基于贝叶斯网的短文本特征扩展方法

刘慧清 郭延哺 李红灵 李维华
(云南大学信息学院 昆明 650500)

摘 要 针对短文本特征词稀疏、表示能力不足等问题,提出了一种基于贝叶斯网的短文本特征扩展方法。该方法根据短文本中特征词之间的依赖关系构建语义贝叶斯网,定义特征词与短文本之间的关联度。基于贝叶斯网的推理计算关联度,将与短文本关联密切的特征词扩展到短文本中,以达到降低短文本的噪声、改善特征稀疏的目的。在此基础上,以短文本分类作为基本的文本分析任务,分析所提方法的可行性和有效性。在 Amazon 评论数据集上进行实验,结果表明所提方法是可行和有效的。

关键词 文本分析,短文本,特征扩展,贝叶斯网
中图法分类号 TP391 **文献标识码** A

Short Text Feature Extension Method Based on Bayesian Networks

LIU Hui-qing GUO Yan-bu LI Hong-ling LI Wei-hua
(School of Information, Yunnan University, Kunming 650500, China)

Abstract Aiming at the problems of feature sparsity and insuffcient representation ability in short text, this paper proposed a feature extension method based on Bayesian networks. Firstly, the semantic Bayesian network is constructed by defining the dependencies between the feature words in the short texts. Then, the correlation degree is defined between the feature word and the short text, and the feature words closely related to the short text are selected. These words are further extended to the short text to reduce the noise and sparsity of short texts. Finally, this paper analyzed the feasibility and effectiveness of the proposed method with the short text classification as the basic task of text analysis. The experimental results on the Amazon product dataset show that the proposed method is feasible and effective.

Keywords Text analysis, Short text, Feature extension, Bayesian network

1 引言

随着互联网技术与移动通信技术的结合,短信、微博、客户评论等短文本成为了一种重要的信息传播方式。相对于长文本,短文本具有字数少、数据量大、实时性高、应用广泛等突出特性。因此,短文本在日常网络沟通及信息获取中扮演着重要的角色。对短文本进行有效地分析并从中发现有用的信息,可以为话题追踪、舆情分析、信息检索、个性推荐等[1-3]方面提供技术基础。随着短文本数据的增加,如何有效利用丰富的数据资产,研究其再生价值,成为当下学术界及工业领域研究的热点之一[4-6]。

有效地表示短文本的特征是短文本分析的基础。文本的特征表示一般以特征词为基础。然而,短文本字数偏少、特征稀疏、没有足够的信息量来进行统计推断,是短文本特征表示中典型的问题,给自动化的短文本分析带来了挑战[7]。

针对特征稀疏这个明显的问题,特征扩展成为解决该问题的一个有效方法,并且得到了广泛的研究。特征扩展目前的研究方向主要有:

(1)基于外部语料库进行扩展。Yu 等[8]针对短文本中的每个术语,通过从概率知识库中获取其概念和共现术语来丰富短文本内容,从而实现对短文本的检索、分类等文本处理。Wang 等[9]基于外部知识库挖掘短文本中隐含的信息,并利用与短文本相关的相关概念构建概念语言模型,以实现短文本的扩展查询。崔婉秋等[10]提出一种将维基百科的社交与概念化语义结合的扩展搜索算法,通过挖掘社交网络独特的社交属性和链接信息 URL,对微博短文本实现进一步的社交语义扩展。然而,该方法过分依赖外部语料,对于一些专业性很强或语言比较特殊的短文本时,则很难找到合适的外部语料。

(2)基于内部信息进行扩展。基于内部信息的扩展方法主要在于挖掘短文本数据内部的语义信息,选取有用的特征对短文本进行扩展。吕超镇等[11]提出一种基于 LDA 的特征扩展方法,利用 LDA 主题模型对短文本进行预测,得到对应的主题分布,把主题中的词作为短文本的部分特征,扩充到原短文本的特征中去。Xu 等[12]采用深度学习算法和神经网络语言模型来挖掘深度词义知识,通过融合字和词来扩展文本特征。Sriram 等[13]根据 Twitter 数据的特点,提出基于作者信息和 Twitter 相关特征结合的方法对 Twitter 短文本进行预归类处理。上述方法大多是针对特定的领域或者分类算

