subseteq Dom_Sub_C$ 时,

$$\begin{aligned} SD(Dom_Sub_A, Dom_Sub_B) &= ID\left(\frac{Dom_Sub_A \cap Dom_Sub_B}{Dom_Sub_A \cup Dom_Sub_B}\right) \\ &= ID(Dom_Sub_A / Dom_Sub_B), \\ SD(Dom_Sub_A, Dom_Sub_C) &= ID\left(\frac{Dom_Sub_A \cap Dom_Sub_C}{Dom_Sub_A \cup Dom_Sub_C}\right) \\ &= ID(Dom_Sub_A / Dom_Sub_C). \end{aligned}$$

由于 $ID\left(\frac{Dom_Sub_A}{Dom_Sub_B}\right) \geq ID\left(\frac{Dom_Sub_A}{Dom_Sub_C}\right)$, 则 $SD(Dom_Sub_A, Dom_Sub_B) \geq SD(Dom_Sub_A, Dom_Sub_C)$ 。

由此可知, $SD(Dom_Sub_A, Dom_Sub_B) = ID\left(\frac{Dom_Sub_A \cap Dom_Sub_B}{Dom_Sub_A \cup Dom_Sub_B}\right)$ 为相似度。

(二) 包含度证明:

$$\begin{aligned} (I) \text{ 非负性: } 0 &\leq ID\left(\frac{Dom_Sub_A \cap Dom_Sub_B}{Dom_Sub_A \cup Dom_Sub_B}\right) \\ &= \frac{Num(Dom_Sub_A \cap Dom_Sub_B)}{Num(Dom_Sub_A \cup Dom_Sub_B)} \leq 1; \end{aligned}$$

(II) 规范性: 由于 $(Dom_Sub_A \cap Dom_Sub_B) \subseteq (Dom_Sub_A \cup Dom_Sub_B)$, 则

$$ID\left(\frac{Dom_Sub_A \cap Dom_Sub_B}{Dom_Sub_A \cup Dom_Sub_B}\right) = 1。$$

(III) 传递性:

当 $(Dom_Sub_A \cap Dom_Sub_B) \subseteq (Dom_Sub_A \cup Dom_Sub_B) \subseteq Dom_Sub_C$,

那么, $Num((Dom_Sub_A \cap Dom_Sub_B) \cap (Dom_Sub_A \cup Dom_Sub_B)) = Num(Dom_Sub_A \cap Dom_Sub_B)$,

$$Num((Dom_Sub_A \cap Dom_Sub_B) \cap Dom_Sub_C) = Num(Dom_Sub_A \cap Dom_Sub_B)。$$

$$\text{因此, } ID(Dom_Sub_A \cap Dom_Sub_B) / Dom_Sub_C = \frac{Num(Dom_Sub_A \cap Dom_Sub_B)}{Num(Dom_Sub_C)},$$

$$\begin{aligned} ID\left(\frac{Dom_Sub_A \cap Dom_Sub_B}{Dom_Sub_A \cup Dom_Sub_B}\right) &= \frac{Num(Dom_Sub_A \cap Dom_Sub_B)}{Num(Dom_Sub_A \cup Dom_Sub_B)}。 \end{aligned}$$

又因为 $(Dom_Sub_A \cup Dom_Sub_B) \subseteq Dom_Sub_C$, 则 $Num(Dom_Sub_A \cup Dom_Sub_B) \leq Num(Dom_Sub_C)$, 所以 $ID((Dom_Sub_A \cap Dom_Sub_B) / Dom_Sub_C) \leq$

$$\begin{aligned} ID\left(\frac{Dom_Sub_A \cap Dom_Sub_B}{Dom_Sub_A \cup Dom_Sub_B}\right) &\text{ 即 } ID\left(\frac{Dom_Sub_A \cap Dom_Sub_B}{Dom_Sub_A \cup Dom_Sub_B}\right) \\ &= \frac{Num(Dom_Sub_A \cap Dom_Sub_B)}{Num(Dom_Sub_A \cup Dom_Sub_B)} \text{ 为包含度。} \end{aligned}$$

综上所述, 基于包含度的相似性度量公式为

$$\begin{aligned} SD(Dom_Sub_A, Dom_Sub_B) &= ID\left(\frac{Dom_Sub_A \cap Dom_Sub_B}{Dom_Sub_A \cup Dom_Sub_B}\right) \\ &= \frac{Num(Dom_Sub_A \cap Dom_Sub_B)}{Num(Dom_Sub_A \cup Dom_Sub_B)}。 \end{aligned}$$

证毕。

例如, 基于 FP-Growth 算法挖掘三个频繁模式 $X: \langle t_1, t_2, t_3, t_4, t_5, t_6, t_8, t_{13}, t_{14}, t_{15} \rangle, Y: \langle t_1, t_2, t_3, t_4, t_5, t_6, t_9, t_{13}, t_{14} \rangle, Z: \langle t_6, t_7, t_8, t_9, t_{10}, t_{11}, t_{12} \rangle$ 。采用相似度式(1)计算 X 与 Y 的相似度以及 X 与 Z 的相似度:

$$SD(X, Y) = \frac{Num(X \cap Y)}{Num(X \cup Y)} = \frac{8}{11};$$

$$SD(X, Z) = \frac{Num(X \cap Z)}{Num(X \cup Z)} = \frac{2}{15},$$

若相似度阈值预先设定为 $SD \geq 0.7$, 则将 X 和 Y 视为相似模式, X 和 Z 视为非相似模式。根据 2.2 节保留较长模式的原则, 在进行冗余频繁模式过滤操作时, 会将 Y 从模式集合中去除。

2.2 基于包含度的冗余文本频繁模式过滤方法

定义 3 频繁模式(Frequent Pattern)指频繁出现在数据集中的模式, 含频繁项集、子序列或子结构。

定义 4 文本频繁模式(Text Frequent Pattern)

