基于频繁项特征扩展的短文本分类方法

靳一凡 傅颖勋 马 礼
(北方工业大学信息学院 北京 100144)

摘 要 短文本具有特征维度高且稀疏等特点,导致将传统的分类方法应用于短文本分类时效果较差。针对此问题,提出基于频繁项特征扩展的短文本分类方法(Short Text Classification Based on Frequent Item Feature Extension,STCFIFE)。首先通过 FP-growth 算法挖掘背景语料库的频繁项集,结合上下文的关联特征,计算出扩展特征权重;然后将新特征加入到原短文本的特征空间中,在此基础上训练 SVM(Support Vector Machine,SVM)分类器,并进行分类。实验结果表明,与传统的 SVM 算法和 LDA+KNN 算法相比,STCFIFE 方法能有效缓解短文本特征不足、高维稀疏的问题,使 $F1$ 值提升了 2%~10%,提高了短文本的分类效果。

关键词 短文本分类,特征扩展,频繁项挖掘,特征权重,支持向量机
中图法分类号 TP391 **文献标识码** A

Method of Short Text Classification Based on Frequent Item Feature Extension

JIN Yi-fan FU Ying-xun MA Li
(College of Information,North China University of Technology,Beijing 100144,China)

Abstract Short text has the characteristics of high feature dimension and sparse,as a result,the traditional classification method is not effective in short text classification. To solve this problem,a short text classification method based on frequent item feature extension called STCFIFE was proposed. First of all,frequent itemsets in the background corpus are mined through FP-growth algorithm,and combining the contextual association feature,the extended feature weight is calculated. Then the new features are added to the feature space of the original short text. On this basis,SVM (Support Vector Machine) classifier is trained for classification. The experimental results show that,compared with the traditional SVM algorithm and the LDA+KNN algorithm,STCFIFE can effectively alleviate problems of feature deficiency and high dimensional sparsity in short text and improves $F1$ value by 2%~10%,improving the classification effect in short text.

Keywords Short text classification,Feature extension,Frequent item mining,Feature weight,Support vector machine

1 引言

随着互联网技术与移动通信技术的飞速发展,我们的生活无时无刻离不开微信、论坛、qq、短信、微博等网络应用。而人们相互交流会产生大量信息,这些信息的表现形式是短文本。短文本的字数一般在 140 字以内,具有词语少、特征维度高且稀疏等特点。现如今,社交媒体、新闻网站被广泛使用,已逐渐成为人们日常沟通交流以及信息获取不可缺少的方式,在日常的信息交流中发挥了重要作用。面对大规模的短文本数据,如何快速准确地获取有价值的信息,需要对其信息进行快速分类。短文本分类技术在自动问答、信息检索、话题追踪和搜索引擎等多个领域具有重要的研究及应用价值[1]并发挥着重要的作用,越来越受到研究者的关注。

与长文本相比,短文本具有特征少、维度高且稀疏等特性[1]。传统的向量空间模型(Vector Space Model,VSM)忽略了短文本自身的特点[1],使得 VSM 模型特征的稀疏性显得更为突出。将传统的机器学习方法直接应用到短文本分类上,准确性往往不高,其难点主要有如下两点[2-3]:

(1)特征不足

短文本篇幅简短,特征较少。若直接采用针对长文本的分类方法(如分词、去停用词),来计算词频-逆文本频率(TF-IDF),上下文依赖性强,特征关联性强,很容易丢失短文本的语义信息[3-4]。

(2)特征稀疏

短文本的篇幅简短,特征维度较高,但包含的特征数目有限,易导致特征向量非常稀疏。将传统的文本分类方法直接应用于短文本分类,未能取得良好的分类效果。

目前常用的文本分类算法有 SVM[5-7]、朴素贝叶斯[8](Native Bayes,NB)、K 近邻算法[9](K-Nearest Neighbor,KNN)和决策树[10](DesionTree,DT)等。研究者对上述算法的性能进行了比较,如 Yang 等[5]采用路透社的 Reuters-21578 语料库预测未知文本的类别,得到 SVM 算法的分类准