本文受云南省应用基础研究计划重点项目(2016FA026),国家自然科学基金项目(61762090),云南大学研究生科研创新基金项目(2018226)资助。

刘慧清(1996-),女,硕士生,主要研究方向为自然语言处理等;**李维华**(1977-),女,博士,副教授,主要研究方向为机器学习,E-mail:lywey@163.com(通信作者)。

法，不具有普适性，并且在特征选择的关键步骤上仍然存在一些缺陷，即只对与短文本具有直接关系的特征进行了扩展。

此外还有采用词向量模型进行特征空间优化的同义词扩展方法，Zhang 等[14]采用 Wordnet 和词向量结合的方法对短文本进行扩展，并利用扩展后的语料进行分类实验；也有研究者采用背景知识来扩展短文本特征[15]；还有研究者利用搜索引擎进行特征扩展，在提取特征后，通过百度搜索引擎检索该关键字，但是检索得到的文档并不理想，很难准确地找出符合关键词特征的结果[16]。

贝叶斯网（Bayesian Network，BN）[17-18]是每个节点都有一张概率表（也称参数）的有向无环图，不仅可以直观地表示节点之间的定量依赖关系，同时也为这些依赖关系提供了有效、可靠的推理和分析。近年来，贝叶斯网络被广泛地用于预测、智能推理、诊断、决策、风险评估、可靠性分析等问题[19]。Tang 等[20]研究了简单的朴素贝叶斯分类器与一种新颖的特征选择框架之间的渐进性质，并用其对文本进行分类。陈为等[21]针对贝叶斯网络的拓扑结构、条件概率对地理空间进行了可视化分析。王双成等[22]采用贝叶斯网络回归模型实现对动态与静态信息的融合，并通过具有不同超父节点贝叶斯网络的集成来降低回归误差和提高泛化能力。

因此，本文提出一种基于贝叶斯网的短文本特征扩展方法。其充分利用贝叶斯网丰富的概率表达能力和强大的推理能力，将短文本的特征以及特征之间的关系定量地表示在贝叶斯网中，并基于贝叶斯网进行特征推理，对短文本的特征进行扩展，以弥补短文本特征稀疏的不足，并为进一步的短文本分析提供支持。

本文的研究工作主要包括以下几个方面：

1）基于贝叶斯网对文本特征之间的定量依赖关系进行建模，建立语义贝叶斯网（Semantic Bayesian Network，SBN），为短文本的特征扩展提供基础。

2）基于语义贝叶斯网以及贝叶斯网的推理算法，设计特征选择及扩展算法，实现短文本的特征扩展。

3）在亚马逊评论数据集上设计一系列实验，对本文所提方法进行分析。实验结果表明基于贝叶斯网的短文本扩展方法具有较好的可行性和有效性。

具体研究路线如图 1 所示。

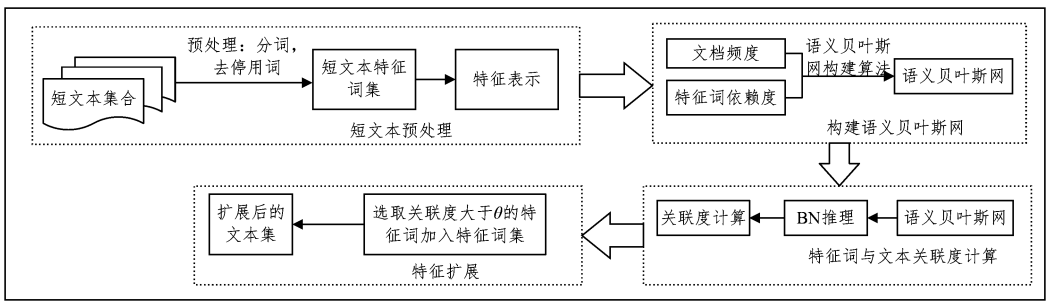

图 1　研究路线图

2　语义贝叶斯网

本文提出的基于贝叶斯网的短文本扩展方法，通过提取短文本的特征，将短文本的特征词抽象为节点，将特征之间的依赖关系定量地表示为贝叶斯网的有向边及参数，建立语义贝叶斯网（Semantic Bayesian Network，SBN），为短文本的特征扩展提供基础，也可以为进一步的短文本特征推理和分析提供支持。

2.1　文本特征提取

为了定量建模短文本之间的特征依赖关系，我们首先提取短文本的特征。对于 n 个短文本的集合 $D=\{d_1,d_2,\cdots,d_n\}$，本文采用 TF-IDF 方法提取短文本的特征词，筛选短文本中代表性较强的词。由于一些特征词是由 n 个连续的单词组成的序列，所以本文采用了 N-gram 模型。