若文档 Td 中某一词集 $WSet_i_Td = \{w_1, w_2, \dots, w_q\} \subseteq W$ 的支持度满足 $Support(WSet_i_Td) \geq MinSupport$, 则称 $WSet_i_Td$ 构成的模式为文本频繁模式, 记作 TFP 。其中 $MinSupport$ 为预定的最小支持度。

定义 5 文本频繁子模式(Text Frequent Sub-pattern)若两个文本频繁模式 TFP_i 和 TFP_j 对应词集 $WSet_TFP_i$ 和 $WSet_TFP_j$ 满足关系 $WSet_TFP_i \subseteq WSet_TFP_j$, 则称 TFP_i 为 TFP_j 的文本频繁子模式, 记为 $TFP_i \subseteq TFP_j$ 。

定义 6 文本频繁交叉模式(Text Frequent Cross Pattern)若两个文本频繁模式 TFP_i 和 TFP_j 对应词集 $WSet_TFP_i$ 和 $WSet_TFP_j$ 满足关系 $WSet_TFP_i \not\subseteq WSet_TFP_j \& WSet_TFP_j \not\subseteq WSet_TFP_i \& (WSet_TFP_i) \cap (WSet_TFP_j) \neq \Phi$, 那么 TFP_i 与 TFP_j 为文本频繁交叉模式, 记为 $TFP_i \not\subseteq TFP_j \& TFP_j \not\subseteq TFP_i \& TFP_i \cap TFP_j \neq \Phi$ 。

频繁模式挖掘过程中不可避免地会受到噪声问题的影响。较长的频繁模式往往包含比较短模式更多的有用信息, 有时甚至可以完全覆盖某些子模式, 因此在模式过滤中留下较长的频繁模式可保留更多与类别相关的信息, 对于类别划分更加有利。

设 $TFPSet = \{TFP_1, TFP_2, \dots, TFP_n\}$ 为文本频繁模式全集, 集合中的频繁模式按照模式长度进行降序排序。文本频繁模式过滤集合初始化为 $Filter_TFPSet = \Phi$ 。从集合 $TFPSet$ 中依次选取频繁模式与 $Filter_TFPSet$ 中的模式做比较。对于 $\forall TFP_i \in TFPSet$, TFSIDFP 算法进行冗余文本频繁模式过滤的过程如下:

(1) 对于 $\forall TFP_j \in Filter_TFPSet$:

① 若 TFP_i 为 TFP_j 的文本频繁子模式, 即 $TFP_i \subseteq TFP_j$, 则执行冗余模式过滤操作 $TFPSet$

$- TFP_i$;

② 若 TFP_i 和 TFP_j 为文本频繁交叉模式, 即 $TFP_i \not\subseteq TFP_j \& TFP_j \not\subseteq TFP_i \& TFP_i \cap TFP_j \neq \Phi$, 则计算其相似度 $SD(TFP_i, TFP_j)$, 若 $SD(TFP_i, TFP_j) \geq \theta$ (θ 为预定相似度阈值), 则执行冗余模式过滤操作 $TFPSet - TFP_i$, 同时归并支持度 $Support(TFP_j) = Support(TFP_j) + Support(TFP_i)$;

③ 否则, 执行文本频繁模式计数器增值操作 TFP_count_i++ 。

(2) 若 $TFP_count_i = |Filter_TFPSet|$, 表示 TFP_i 与 $Filter_TFPSet$ 中任意文本频繁模式 TFP_j 均不存在子模式或高相似度交叉模式关系, 则将 TFP_i 归入 $Filter_TFPSet$, 并从 $TFPSet$ 中去除。

(3) 重复执行过程(1)(2), 直至 $TFPSet = \Phi$ 。

经过冗余文本频繁模式过滤, 可明显缩减文本频繁模式集合容量, 提高文本频繁模式挖掘效率, 进而提升文本特征选择的性能。

2.3 基于关联度的文本特征选择方法

本节在经过过滤优化处理后的非冗余文本频繁模式基础上, 基于特征与文档的不同关联度对特征进行类别划分及权重分配, 以此实现文本特征选择。

定义 7 关联文档和非关联文档(Correlated Document and Uncorrelated Document) 指定类别 C , 若文本文档 Td 满足 $Td \in C$, 则称 Td 为关联文档。所有关联文档集合表示为 $TD_{cor} = \{Td | Td \in C\}$ 。若文档 Td 满足 $Td \notin C$, 则称 Td 为非关联文档, 所有非关联文档集合表示为 $TD_{uncor} = \{Td | Td \notin C\}$ 。 Td 的训练集合为 $TD = TD_{cor} \cup TD_{uncor}$ 。

定义 8 嵌入式文档(Embedded Document)

$WSet_TD_{cor}$ 表示关联文档集合 TD_{cor} 的词集。对于任意词语 $w \in WSet_TD_{cor}$, 有

$$Emb_TD_{cor}(w) = \{Td \in TD_{cor} | w \in Td\} \quad (2)$$

称为词语 w 的嵌入式关联文档集。

$$Emb_TD_{uncor}(w) = \{Td \in TD_{uncor} | w \in Td\} \quad (3)$$

称为 w 的嵌入式非关联文档集。

定义 9 关联度函数(Correlative Degree Function) 在训练集 $TD = TD_{cor} \cup TD_{uncor}$ 中, 词语 w 与文档间的关联度函数为:

$$CorDeg(w) = \frac{|Emb_TD_{cor}| - |Emb_TD_{uncor}(w)|}{n} \quad (4)$$

其中, $n = |TD_{cor}|$ 为关联文档数目。 $CorDeg(w)$ 值越大, 代表 w 与预定类别关联度越大。 $CorDeg$

$(w) > 0$ 表示 w 较常描述关联文档;反之,则说明 w 描述非关联文档较多。

定义 10 关联特征词语和普通特征词语 (Cor-related Feature Word and General Feature Word)