本文受国家自然科学基金(61702013),北京市优秀人才培养资助项目(2016000020124G016),北京市教委科技计划项目(KM201710009008),北方工业大学科研启动项目资助。

靳一凡(1994-),男,硕士生,主要研究方向为分布式信息处理等,E-mail:1009542253@qq.com(通信作者);**傅颖勋**(1986-),讲师,博士,CCF 会员,主要研究方向为云/分布式存储可靠性等;**马 礼**(1968-),教授,CCF 高级会员,主要研究方向为无线传感器网络、嵌入式技术等。

确性优于 KNN 等其他传统的分类算法的结论。Joachim[6] 通过实验说明 SVM 算法应用在文本分类中克服了过拟合因素的影响，泛化性能较好，但无法有效解决缺失特征。

针对短文本分类中存在的问题，目前研究主要采用特征扩展来解决短文本特征维度高且稀疏的问题，具体研究方向如下：

(1)利用知识库进行特征扩展。例如文献[11-12]利用 WordNet 发现了短文本中词语间的语义关系，不足之处在于其仅对收录的词语有效。

(2)文献[13-14]借助搜索引擎的返回结果对文本进行了特征扩展。搜索引擎的质量对分类结果影响较大，返回了大量冗余数据，且时间长，复杂度高，难以实现。

针对上述研究存在的问题，Chen 等[15] 采用 LDA(Latent Dirichlet Allocation)主题模型对短文本进行建模，然后使用 SVM 算法对建模后的文本进行分类，在降维幅度 90% 下提高了文本分类的准确率，但短文本特征稀疏又严重影响了 LDA 主题模型对短文本建模的效果，造成了分类不准确，模型的困惑度(Perplexity)普遍偏高。Xue 等[16] 考虑了文档类别信息对 LDA 模型的改进，结合 SVM 的核函数，提出了一种有监督的中文短文本分类方法——Labeled LDA-Kernel SVM，但面对特征高维稀疏的短文本，仍无法有效地解决短文本特征间的关联性与依赖性。Yuan 等[17] 采用挖掘背景语料库中的频繁模式的方法，利用背景语料库的有效特征丰富了短文本的信息，在特征处理上，相对 LDA 模型其计算量相对较小，因此普适性较强，复杂度较低，但此方法忽略了短文本原有特征和扩展特征权重的计算，虽然简化了模型的复杂度，但无法准确反映特征词对类别的贡献程度，造成文本表示不准确[18-19]。

本文以上述文献为研究基础，根据短文本的单词数量少、描述特定信息弱的特点，提出了基于频繁项特征扩展的短文本分类方法 STCFIFE。将挖掘出的背景语料库的频繁项集作为扩展特征，计算扩展特征权重，并加入到短文本特征向量空间中，构建 SVM 分类器进行分类。该方法考虑了上下文之间的特征关联，解决了短文本特征不足稀疏等问题，提高了分类的准确性。

2　相关工作

2.1　频繁项集挖掘算法

频繁项集挖掘算法有 Apriori 算法[20] 和 FP-growth 算法[21] 等。Apriori 算法是通过引入候选集计算频繁项集的。针对每一个候选集，扫描一遍数据库，多次读写数据库需要消耗许多时间。由于 FP-growth 算法挖掘语料库的频繁项集不需要创建候选集，且只需要遍历两次数据集。第一次扫描是统计元素项的支持度，元素项按支持度降序排列，第二次扫描构建 FP 树，挖掘频繁项集。当挖掘完包含某个元素项的频繁项集时，FP 树就不会遍历这个元素项，它所占用的内存空间会被立刻释放。对于大规模的文本挖掘频繁项集，FP-growth 算法的内存利用率和挖掘效率显然要优于 Apriori 算法。因此本文采用 FP-growth 算法挖掘背景语料库的频繁项集。