TF-IDF 中 tf 表示词语在文本中出现的频率，idf 表示逆文本频率指数，idf 越大，则说明词语具有越好的类别区分能力。词项的 $TF\text{-}IDF$ 值通常用式(1)和式(2)度量。

$$TF\text{-}IDF(t_i)=tf(t_i)\times idf(t_i) \tag{1}$$

$$idf(t_i)=\log(N/df(t_i)) \tag{2}$$

其中，t_i 表示第 i 个特征词，$tf(t_i)$ 表示 t_i 在当前文本中所出现的频率，N 表示文本集合中所有的文本数，$df(t_i)$ 表示出现词项 t_i 的文本数。

对包含 n 个短文本的集合 $D=\{d_1,d_2,\cdots,d_n\}$ 经过特征提取，最终得到 m 个特征词，表示为 $T=\{t_1,t_2,\cdots,t_m\}$，D 中每个文本 d_i 的特征词表示为 $T_i=\{t_{i1},t_{i2},\cdots,t_{ia}\}$。

2.2　语义贝叶斯网

本文基于贝叶斯网对短文本集进行建模，构建语义贝叶斯网，通过贝叶斯网为短文本集中特征词及其之间的关系提供一个直观的表示框架，也为进一步的特征推理和分析提供基础和支持。

语义贝叶斯网由一个特征词组成的有向无环图以及条件概率表构成，下面给出了语义贝叶斯网的定义。

定义 1（语义贝叶斯网）　对于短文本集合 $D=\{d_1,d_2,\cdots,d_n\}$，$T=\{t_1,t_2,\cdots,t_m\}$ 是特征词集合，D 上的语义贝叶斯网是一个二元组 (G,P)，其中：

（1）$G=(U,E)$ 是一个有向无环图，且 $U=\{u_1,u_2,\cdots,u_m\}$，每个节点表示 T 中的一个特征词，且取值空间为 $\{0,1\}$，分别表示特征词出现和不出现。E 表示节点之间的有向边集合，若存在有向边 $u_j\rightarrow u_i$，则称 u_j 是 u_i 的父节点，$pa(u_i)$ 表示 u_i 的父节点集合。

（2）$P=\{p(u_i\mid pa(u_i))\mid u_i\in U\}$ 表示条件概率表集合，$p(u_i\mid pa(u_i))$ 表示每个节点 u_i 在其父节点 $pa(u_i)$ 条件下的条件概率表。

语义贝叶斯网是一个有向无环图，节点是短文本的特征，节点之间的边和参数反映了特征之间的定量关系。本文在确定节点之间的边和方向时，不仅考虑了特征词之间的相似关系，还考虑了特征词在给定上下文背景下的语义关联，所以本文进一步定义文档频度以及特征依赖度。下面给出了文档频度和特征依赖度的具体概念。

定义 2（文档频度）　对于短文本集合 D，特征词集合 T，

设 x 是一个特征词组合序列；对于特征词序列 x 中的特征词 $t_i \in T$，t_i 的取值为 0 或 1，分别表示特征词 t_i 是否出现；x 在 D 中出现的文档频数称为 x 的文档频度，用 $c(x)$ 表示。

例如，$c(t_i=1,t_j=1)=5(t_i,t_j \in T)$，表示在短文本集合 D 中，两个特征词 t_i 和 t_j 都出现的文档数为 5。

定义 3（特征依赖度） 对于短文本集合 $D=\{d_1,d_2,\cdots,d_n\}$，设 T 是特征词集合，设任意两个特征词 $t_i,t_j \in T$，则 t_i 和 t_j 之间的依赖程度 $I(t_i,t_j)$ 定义为：

$$I(t_i,t_j)=\log_2\left(\frac{n\times c(t_i=1,t_j=1)}{c(t_i=1)\times c(t_j=1)}+1\right) \tag{3}$$

其中，$I(t_i,t_j)$ 表示整个文本集上任意特征词之间的依赖程度，其值越大，则说明特征词之间的依赖程度越高。将特征依赖度作为特征词之间建立关系的依据。

2.3 语义贝叶斯网的构建

根据语义贝叶斯网的概念构建语义贝叶斯网的基本过程是：1）在短文本集合 D 中，将从短文本中提取的特征词 $T=\{t_1,t_2,\cdots,t_m\}$ 抽象为贝叶斯网中的节点；2）根据特征依赖度 $I(t_i,t_j)$ 和依赖阈值 ε，确定节点之间的边以及方向；3）计算节点的参数。算法 1 具体地描述了短文本的语义贝叶斯网的构建过程。