频繁出现在关联文档中且较少出现在非关联文档中的词语称为**关联特征词语**,如式(5)所示。

$$CorFW^+ = \{w \in WSet \mid CorDeg(w) \geq \delta\} \quad (5)$$

频繁出现在关联和非关联文档中的词语称为**普通特征词语**,如式(6)所示。

$$GenFW^0 = \{w \in WSet \mid CorDeg(w) < \delta\} \quad (6)$$

其中, δ 表示 $CorFW$ 和 $GenFW$ 的关联度界限。

定义 11 特征选择支持度 (Feature Selection Support) 词语 w_j 的特征选择支持度定义,如式

$$Weight(w) = \begin{cases} FS_Support(w, TD_{cor})(1 + CorDeg(w)), & w \in CorFW^+ \\ FS_Support(w, TD_{cor}), & w \in GenFW^0 \end{cases} \quad (8)$$

例如,假设训练集中包含的文档总数为 5,其中,3 个关联文档 Td_1, Td_2, Td_3 中包含特征词 w_2 , 且有 1 个非关联文档 Td_4 也包含 w_2 。从 Td_1, Td_2, Td_3 中提取的频繁模式如表 1 所示(符号 $<$ $>$ 脚标为频繁模式对应支持度):

表 1 文档与对应的频繁模式

关联文档	频繁模式
Td_1	$TFP_{11} = \langle w_1, w_2, w_5 \rangle_4, TFP_{12} = \langle w_2, w_3, w_4 \rangle_3, TFP_{13} = \langle w_3, w_8 \rangle_3$
Td_2	$TFP_{21} = \langle w_2, w_6 \rangle_5, TFP_{22} = \langle w_5, w_9 \rangle_3$
Td_3	$TFP_{31} = \langle w_2, w_7, w_8 \rangle_3$

那么,特征词语 w_2 的权重计算如下:

$$(1) \text{ 特征关联度: } CorDeg(w_2) = \frac{3-1}{5} = 0.4。$$

$$(2) \text{ 特征选择支持度: } FS_Support(w_2) =$$

$$\sum_{i=1}^3 \omega(w_{i2}) = \frac{7}{27} + \frac{5}{16} + \frac{1}{3} = 0.905。$$

其中, $\omega(w_{12}) =$

$$\frac{n(w_{12})}{n(w_{11}) + n(w_{12}) + n(w_{13}) + n(w_{14}) + n(w_{15}) + n(w_{18})} = \frac{4+3}{4+(4+3)+(3+3)+3+4+3} = \frac{7}{27},$$

$$\omega(w_{22}) = \frac{n(w_{22})}{n(w_{22}) + n(w_{25}) + n(w_{26}) + n(w_{29})} = \frac{5}{5+3+5+3} = \frac{5}{16},$$

$$\omega(w_{32}) = \frac{n(w_{32})}{n(w_{32}) + n(w_{37}) + n(w_{38})} = \frac{3}{3+3+3}$$

(7)所示。

$$FS_Support(w_j) = \sum_{i=1}^n \omega_{ij} \quad (7)$$

其中, $\omega_{ij} = n_{ij} / \sum_{j=1}^{N_i} n_{ij}$, n_{ij} 为 w_{ij} 所在文本频繁模式的支持度之和, N_i 为 Td_i 中文本频繁模式数目, n 为关联文档 $Td_i \in TD_{cor}$ 数目。

定义 12 特征权重分配函数 (Feature Weight Distribution Function) 词语 w 在关联文档集合 TD_{cor} 中的特征选择支持度为 $FS_Support(w, TD_{cor})$, 与预定类别的关联度为 $CorDeg(w)$, 则 w 的特征权重分配函数定义,如式(8)所示。

$$= \frac{1}{3}。$$

$$(3) \text{ 特征权重: } Weight(w_2) = FS_Support(w_2, TD_{cor})(1 + CorDeg(w_2)) = 0.905 \times (1 + 0.4) = 1.267。$$

2.4 算法伪代码

算法 1 为基于包含度和频繁模式的文本特征选择算法 TFSIDFP。步骤 1-26 为冗余文本频繁模式过滤过程,步骤 27-42 为文本特征选择过程。其中,步骤 1 初始化文本频繁模式过滤集合 $Filter_TFPSet$ 和文本频繁模式计数器 TFP_count_i ; 步骤 2 利用 FP-Growth 算法挖掘所有文本频繁模式,并按长度进行降序排序;步骤 3-6 判断集合 $Filter_TFPSet$ 是否为空,将 $TFPSet$ 中首个文本频繁模式 TFP_1 从集合中删除,加入 $Filter_TFPSet$ 中;步骤 7-20 为冗余模式过滤过程,将 $TFPSet$ 与 $Filter_TFPSet$ 中模式逐一比较,若 $TFPSet$ 中模式为 $Filter_TFPSet$ 中模式的子模式或二者相似度大于预定阈值,则将其从 $TFPSet$ 中删除,否则加入 $Filter_TFPSet$ 中;步骤 21-26 将非冗余文本频繁模式加入 $Filter_TFPSet$,判定 TFP_i 并非 $Filter_TFPSet$ 中任意文本频繁模式 TFP_j 的子模式或相似度较高的交差模式,则将 TFP_i 选入集合 $Filter_TFPSet$,并从 $TFPSet$ 中删除。步骤 27-30 定义变量及集合的值;步骤 31-34 计算文本特征词语支持度及关联度;步骤 35-36 为特征词语类别划分,采用聚类方式确定关联度界限 δ ;步骤 37-42 为文本特征词语加权;步骤 43 返回文本频繁模式过滤集合及特征词语权重。