2.2　SVM 算法

SVM 算法[5-7] 是目前公认的文本分类效果较好的机器学习算法。SVM 模型本质上是一个二元分类器，将学习问题归结为一个凸二次规划问题，得到的是全局最优解，解决了在神经网络方法中的局部极值问题，对于线性不可分的数据，可通过核函数将数据映射到高维空间，使数据在高维空间用线性函数分类，解决了“维度灾难”问题。SVM 由于性能上的优势，近年来一直是数据挖掘和信息检索领域的研究热点[19-21]。SVM 算法应用在文本分类中，可以避免数据集分布不均、冗余特征以及过拟合等因素的影响，具有很好的泛化能力。SVM 分类模型在处理多分类任务时，需要将多个训练好的 SVM 二元分类器组合成多元分类器。

3　频繁项特征扩展的短文本分类方法

3.1　相关定义

定义集合 $ST=\{d_1,d_2,\cdots,d_T\}$ 为短文本数据集，是短文本新闻标题。$B=\{b_1,b_2,\cdots,b_m\}$ 是新闻正文内容，为与 ST 相关的背景语料库。ST 的特征集合为 $V=\{v_1,v_2,\cdots v_m\}$；文本 $d_i\in ST$ 的向量表示为 $V^i=\{w_1^i,w_2^i,\cdots,w_m^i\}$，其中 w_k^i 为 v_k 在短文本 d_i 的权重。tf_{ki} 表示词 v_k 在 d_i 出现的频次，df_{ki} 代表 ST 含有词 v_k 的文本数目。w_k^i 的计算公式如下：

$$w_k^i=tf_{ki}*\ln[m/(df_{ki}+1)] \tag{1}$$

定义 1　支持度 $sup(X)$ 为特征词 X 出现的文本数目 $num(X)$ 与语料库 B 文本数目 $num(B)$ 的比值。计算公式为：

$$sup(X)=num(X)/num(B) \tag{2}$$

定义 2(置信度)　当特征词 X 出现时，特征词 Y 会以某概率出现，特征 Y 出现的概率称为 $X->Y$ 的置信度，记为 $conf(X->Y)$。其计算公式如下：

$$conf(X->Y)=P(Y|X)=sup(X\cup Y)/sup(X) \tag{3}$$

3.2　挖掘频繁项集方法

采用 FP-growth 算法挖掘频繁项集的步骤如下：

步骤 1　第一次遍历背景语料库，统计每篇文本特征词出现的次数，创建头指针表 ht，移除头指针表中小于 num 的元素项。

步骤 2　第二次遍历背景语料库，初始化 FP 树为空集。过滤掉低于最小次数的特征词，按次数重排序。

步骤 3　更新 FP 树，更新头指针表。按顺序创建频繁项条件 FP 树，根据频繁项前缀路径计算频繁项集，加入频繁项集合。

步骤 4　若条件 FP 树不为空，返回步骤 1；若 FP 树为空，进行步骤 5。

步骤 5　返回频繁项集合。

3.3　特征扩展及分类

短文本特征少和高维稀疏的特点，导致特征抽取及特征补充都存在许多问题，使分类效果表现不佳。本方法借助背景语料库，采用频繁项挖掘的方式，对短文本进行特征扩展。

特征扩展和分类的步骤如下：

步骤 1　对短文本和背景语料库进行预处理。包括分词、去停用词，过滤冗余信息，构建文档向量空间模型。

步骤 2　选择对类别贡献最大的前 K 个特征为原始特征，由获取的频繁项集计算关联规则“$X->Y$”的置信度 $conf(X->Y)$，筛选大于阈值的规则。

步骤 3　遍历关联规则集合，对于每个规则，若 X 是短文本集 ST 的特征，而 Y 不是 ST 的特征，则将 Y 加入 ST 的扩

展特征集 *exd_V* 中，新加入的特征权重为短文本对应特征权重与关联规则置信度的乘积，生成特征矩阵。

步骤 4　构建 SVM 分类器，采用一对多法(one to rest，1 to R)训练 SVM 多分类模型，一类为正例，其余类别为反例，将多分类问题转化为二元 SVM 分类问题，再评估分类效果。