算法 1 CSBN(Constructing Semantic Bayesian Network)

输入：短文本数据集 $D=\{d_1,d_2,\cdots,d_n\}$，特征提取后得到的特征词集 $T=\{t_1,t_2,\cdots,t_m\}$，特征依赖度阈值 ε

输出：语义贝叶斯网(G=(U,E),P)

步骤 1 初始化语义贝叶斯网的节点集合 U，将特征词集中的特征词加入到集合 U 中，即 $U=\{u_1,u_2,\cdots,u_m\}$，初始化无向边节点集合 E1 和有向边节点集合 E，以及概率集合 P。

步骤 2 针对节点集合 U，依次遍历集合中的任意节点 u_i,u_j，分别统计对应的文档频度 $c(u_i=1,u_j=1),c(u_i=1),c(u_j=1)$。

步骤 3 根据式(3)计算节点之间的依赖度 $I(u_i,u_j)$。

步骤 4 根据依赖度的值，判断 $I(u_i,u_j)$ 是否大于依赖度阈值 ε，若大于，则节点 u_i,u_j 之间存在边，将 u_i-u_j 加入到节点集合 E1 中；否则，转步骤 2。

步骤 5 如果节点 u_i,u_j 之间存在边，则根据文档频度分别计算 $\tau(u_i|u_j)=\frac{c(u_i=1,u_j=1)}{c(u_j=1)}$，$\tau(u_j|u_i)=\frac{c(u_i=1,u_j=1)}{c(u_i=1)}$。若 $\tau(u_i|u_j)>\tau(u_j|u_i)$，则构建边的方向为 $u_j\rightarrow u_i$，否则，$u_i\rightarrow u_j$，将节点以及节点对应的边和方向加入到有向边节点集合 E 中，最后得到有向无环图 G(U,E)。

步骤 6 遍历下一个特征词，重复步骤 2－步骤 5，直到所有特征词组合处理完毕。

步骤 7 遍历 U 中的节点，采用似然估计的方法计算每个节点 u_i 在其父节点 $pa(u_i)$ 条件下的条件概率 $p_{ui}=p(u_i|pa(u_i))$，加入到概率集合 P 中，最后得到语义贝叶斯网(G,P)。

算法 1 在短文本特征提取的基础上，将文本特征抽象为贝叶斯网的节点，根据特征之间的依赖度确定特征词之间是否有相似关系，在有相似关系的特征词中，通过 $\tau(u_i|u_j)$ 和 $\tau(u_j|u_i)$ 来判断特征词之间的依赖程度，若 $\tau(u_i|u_j)>\tau(u_j|u_i)$，则说明 u_i 对 u_j 的依赖程度高于 u_j 对 u_i 的依赖程度，从而确定节点之间边的方向，构建短文本集合的语义贝叶斯网结构 G。在得到结构之后，根据每个节点 u_i 以及对应的父节点集合 $pa(u_i)$，计算出每个节点的参数。最终得到短文本集合 D 上的语义贝叶斯网。

例 1 采用亚马逊评论集中的 Book 数据集，对其中 6 个评论进行特征提取并计算特征频数，结果如表 1 所列。

表 1 文档特征频数

doc	boring	book	horrible	waste	time	bad	life
d1	0	1	1	1	1	0	0
d2	0	1	0	1	0	0	1
d3	0	1	0	0	0	1	0
d4	0	0	0	1	0	0	0
d5	1	0	1	0	0	1	0
d6	1	0	0	0	0	0	1

在此基础上，根据定义 3 计算特征依赖度，并选择合适的依赖度阈值，按照算法 1 首先构建出对应的无向图，再确定边的方向，最后得到的语义贝叶斯网的结构如图 2 所示。