算法 1 基于包含度和频繁模式的文本特征选择算法 TFSIDFP

INPUT: 关联文档集合 TD_{cor} 和非关联文档集合 TD_{uncor} , 其中 $Td_i \in TD_{cor}$; 相似度阈值 θ ;

OUTPUT: 文本频繁模式过滤集合 $Filter_TFPSet$; 文本特征词语权重: $Weight(w)$ 。

METHOD:

```

    /* 冗余模式过滤 */
(1)  $Filter\_TFPSet = \Phi, TFP\_count_i = 0$ 
(2)  $TFPSet = \text{procedure FP\_Growth}(Td_i)$  /* 挖掘频繁模式, 并按模式长度降序排序 */
    /* 判断  $Filter\_TFPSet$  是否为空, 将  $TFPSet$  中第一个模式  $TFP_1$  从集合中删除, 并加入  $Filter\_TFPSet$  */
(3) IF  $Filter\_TFPSet = \Phi$  THEN
(4)      $TFPSet = TFPSet - TFP_1$ 
(5)      $Filter\_TFPSet = Filter\_TFPSet \cup P_1$ 
(6) END IF
    /* 冗余模式过滤过程 */
(7) FOREACH  $TFP_i$  IN  $TFPSet$  DO
(8) FOREACH  $TFP_j$  IN  $Filter\_TFPSet$  DO
(9)     IF  $TFP_i \subseteq TFP_j$  THEN /* 子模式 */
(10)         $TFPSet = TFPSet - TFP_i$ 
(11)     ELSE IF  $TFP_i \not\subseteq TFP_j \& TFP_j \not\subseteq TFP_i \& TFP_i \cap TFP_j \neq \Phi$  THEN /* 交差模式 */
(12)         $SD(TFP_i, TFP_j) = \frac{Num(TFP_i \cap TFP_j)}{Num(TFP_i \cup TFP_j)}$ 
(13)        IF  $SD(TFP_i, TFP_j) \geq \theta$  THEN
(14)             $TFPSet = TFPSet - TFP_i$ 
(15)             $Support(TFP_j) += Support(TFP_i)$ 
(16)        ELSE
(17)             $TFP\_count_i += 1$ 
(18)        END IF
(19)     END IF
(20) END IF
(21) IF  $TFP\_count_i = |Filter\_TFPSet|$  THEN /* 将非冗余模式并入  $Filter\_TFPSet$  */
(22)      $Filter\_TFPSet = Filter\_TFPSet \cup TFP_i$ 
(23)      $TFPSet = TFPSet - TFP_i$ 
(24) END IF
(25) END FOR
(26) END FOR
    /* 文本特征选择 */
(27)  $n = |TD_{cor}|$  /* 关联文档数目 */
(28)  $WSet\_Filter\_TFPSet = \{w | w \in TFP, TFP \in Filter\_TFPSet\}$  /* 文本频繁模式过滤集合词集 */
(29)  $Emb\_TD_{cor}(w) = \{Td | Td \in TD_{cor}, w \in Td\}$  /* 嵌入式关联文档 */
(30)  $Emb\_TD_{uncor}(w) = \{Td | Td \in TD_{uncor}, w \in Td\}$  /* 嵌入式非关联文档 */
(31) FOREACH  $w$  IN  $WSet\_Filter\_TFPSet$  DO
(32)      $FS\_Support(w_j) = \sum_{i=1}^n (n_{ij} / \sum_{j=1}^{N_i} n_{ij})$  /* 计算特征选择支持度 */
(33)      $CorDeg(w) = \frac{|Emb\_TD_{cor}(w)| - |Emb\_TD_{uncor}(w)|}{n}$  /* 计算特征关联度 */
(34) END FOR
(35)  $CorFW^+ = \{w \in WSet | CorDeg(w) \geq \delta\}$  /* 词语类别划分 */
(36)  $GenFW^0 = \{w \in WSet | CorDeg(w) < \delta\}$  /* 词语类别划分 */
(37) FOREACH  $w$  IN  $w \in CorFW^+$  DO /* 关联特征词语加权 */
(38)      $Weight(w) = FS\_Support(w) * (1 + CorDeg(w))$ 
(39) END FOR
(40) FOREACH  $w$  IN  $GenFW^0$  DO /* 普通特征词语加权 */

```

(41) $Weight(w) = FS_Support(w)$

(42) **END FOR**

(43) **RETURN** $Filter_TFPSet, Weight(w)$

3 实验

对比实验分为与经典方法比较和与新方法比较。经典方法选取信息增益 IG、卡方统计 χ^2 和互信息 MI 三种特征选择方法与 TFSIDFP 方法作对比,采用分类器 SVM、KNN 和 NB(Naïve Bayes)评估分类效果。新方法选用近几年发表在国际期刊上的最新特征选择与特征抽取方法。

分类性能评价指标为准确率(Accuracy)、精确率(Precision)、召回率(Recall)和 F1 值(F1-Measure),及各自对应的宏平均值。

硬件环境: CPU 3.40Hz;内存 4G。软件环境: 操作系统 Windows7 32 位;开发环境 Eclipse JDK 1.6,Pydev 3.9;开发语言 Python 2.7。

3.1 数据集

数据集选取公共语料库 Reuters-21578: Acq(2 369 篇),Crude(578 篇),Earn(3 964 篇),Grain(1 102 篇),Interest(478 篇),Money(717 篇),Ship(286 篇),Trade(486 篇)。其中,训练样本与测试样本比例为 7:3。