频繁项特征扩展的短文本分类流程如图 1 所示。

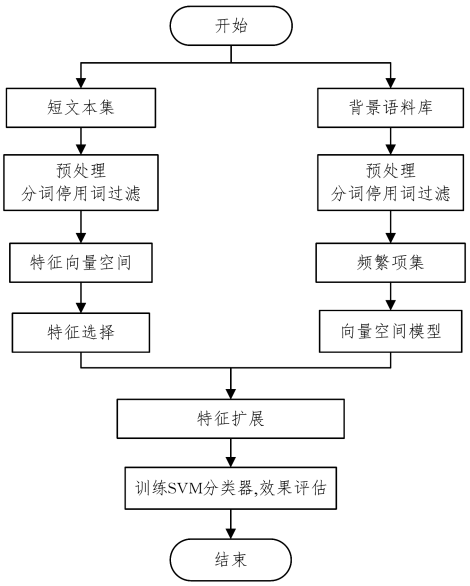

图 1　短文本分类流程图

例如：在 IT 类别的短文本“苹果使用流畅，但是价格很贵”，经过预处理后，存在“苹果、使用、流畅、但是、价格、很贵”这 6 个特征，利用式(1)计算特征权重分别为 0.3，0.1，0.2，0.05，0.15，0.1。通过挖掘背景语料库的频繁项集，存在(苹果，手机)、(流畅，系统)、(价格，购买)这 3 个频繁项集，对应的关联规则置信度分别为：

$$conf(\text{苹果}->\text{手机})=P(\text{手机}|\text{苹果})=0.8$$

$$conf(\text{流畅}->\text{系统})=P(\text{系统}|\text{流畅})=0.6$$

$$conf(\text{价格}->\text{购买})=P(\text{购买}|\text{价格})=0.3$$

置信度阈值设为 0.5，(价格，购买)对应的关联规则置信度为 0.3，小于置信度阈值 0.5，因此直接舍弃。同理，保留(苹果，手机)、(流畅，系统)。扩展特征权重是：

$$V(\text{手机})=V(\text{苹果})*conf(\text{苹果}->\text{手机})=0.3*0.8=0.24$$

$$V(\text{系统})=V(\text{流畅})*conf(\text{流畅}->\text{系统})=0.2*0.6=0.12$$

利用频繁项特征扩展后的短文本特征为“苹果 使用流畅 但是 价格很贵 手机 系统”这 8 个特征，对应特征权重分别为 0.3，0.1，0.2 ，0.05，0.15，0.1，0.24，0.12。由于频繁项关联特征加入到短文本的特征空间，缓解了短文本的特征稀疏性。在此特征空间上训练 SVM 分类器，分类准确性得到了提高。

算法 1　频繁项特征扩展算法伪代码

Input：短文本特征集 $V=\{\langle t_1,w_1\langle t_2,w_2\rangle\rangle,\cdots,\langle t_n,w_n\rangle\}$，频繁项集 FPI

Output：扩展后短文本特征集 exd_V

Begin：

1. 初始化 exd_V 空集
2. for each m，m 取值为[1，n]
3.　for each $(t_i,t_j)\in$ FPI
4.　if ($t_m=t_j$ and t_m not in V U exd_V)，do
5.　exd_V=exd_V U $\{t_j,\langle Conf(t_m->t_j)*V(t_m)\rangle\}$
6.　end if
7.　end for each
8.　exd_V=exd_V U V
9. end for each
10. return exd_V
11. End

4　短文本分类实验

实验数据集是搜狗实验室提供的新闻语料库[21]，包含了新浪、搜狐等各大知名网站的短文本。每个类别选取 20 000 篇为训练集，5 000 篇为测试集，包含军事、汽车、健康、文化、IT 5 个类别。