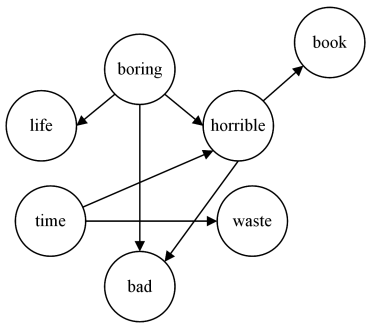

图 2 SBN 的有向无环图结构

3 基于 SBN 的短文本特征扩展

在语义贝叶斯网的基础上，如何准确、高效地找出与短文本特征相关的候选特征是特征扩展的关键，本文利用 SBN 模型及 BN 的推理机制，推理与短文本存在紧密联系的候选特征词，并将其扩展到短文本中，以弥补短文本特征稀疏的不足。

3.1 基于 SBN 的特征推理

BN 的推理就是在给定贝叶斯网中，计算一个我们感兴趣的概率[23]。本文通过 BN 推理找出存在直接或间接联系的特征词，为短文本的特征扩展奠定基础。

BN 的推理可以分为精确推理和近似推理两大类。精确推理算法取决于网络的拓扑结构，所以当网络节点多、连接稠密时，BN 的精确推理效率很低；而此时近似推理就成为提高 BN 推理效率的替代方法。Gibbs 采样是一种常用的近似推理方法，它是一种随机采样方法，在确定证据变量与非证据变量的条件下，首先随机产生一个与证据变量一致的样本作为初始点，然后再从当前样本出发产生下一个样本，依次对非证据变量逐个进行采样，最后近似估算出后验概率。算法 2 给出了基于 Gibbs 采样的 SBN 近似推理过程。

算法 2 Gibbs SBN

输入：语义贝叶斯网 B=(G,P)，证据变量 E(当前特征词节点)，证据变量的取值 e，查询变量(拟推断的特征词节点)Q，查询变量的取值 q，q 值为 1 的样本数 m，采样次数 s

输出：$p(q=1|E=e)$

步骤 1 给 m 赋值为 0；

步骤 2 随机生成一个样本 C，令其状态 $c_0=e\cup q$；

步骤 3 遍历采样次数，根据 B 计算 $P_B(Q_i|C=c_0)$；

步骤 4 通过 $P_B(Q_i|C=c)$ 对 Q_i 进行采样，获取 Q_i 的值，并赋值给 q_i^m；

步骤 5 将 q_i^m 赋值给 q；

步骤 6 若 $q_m=q$，则 $m_q=m_q+1$；

步骤 7　最后得到 p(q=1|E=e)的值为 m_q/s。

根据算法 2 可以得到 $p(q=1|E=e)$的值，即在给定证据变量的条件下，得到待查询变量的后验概率。其中，待查询变量为待扩展特征词节点，证据变量为短文本的特征词节点，通过 Gibbs 采样即得到待扩展节点与短文本中的特征词节点的后验概率，从而得出与短文本特征相关的特征词。

3.2　基于 SBN 的特征扩展

基于 Gibb SBN 可以有效地在 SBN 上对短文中的特征词及其关联程度进行定量地分析和推理，获得特征词之间直接或者间接的概率联系。本文根据这种概率联系研究短文本特征词扩展的方法，将与短文本关联程度大的特征词扩展到短文本中，从而解决短文本特征稀疏的问题。

为了定量描述一个特征与一个短文本之间的关联程度，本文定义关联度的概念。

定义 4(关联度)　对于短文本集合 D，设 $B=(G,P)$是 D 上的一个 SBN，u_k 为 B 中的一个节点，d_i 是任意一个短文本，且其特征词集为 $T_i=\{t_{i1},t_{i2},\cdots,t_{i\alpha}\}$，$t_{ij}$ $(1\leqslant i\leqslant m,1\leqslant j\leqslant\alpha)$为特征词集中的特征词，$u_{ij}$ 为特征词 t_{ij} 对应的节点，u_k 与 d_i 的关联度 $score(u_k,d_i)$定义为：

$$score(u_k,d_i)=\sum_{j=1}^{\alpha}p(u_k=1|u_{ij}=1) \tag{4}$$

关联度通过短文本 d_i 的每个特征条件下特征 t_k 的后验概率来综合度量 t_k 与 d_i 的关联程度，其中 $p(u_k|u_{ij})$通过 Gibb SBN 算法获得。计算出 $score$ 值就可以将与短文本 d_i 联系最紧密的特征扩展到 d_i 的特征中。

下面的算法 3 概括地描述了基于语义贝叶斯网特征选择和扩展的过程。

算法 3　FE(Feature Extension)