3.2 实验结果

3.2.1 参数分析

为验证冗余文本频繁模式过滤方法有效性,令相似度阈值 θ 在最小支持度 min_sup 取不同值,得到非冗余频繁模式数量占模式总数的比重,如图 2 所示。可知 θ 取值不同,文本频繁模式过滤集合中模式数量在模式总数中占比均有明显下降,证明冗余模式过滤对提升频繁模式挖掘效率具有重要作用。为保证频繁模式尽可能多地保留与文档关联的信息,将 min_sup 设为较小值。由于 FP-Growth 算法和 TFSIDFP 方法复杂度较低, min_sup 较小并不会明显提升时间复杂度。随着 θ 设置增高,文本频繁模式过滤集合中的模式数量逐渐增多。 θ 设置过高,会保留大量冗余频繁模式; θ 设置过低,会过滤掉过多与文档关联的频繁模式。由图 2 可知,当 θ 取值为 0.7 左右,频繁模式数量相对稳定。因此,设定 $min_sup = 0.2, \theta = 0.7$ 。

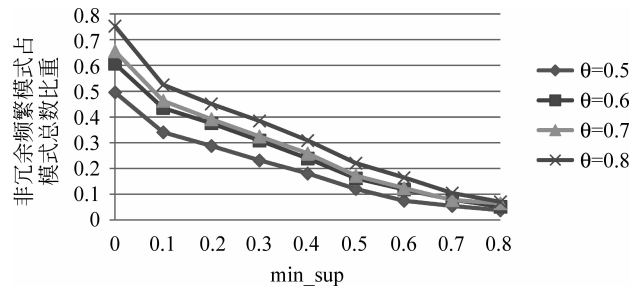


图 2 非冗余频繁模式占所挖掘模式总数的比重

3.2.2 特征选择性能评价

(1) 基于信息熵的性能评价

信息熵是一种常用的特征选择评价方法。假设将类别 $C_i (i = 1, \dots, n)$ 看作是一系列随机事件,对特征的某一取值 x ,样本属于各类的后验概率为 $P(C_i | x)$ 。则该特征的信息熵定义为: $H_{entropy} = - \sum_{i=1}^n P(C_i | x) \log_2 P(C_i | x)$,熵值越低,分类性能越好。

在数据集 Reuters-21578 中,比较基于关联度进行特征词语类别划分对特征熵值的影响。计算前 $k (k = 10, \dots, 2\,000)$ 个特征的平均熵值。如图 3 所示, unCor 表示仅参照词语支持度进行特征选择, CorDeg 表示在支持度基础上利用关联度进行词语类别划分和权重分配后进行特征选择。基于关联度划分特征词语后,关联特征词语 CorFW⁺ 相对该类的关联度加强,对类别区分力增强,错误率下降,对应熵值降低。由图 3 可知,前 200 个特征主要为关联特征,其对应熵值的平均值明显低于未使用关联度函数的特征;随着特征数目增加,普通特征数目增多,其取值差异较小,无法有效区分类别,平均熵值逐渐增大。因此,在所选特征数目有限的条件下,基于关联度进行特征选择,对类别划分更有效。

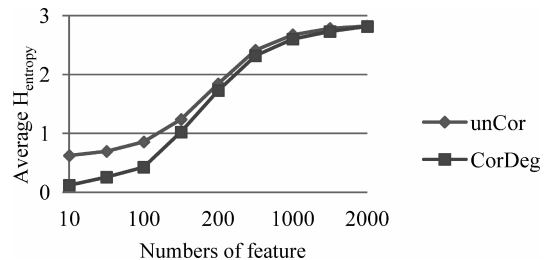


图 3 基于信息熵的特征选择方法性能对比

(2) 基于关联度的特征词语分类模型性能评价

在数据集 Reuters-21578 中,将基于关联度的特征词语分类模型在分类器 SVM 上进行验证,如图 4 所示。由图可知,关联特征词语 CorFW⁺和普通特征词语 GenFW⁰ 同时使用可以明显提升分类精度。相较 CorFW⁺ ∪ GenFW⁰ 而言,仅将 CorFW⁺ 用于分类效果欠佳,原因在于 CorFW⁺ 对所属类别区分性较好,却不足以完整描述该文档,需要

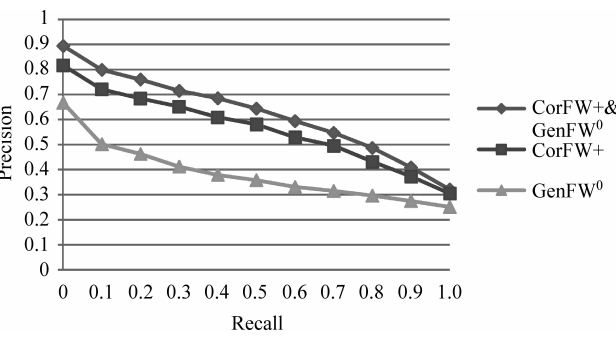


图 4 SVM 分类器采用不同特征词语的分类 PR 曲线

加入 GenFW⁰ 来辅助分类;若仅用 GenFW⁰,分类精度将大幅降低,这是由于 GenFW⁰ 频繁出现在关联和非关联文档中,无法有效划定文档类别。

3.2.3 与经典特征选择方法的比较

表 2 为数据集 Reuters-21578 中各分类器采用不同特征选择方法的性能对比。由此可看出,SVM 采用 TFSIDFP,准确率比 IG、 χ^2 、MI 分别高出 2.11%、2.37%、3.13%。相仿,KNN 采用 TFSIDFP 时分别高出 1.72%、1.94%、2.38%,NB 采用 TFSIDFP 时分别高出 2.57%、1.94%、2.99%。SVM 采用 TFSIDFP,P、R 和 F1 值分别高出其他三种方法中的最高值 9.31%、7.95%、8.60%;相仿,KNN 采用 TFSIDFP,P、R 和 F1 值分别高出其他三种方法中的最高值 7.90%、7.62%、7.87%;NB 采用 TFSIDFP,P、R 和 F1 值分别高出其他三种方法中的最高值 8.49%、9.46%、9.01%。可见 TFSIDFP 方法在这四个评价指标上性能均优于其他三种特征选择方法。