分类评价指标有：精确率、召回率和 F1 值[19]。F1 值是精确率与召回率的调和平均值[19]，评价的是分类的效果。计算公式如下：

$$Pecision=\text{预测正确数目}/\text{预测总数} \tag{4}$$

$$Recall=\text{召回的文档数}/\text{文档总数} \tag{5}$$

$$F1=2*Pecision*Recall/(Pecision+Recall) \tag{6}$$

4.1　置信度对分类的影响

频繁项集(X,Y)的数目是由关联规则 $X->Y$ 的置信度阈值决定的。为了验证置信度阈值对频繁项集和对分类效果的影响，取置信度阈值 β 分别为 0.5，0.6，0.7，0.8，0.9，对比挖掘出的频繁项集数和分类 $F1$ 值。

表 1　置信度阈值的选取

置信度阈值 β	频繁项集数目	分类 $F1$ 值
0.5	10352	74.52
0.6	7396	77.34
0.7	6538	78.15
0.8	5723	78.86
0.9	4842	77.92

实验表明：随着置信度阈值的增加，频繁项集的数目减少。例如：置信度阈值为 0.5 时对应的频繁项集的数目是置信度阈值为 0.9 时的两倍多。分类 $F1$ 值随着置信度阈值的提升呈先上升再下降的趋势。原因是置信度阈值过低会导致扩展特征冗余，分类效果差。随着置信度阈值的提升，频繁项集的数目降低，特征关联性提高，分类效果明显改进。如果置信度阈值过，则高频繁项集过少，短文本的特征空间变化不大，分类效果无明显提升。综合考虑选择置信度阈值为 0.8，分类效果为最好。

4.2　分类效果的实验对比

为了验证本文提出的基于频繁项特征扩展的短文本分类方法的有效性，进行 3 组对比实验，STCFIFE 方法与无特征扩展的 SVM 算法和文献[15]提出的 LDA+KNN 算法分别进行对比。选取置信度阈值为 0.8，实验结果如表 2 所列，对比情况如图 2 所示。

表 2　不同算法的 $F1$ 值对比表

类别	STCFIFE	LDA+KNN	无扩展的 SVM
汽车	80.22	76.69	72.51
IT	82.14	77.25	73.79
健康	71.62	69.53	66.76
文化	69.21	67.38	66.87
军事	84.08	81.77	73.91

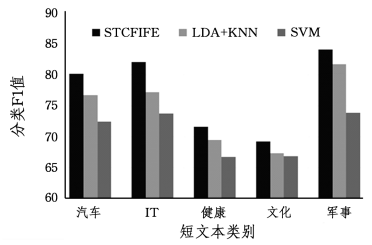

图 2　不同算法的 $F1$ 值对比图

由实验结果可知，采用频繁项特征扩展的分类算法 STCFIFE 的分类效果要优于无特征扩展的传统 SVM 算法与 LDA＋KNN 分类算法，各类别的分类 $F1$ 值都得到了明显提升。与无特征扩展的 SVM 算法相比，STCFIFE 方法的分类 $F1$ 值提升约 2%～10%；与 LDA＋KNN 算法相比，STCFIFE 方法的分类 $F1$ 值提升约 2%～5%。这说明本文提出的基于频繁项特征扩展的短文本分类方法是切实有效的，在一定程度上提升了短文本分类的准确性。

结束语　针对短文本分中类存在特征少且稀疏等问题，提出了基于频繁项特征扩展的短文本分类方法 STCFIFE。其核心思想是挖掘背景语料库的频繁项作为扩展特征，并加入到原短文本特征空间中，使之适合短文本分类。实验结果表明，置信度对频繁项集规模有很大影响，选择合适的置信度有助于提升分类效果。STCFIFE 方法与传统的 SVM 算法和 LDA＋KNN 算法相比，各类别短文本的分类 $F1$ 值均有一定的提高。但随着训练样本数目的增多，分类效率有一定下降。在特征扩展阶段对新权重的计算还有待改进，后期将继续设计可行的特征平滑公式对扩展特征权重的计算方式进行优化，以寻找改进的方法。

参 考 文 献

[1] 张志飞，苗夺谦，高灿. 基于 LDA 主题模型的短文本分类方法[J]. 计算机应用，2013，33(6)：1587-1590.