输入：语义贝叶斯网 B=(G,P)，短文本数据集 $D=\{d_1,d_2,\cdots,d_n\}$，关联度阈值 θ

输出：扩展后的短文本集 D′

步骤 1　针对语义贝叶斯网中的节点集合 U，依次遍历集合中的节点 t_k；

步骤 2　对短文本数据集 D，依次遍历每条短文本 $d_i(1\leqslant i\leqslant n)$，获取特征词 $T_i=\{t_{i1},t_{i2},\cdots,t_{i\alpha}\}$；

步骤 3　遍历短文本中的特征词 t_{ij}，在 B 上推理 $p(u_k|u_{ij})$；

步骤 4　根据式(4)计算贝叶斯网中的节点特征与短文本的关联度 $score(u_k,d_i)$；

步骤 5　根据关联度的值判断 score 是否大于关联度阈值 θ，如果 $score>\theta$，则将特征词加入到短文本的特征词集 T_i 中，得到新的特征词集 $T_i'=\{t_{i1},t_{i2},\cdots,t_{i\alpha},t_k\}$，否则，转步骤 3；

步骤 6　遍历下一条短文本 d_{i+1}，重复步骤 2－6 直到短文本集 D 中所有文本处理完毕，最后得到扩展后的短文本集 D′。

算法 3 基于贝叶斯推理，计算特征与短文本之间的关联度 $score$，通过关联度阈值 θ 选择与短文本关联密切的特征，并将这些特征词扩展到短文本 d_i 的特征中。

例如，例 1 中短文本特征词集为{boring，book，horrible，waste，time，bad，life}，设经过算法 2 计算得到特征词 sad 与 d_1 的关联度最高，那么将 sad 扩展到 d_1 的特征中，得到特征集合为{boring，book，horrible，waste，time，bad，life，sad}。扩展特征对应的特征值可以取 1，所以 d_1 的特征向量表表示为(0，1，1，1，1，0，0，1)。

4　实验以及实验结果

为了评价本文方法的可行性和有效性，我们采用 Amazon[1)] 评论数据集[24]来评价本文所提方法。Amazon 数据集中每个文本的长度大概为 20～40 个词，属于比较典型的短文本数据。本文在 Amazon 数据集中选取其中的 6 个类别，分别是 Book，DVD，Electric，Kitchen，Music 和 Video，每个类别都有 2000 条数据。

本文采用文本分类作为基本的文本分析任务，以 SVM 作为分类器，分析基于 SBN 的短文本特征扩展对文本特征表示以及文本分析的影响，并将准确率(P)、召回率(R)、F1 值作为评价的主要指标，其公式如式(5)－式(7)所示。

$$precision=\frac{n_{\text{correct}}}{n} \tag{5}$$

$$recall=\frac{n_{\text{correct}}}{N} \tag{6}$$

$$F1=\frac{precision\times recall}{precision+recall} \tag{7}$$

其中，n_{correct}为真正分类正确的文本数，n 为分类器分类为正确的文本数，N 为该类所有文本数。

本文在此数据集的基础上设计了 4 个实验：1)分别比较扩展前后 6 个类别的分类效果；2)通过划分训练集与测试集的比例来比较不同训练集规模下扩展后的分类效果；3)将本文方法与基于 LDA 的扩展方法[11]进行比较；4)分析不同关联度阈值 θ 对本文方法性能的影响。

4.1　扩展前后的分类效果

为了验证本文所提出的特征扩展方法能有效地解决短文本特征稀疏等问题，我们对数据集中的 6 个类别进行特征扩展，并通过分类结果来说明扩展方法的有效性。

在此实验中，将训练集与测试集以 3∶7 和 7∶3 的比例划分，设置关联度阈值 $\theta=0.8$，并使用 SVM 作为分类器。实验结果如表 2 所列，表中的 Average 表示各个类别分类性能的平均值。

表 2　扩展前后性能比较

类别	V=3∶7						V=7∶3					
	扩展前			扩展后			扩展前			扩展后		
	P	*R*	*F1*	*P*	*R*	*F1*	P	R	F1	P	R	F1
Book	0.70	0.68	0.69	0.84	0.78	0.77	0.78	0.78	0.78	0.79	0.77	0.77
DVD	0.76	0.75	0.75	0.80	0.80	0.80	0.79	0.79	0.78	0.87	0.87	0.87
Electric	0.75	0.75	0.75	0.83	0.75	0.74	0.79	0.79	0.71	0.84	0.81	0.80
Kitchen	0.77	0.77	0.77	0.84	0.78	0.77	0.81	0.81	0.81	0.90	0.88	0.88
Music	0.75	0.75	0.75	0.83	0.83	0.82	0.76	0.76	0.76	0.86	0.86	0.85
Video	0.76	0.75	0.75	0.77	0.77	0.77	0.79	0.79	0.79	0.80	0.80	0.80
Average	0.748	0.741	0.743	0.818	0.785	0.778	0.78	0.786	0.771	0.843	0.831	0.828