表 2 Reuters-21578 数据集中精确率、召回率和 F1 值对比

Reuters-21578 Data Sets	Model	TFSIDFP				IG				χ^2				MI			
		Acc/%	P/%	R/%	F1/%	Acc/%	P/%	R/%	F1/%	Acc/%	P/%	R/%	F1/%	Acc/%	P/%	R/%	F1/%
Grain	SVM	95.39	79.94	77.51	78.71	93.78	71.87	71.43	71.65	92.94	70.49	61.70	65.80	93.51	72.19	66.26	69.10
	KNN	94.82	77.02	75.38	76.19	92.64	67.87	62.92	65.30	92.98	68.89	65.96	67.39	92.18	66.44	58.36	62.19
	NB	95.02	78.66	75.08	76.83	91.78	63.26	60.18	61.68	92.68	68.21	62.61	65.29	91.74	64.24	56.23	59.97
Acq	SVM	92.58	86.09	81.97	83.98	89.43	80.03	73.94	76.86	87.89	76.36	70.99	73.58	86.26	71.70	69.58	70.62
	KNN	91.94	85.18	80.14	82.58	88.90	77.63	74.79	76.18	88.83	77.73	74.23	75.74	87.33	74.37	71.13	72.71
	NB	91.31	83.38	79.15	81.21	86.13	73.30	65.35	69.10	87.09	74.32	66.90	70.42	85.52	71.27	65.35	68.19
Crude	SVM	95.39	61.15	55.49	58.18	94.38	51.48	50.29	50.88	94.28	50.61	47.98	49.25	94.25	50.31	47.40	48.81
	KNN	94.92	56.52	52.60	54.49	94.42	51.81	49.71	52.28	95.19	59.35	53.18	56.10	93.95	47.47	43.35	45.32
	NB	95.09	57.74	56.07	56.89	93.98	47.58	39.88	43.39	93.75	45.73	43.35	44.51	93.78	45.86	41.62	43.64
Earn	SVM	91.64	90.72	87.97	89.32	85.09	84.24	76.87	80.39	84.65	84.63	75.02	79.54	82.68	80.58	74.35	77.34
	KNN	90.14	88.87	85.95	87.39	85.86	85.86	77.12	81.86	84.15	83.51	74.94	78.99	84.39	84.31	74.60	79.16
	NB	88.87	87.94	83.43	85.63	81.78	80.64	71.24	75.66	84.55	84.32	75.11	79.45	79.30	75.40	71.15	73.21
Interest	SVM	95.42	52.38	46.15	49.07	94.21	39.44	39.16	39.30	94.55	42.75	41.26	41.99	94.32	40.43	39.86	40.14
	KNN	95.32	51.16	46.15	48.53	94.12	38.30	37.76	38.03	93.68	33.09	31.47	32.25	93.98	34.45	28.67	31.30
	NB	94.18	39.01	38.46	36.73	93.41	30.99	30.74	30.86	93.65	33.33	32.86	33.09	93.48	31.69	31.46	31.57
Money	SVM	95.12	67.69	61.40	64.39	92.91	50.78	45.58	48.04	93.92	58.12	54.88	56.45	92.31	46.31	43.72	44.98
	KNN	94.05	59.20	55.35	57.21	92.84	50.26	44.18	47.02	92.34	46.82	47.91	47.36	91.88	43.20	41.39	42.28
	NB	93.51	55.33	50.70	52.91	91.71	42.11	40.93	41.51	91.88	42.93	39.54	41.16	91.54	40.95	40.00	40.47
Ship	SVM	96.82	44.00	38.37	40.99	96.77	42.25	34.88	38.21	96.59	39.74	36.05	37.81	95.85	30.21	33.72	31.87
	KNN	96.42	37.35	36.05	36.69	96.09	31.76	31.40	31.58	96.16	32.94	32.56	32.75	96.19	33.33	32.56	32.94
	NB	96.42	37.35	36.05	36.69	96.02	30.59	30.23	30.41	95.95	29.41	29.07	29.24	95.89	28.23	27.91	28.07
Trade	SVM	96.12	61.54	54.79	57.97	95.01	48.95	47.94	48.44	94.68	45.04	40.41	42.60	94.24	40.97	40.41	40.69
	KNN	95.05	49.25	45.21	47.14	93.98	37.86	36.30	37.06	93.78	36.11	35.62	35.86	93.71	35.42	34.93	35.17
	NB	94.32	41.26	40.41	40.83	93.31	31.25	30.82	31.03	93.61	34.48	34.25	34.36	93.55	33.79	33.56	33.67

续表

Reuters-21578 Data Sets	Model	TFSIDFP				IG				χ^2				MI			
		Acc/%	P/%	R/%	F1/%	Acc/%	P/%	R/%	F1/%	Acc/%	P/%	R/%	F1/%	Acc/%	P/%	R/%	F1/%
宏平均	SVM	94.81	67.94	62.96	65.36	92.70	58.63	55.01	56.76	92.44	58.47	53.54	55.89	91.68	54.09	51.91	52.98
	KNN	94.08	63.07	59.60	61.29	92.36	55.17	51.77	53.42	92.14	54.81	51.98	53.36	91.70	52.37	48.12	50.16
	NB	93.59	60.08	57.42	58.72	91.02	49.97	46.17	47.99	91.65	51.59	47.96	49.71	90.60	48.93	45.91	47.37

3.2.4 与新方法的比较

(1) 与新特征选择方法的比较

Filter 和 Wrapper 是两种主流的特征选择模式。基于 Filter 模式的特征选择方法基于原始数据评价特征性能,无需考虑具体分类器;与之不同,Wrapper 模式的特征选择方法依托具体分类器的分类性能对特征进行评价。