[2] 王雯，赵衍衍，李翠平，等. Spark 平台下的短文本征扩展与分类研究[J]. 计算机科学与探索，2017，34(5)：1-9.

[3] 王振振，何明，杜永萍. 基于 LDA 主题模型的文本相似度计算[J]. 计算机科学 2013，40(12)：229-232.

[4] 石晶，李万龙. 基于 LDA 模型的主题分析[J]. 自动化学报，2009，35(12)：1586-1593.

[5] YANG Y，ZHANG J，KISIEL B. A scalability analysis of classifiers in text categorization [C]// Proceedings of the 26th ACM International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-03). Toronto：ACM Press，2003：96-103.

[6] JOACHIMS T. Text Categorization with Support Vector Machines：Learning with Many Relevant Features[J]. Machine Learning，1998，1398(23)：137-142.

[7] CALMA A，REITMAIER T，SICK B. Semi-Supervised Active Learning for Support Vector Machines：A Novel Approach that Exploits Structure Information in Data[J]. Information Sciences，2018，456：13-22.

[8] 徐光美，刘宏哲，张敬尊. 基于特征加权的多关系朴素贝叶斯分类模型[J]. 计算机科学，2014，41(2)：283-285.

[9] 胡元，石冰. 基于区域划分的 KNN 文本快速分类算法研究[J]. 计算机科学，2012，39(10)：182-186.

[10] 季一木，张永潘，郎贤波，等. 面向流数据的决策树分类算法并行化[J]. 计算机研究与发展，2017，54(9)：1945-1957.

[11] SHIRAKAWA M，NAKAYAMA K，HARA T，et al. Wikipedia-Based Semantic Similarity Measurementsfor Noisy Short Texts Using Extended Naive Bayes[J]. IEEE Transactionson Emerging Topics in Computing，2015，3(2)：1.

[12] LIU W S，CAO Z W，WANG J，et al. Short text classification based on Wikipedia and Word2vec[C]// 2nd IEEE International Conference on Computer and Communications (ICCC). 2016.

[13] HE H，CHEN B，XU W，et al. Short Text Feature Extraction and Clustering for Web Topic Mining[C]// Proceedings of the Third International Conference on Semantics，Knowledge and Grid. IEEE Computer Society，2007：382-385.

[14] LIU J L，YAN Y Y. SMS Text Classification Method Based on Context[J]. Computer Engineering，2011，37(10)：41-43.

[15] CHEN Q U，YAO L X，YANG J. Short text classification based on LDA topic model[C]// International Conference on Audio，Language and Image Processing (ICALIP). 2016.

[16] WANG X L，WANG J，YANG Y. Labeled LDA-Kernel SVM：A Short Chinese Text Supervised Classification Based on SinaWeibo[C]// 4th International Conference on Information Science and Control Engineering(ICISCE). 2017.

[17] YUAN M. Feature Extension for Short Text Categorization Using Frequent Term Sets[J]. Elsevier Procedia Computer Science，2014，31：663-670.

[18] FENG G，LI S，SUN T，et al. A Probabilistic Model Derived Term Weighting Scheme for Text Classification[J]. Pattern Recognition Letters，2018，110：23-29.

[19] MIROЙCZUK M M，PROTASIEWICZ J. A Recent Overview of the State-of-the-Art Elements of Text Classification[J]. Expert Systems with Applications，2018，106：36-54.

[20] LI H，WANG Y，ZHANG D，et al. Pfp：parallel fpgrowth for query recommendation[C]// Proceedings of the 2008 ACM Conference on Recommender Systems. ACM，2008：107-114.

[21] SOGOULABS. SogouCS，version：2012[OL]. http：//www.sogou.com/ labs/resource/cs.php.