1) http://www.cs.jhu.edu/_mdredze/datasets/sentiment/

由表 2 可以看出，扩展后对 Book 类别的分类性能提升较多，各性能最大提高了 10%，Video 类相对提升较少，最大提高了 2%。尽管在不同类别，扩展后的性能提升存在着差异，但无论是整体还是单个类别，扩展后的准确率、召回率和 F1 值都有比较明显的提升。并且按照不同比例将训练集和测试集进行划分，各性能也有不同的影响，因此，4.2 节进一步研究了不同规模训练集对分类性能的影响。

4.2 不同训练集规模的扩展效果

为了分析在不同规模的训练集上特征扩展对分类效果的影响，本文在保证其他实验条件相同的条件下，分别将训练集和测试集划分为不同的比例进行测试，对比不同规模的训练集对扩展结果分类效果的影响。图 3－图 5 为不同规模训练集分类得到的准确率、召回率和 F1 值。

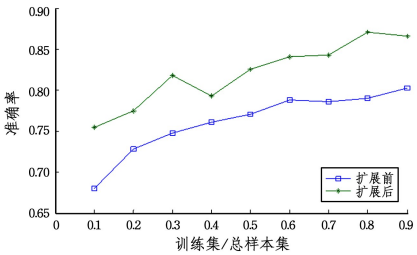

图 3 特征扩展前后的准确率比较

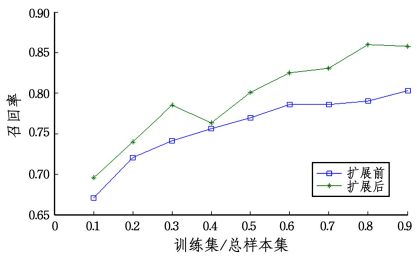

图 4 特征扩展前后的召回率比较

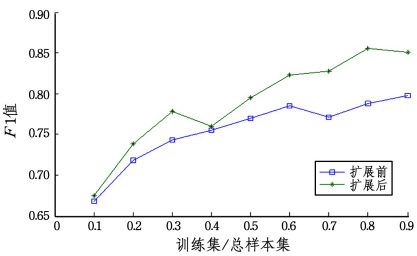

图 5 特征扩展前后的 $F1$ 值比较

由图 3－图 5 可知，训练集规模增大时，特征扩展对分类效果的提升较为明显，无论是准确率、召回率，还是 F1 值都有 2%～10%的变化。总体而言，扩展后的分类性能与训练集的规模是呈线性正相关的，即随着训练样本数的增加，扩展后的分类性能会得到有效的提升，且扩展后的性能上升幅度比扩展前高。实验结果说明，本文提出的基于贝叶斯网的特征扩展可以作为原有特征的有效扩展，并且训练集的增加对扩展后的短文本提升作用更加明显。

4.3 不同扩展方法的效果对比

上文的实验分析说明了本文提出的扩展方法能为短文本提供有效的扩展，弥补短文本特征稀疏的缺陷。为综合评价本文方法，将所提方法与基于 LDA[11] 的特征扩展分类算法进行比较，实验结果如表 3 所列。

表 3 不同方法的性能比较

方法	准确率	召回率	$F1$
LDA	0.75	0.76	0.75
SBN	0.81	0.78	0.77

从表 3 可以看出，虽然两种扩展算法的各项指标都能达到 60%以上，但基于贝叶斯网的特征扩展方法相比基于 LDA 的特征扩展分类算法，在各项指标上都有 1%～2%的提升。

基于 LDA 的短文本特征扩展方法，首先使用一个大文档集来训练 LDA 模型，然后对训练集中的某一篇文档使用训练好的 LDA 模型进行预测，将概率最大的 topic 主题词作为文档的扩展特征。对于干扰特征多的短文本，该方法会将无关的主题词作为扩展特征，因此无法对这类文本进行准确的分类。而对于本文方法，基于 BN 的推理能找到与短文本中特征词直接或间接相关的扩展特征词，即使短文本中区分度较高的特征词较少，也能通过推理扩展大量的相关词汇，从而提高区分文本类别的特征词的影响力。因此，本文方法的实验结果较好。