① 与 Filter 类型的特征选择方法的比较

Y Gao 等^[6]在信息过滤领域提出基于最大匹配模式的主题模型 MPBTM,其中使用了 Filter 类型的特征选择方法。MPBTM 模型使用模式表示各主题。这些模式依据统计和分类特性从主题模型中生成并组织,然后再选出最具代表性和区分力的最大匹配特征来判定文档与用户信息间的关联性,以此过滤不相关文档,提高文本分类性能。TFSIDFP 方法与 MPBTM 模型对比如图 5(a)所示,可知 TFSIDFP 性能优于 MPBTM。

为进一步验证 TFSIDFP 方法性能,使用 McNemar^[18]统计测试对 TFSIDFP 方法与 MPBTM 模型做统计显著性检验。分类器选用 SVM、KNN($k=1$)和 NB(Naive Bayes),显著性水平设定为 0.05。为获得稳定结果,每个算法均运行 10 次,验证结果如表 3 所示。其中,“Win”表示 TFSIDFP 性能明显优于 MPBTM;“Lose”表示 TFSIDFP 比 MPBTM 性能明显较差;“Tie”表示二者性能没有明显差别。由表可知,TFSIDFP 性能优于 MPBTM。

表 3 TFSIDFP 方法与 MPBTM 模型的统计显著性检验结果

Classifier	Win	Tie	Lose
SVM	6	4	0
KNN ($k=1$)	4	5	1
NB	2	7	1
Sum	12	16	2

② 与 Wrapper 类型的特征选择方法的比较

B Tang 等^[19]提出两种基于贝叶斯分类器的 Wrapper 类型的特征选择方法 MD 和 MD- χ^2 。以

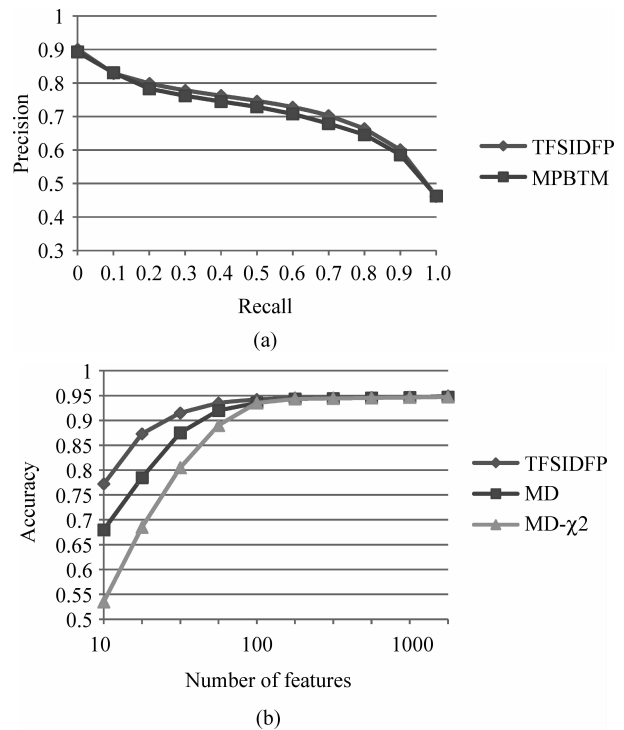


图 5 TFSIDFP 算法与新特征选择方法的比较

((a)与 Filter 模式的特征选择方法 MPBTM 的性能比较;
(b)与 Wrapper 模式的特征选择方法 MD 和 MD- χ^2 的性能比较)

特征对类别的区分力对其排序,选取适用于贝叶斯分类器的特征,以此进行分类。将 TFSIDFP 方法与 MD 和 MD- χ^2 方法做对比,特征选择范围选定为[10,2 000],分类准确率如图 5(b)所示。由图可知,分类精度随特征数目的增多均呈增高趋势,与 MD 和 MD- χ^2 方法相比,当所选特征数目较少时,TFSIDFP 方法展现了其优越性能。这是由于基于关联度的特征类别划分加强了关联特征权重,可明显提高关联特征的作用,提高分类精度。

(2) 与新特征抽取方法的比较

作为文本挖掘领域两种典型的特征选取方式,特征选择(Features selection)和特征抽取(Features extraction)均能有效地降低特征空间维数。特征选择是从 D 个特征中选出使准则函数最优的 d ($d < D$) 个特征,即选择特征子集;特征抽取是通

过适当变换将 D 个特征转换为 d ($d < D$) 个新特征。

M Khabbaz 等^[20]提出一种基于软聚类和信息增益特征约简的特征抽取方法 Cluster BOW-Inforgain。首先,软聚类方法使用模糊 C 均值将每一个词语依据不同组内关联度划分至多个聚类中,将每个聚类作为一个特征;然后利用信息增益进行特征约简。这样在传统词袋基础上,每篇文档被表示成一个经过软聚类及信息增益特征约简的特征向量。将 TFSIDFP 方法用于 SVM 分类器,与 Cluster BOW-Inforgain 方法的对比结果如图 6 所示。由图可知,当所选特征数目有限时,TFSIDFP 方法性能优于 Cluster BOW-Inforgain。这是由于特征提取是将所有词语进行转换从而降低维度,词语数目并未发生巨大缩减,Cluster BOW-Inforgain 方法每个聚类特征中均包含多个词语,因此分类需要的词语数目巨大。同时,由于 TFSIDFP 方法增大了关联特征词语强度,能有效提升分类精度。数据维度过高会增加系统开销,因此若能利用少量特征得到较高的分类精度,可明显提高分类性能和效率。因此当所选特征数目受限时,TFSIDFP 方法性能明显优于 Cluster BOW-Inforgain。

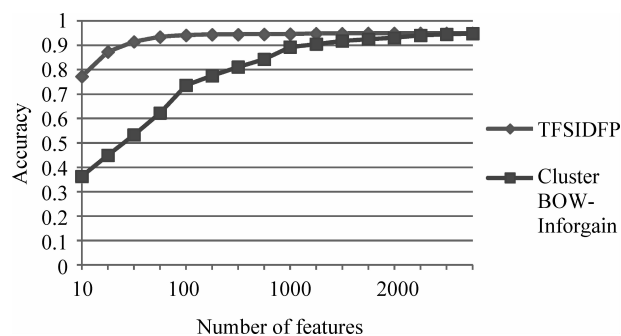


图6 TFSIDFP 方法与新特征抽取方法 Cluster BOW-Inforgain 的性能对比

4 总结

在文本数据量呈爆炸式增长的大数据时代,进行文本特征选择可快速并准确提取文本主题信息,提升文本分类精度。传统基于词语的文本特征选择方法被噪声问题影响,导致分类精度受到制约。提出基于包含度和频繁模式的文本特征选择方法。首先,定义基于包含度的相似性度量原理;然后,提出基于包含度的冗余文本频繁模式过滤方法;最后,提出基于关联度的文本特征选择方法。该方法基于包

含度原理度量文本频繁模式间相似性,去除冗余模式,提升文本频繁模式挖掘性能;基于冗余去噪后的非冗余模式选择文本特征,并利用特征与文档的关联度进行特征类别划分与权重分配,所选特征与文档关联度更强,对分类贡献度更大。该方法与传统基于词语文本特征选择方法相比,可以利用文本频繁模式中词语间关联性,很好地解决基于词语方法因无法有效克服噪声问题而导致的分类性能下降问题。对解决大数据时代的“数据爆炸”问题具有重要影响。此外,在进行特征选择时,还未深入考虑冗余特征词语对文本分类性能的影响,以后将深入研究特征词语去冗余方法,进一步提升文本特征选择质量及分类精度。

参考文献

- [1] Zhao Z, Wang L, Liu H, et al. On similarity preserving feature selection [J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(3): 619-632.
- [2] Zhuang Y T, Gao H D, Wu F, et al. Probabilistic word selection via topic modeling [J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(6): 1643-1655.
- [3] Song Q, Ni J, Wang G. A fast clustering-based feature subset selection algorithm for high-dimensional data[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(1): 1-14.
- [4] Li Y F, Algarni A, Albathan M, et al. Relevance feature discovery for text mining[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(6): 1656-1669.
- [5] 张延祥, 潘海侠. 一种基于区分能力的多类不平衡文本分类特征选择方法 [J]. 中文信息学报, 2015, 29(4): 111-119.
- [6] Gao Y, Xu Y, Li Y F. Pattern-based topics for document modelling in information filtering [J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(6): 1629-1642.
- [7] Zhao Z, Yan D, NG W. Mining probabilistically frequent sequential patterns in large uncertain databases [J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(5): 1171-1184.
- [8] Kessl R. Probabilistic static load-balancing of parallel mining of frequent sequences [J]. IEEE Transaction on Knowledge and Data Engineering, 2016, 28(5): 1299-1311.
- [9] Pumjun N, Kreesuradej W. Maintenance of multi-level

association rules discovery in dynamic database under a change of support threshold[C]//Proceedings of the 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2015: 618-623.

- [10] Zhang J Y, Yang C H. Sequence pattern mining based on Markov chain[C]//Proceedings of the 7th International Conference on Information Technology in Medicine and Education, 2015: 234-238.
- [11] 吐尔地·托合提, 维尼拉·木沙江, 艾斯卡尔·艾木都拉. 基于统计和浅层语言分析的维吾尔文语义串快速抽取[J]. 中文信息学报, 2017, 31(4): 70-79.
- [12] Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data[J]. Journal of Bioinformatics and Computational Biology, 2005, 3(2): 185-205.
- [13] Wang D, Nie F P, Huang H. Feature selection via global redundancy minimization[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(10): 2743-2755.
- [14] Gong M, Li H. Nonparametric statistical active contour based on inclusion degree of fuzzy sets[J]. IEEE Transactions on Fuzzy Systems, 2016, 24(5): 1176-1192.
- [15] Ma Q, Mi H H. The inclusion degree and similarity

degree of fuzzy rough sets defined by Nanda[C]//Proceedings of the 7th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2010), 2010: 354-357.

- [16] Liu W J, Fei Y. A sample classification algorithm based on inclusion degree[C]//Proceedings of the 7th International Conference on Fuzzy Systems and Knowledge Discovery, 2010: 1489-1491.
- [17] 李阳, 高大启. 知识图谱中实体相似度计算研究[J]. 中文信息学报, 2017, 31(1): 140-146, 154.
- [18] Dietterich T G. Approximate statistical tests for comparing supervised classification learning algorithms[J]. Neural Computation, 1984, 10(7): 1895-1923.
- [19] Tang B, Kay S, He H B. Toward optimal feature selection in Naïve Bayes for text categorization[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(9): 2508-2521.
- [20] Khabbaz M, Kianmehr K, Alhadj R. Employing structural and textual feature extraction for semi-structured document classification[J]. IEEE Transactions on Systems, Man, and Cybernetics, 2012(42): 1566-1578.



池云仙(1987—), 博士研究生, 主要研究领域为数据挖掘、智能信息处理、社会文化地理信息系统。

E-mail: chiyunxian_hebtu@163.com



李仁杰(1975—), 教授, 博士生导师, 主要研究领域为社会文化地理信息系统。

E-mail: lrjgis@hebtu.edu.cn



赵书良(1967—), 通信作者, 教授, 博士生导师, 主要研究领域为数据挖掘、智能信息处理。

E-mail: zhaoshuliang@sina.com