4.4 关联度阈值的影响

为了测试关联度阈值对扩展结果的影响，我们对数据集中的 Video 类别集选取不同关联度阈值进行特征扩展，并用 SVM 分类器对已扩展文本进行分类，图 6 展示了当 θ 在 0～4 内变化时，短文本分类准确率的变化。由图 6 可知，关联度阈值不同，扩展后的分类效果也不同，关联度阈值过高或者过低，分类效果都不是很理想。关联度阈值过高，则扩展的特征词较少，不能准确地表达短文本的语义信息；而关联度阈值过低，扩展的特征词较多，伴随的噪声也增大，影响实验分类效果。根据图 6 中的结果，关联度阈值 $\theta=0.8$ 时达到最好的分类效果，说明关联度阈值取 0.8 时既能提供有效的扩展特征词，又保证了不扩展入大量无用特征，因此本文实验中选取 $\theta=0.8$ 作为关联度阈值。

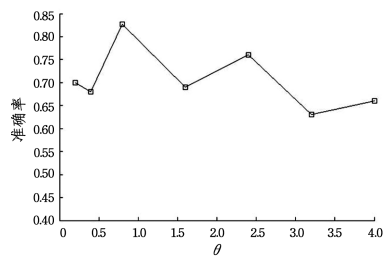

图 6 不同关联度阈值的分类效果

结束语 由图 3－图 5 可知，训练集规模增大时，特征扩展对分类效果的提升较为明显，无论是准确率、召回率，还是 F1 值都有 2%～10%的变化。总体而言，扩展后的分类性能与训练集的规模是呈线性正相关的，即随着训练样本数的增加，扩展后的分类性能会得到有效的提升，且扩展后的性能上升幅度比扩展前高。实验结果说明，本文提出的基于贝叶斯网的特征扩展可以作为原有特征的有效扩展，并且训练集的增加对扩展后的短文本提升作用更加明显。

对短文本进行分析和应用，是目前文本研究的热点之一，短文本特征扩展是解决短文本特征稀疏的一个手段。贝叶斯网是一种不确定性知识建模的有效手段，本文基于贝叶斯网对短文本特征依赖进行建模，定义短文本的语义贝叶斯网，定量表示短文本特征之间的依赖关系，并基于语义贝叶斯网进行语义推理，找到与待扩展短文本相关的特征词，并用其对短文本进行扩展。在 Amazon 数据集上进行实验，结果表明，本

文所提方法具有较好的有效性和可行性。

本文从短文本特征词之间的依赖度出发，基于贝叶斯网进行短文本特征扩展。如何将此方法用于不同领域的短文本是以后要开展的工作。

参 考 文 献

[1] SEVERYN A，MOSCHITTI A. Learning to Rank Short Text Pairs with Convolutional Deep Neural Networks[C]// The International ACM SIGIR Conference. 2015：373-382.

[2] ZHANG W，XUE G R，XUE G R，et al，Advertising Keywords Recommendation for Short-Text Web Pages Using Wikipedia [J]. Acm Transactions on Intelligent Systems & Technology，2012，3(2)：36：1-36：25.

[3] NGUYEN T H，GRISHMAN R. Relation Extraction：Perspective from Convolutional Neural Networks[C]// The Workshop on Vector Space Modeling for Natural Language Processing. 2015：39-48.

[4] MA H，JI Y，LI X，et al. A Microblog Hot Topic Detection Algorithm Based on Discrete Particle Swarm Optimization[C]// Pacific Rim International Conference on Trends in Artificial Intelligence. 2016：271-282.

[5] MA J L，LIU J L，YU C H. An efficient algorithm for Chinese text clustering[J]. Computer Engineering & Science，2013，35(2)：103-108.

[6] 高永兵，钟振华，王宇，等. 基于混合方法的中文微博自动摘要技术研究[J]. 计算机工程与科学，2016，38(6)：1257-1261.

[7] 王仲远，程健鹏，王海勋，等. 短文本理解研究[J]. 计算机研究与发展，2016，53(2)：262-269.

[8] YU Z，WANG H，LIN X，et al. Understanding short texts through semantic enrichment and hashing[J]. IEEE Transactions on Knowledge & Data Engineering，2016，28(2)：566-579.

[9] WANG Y，HUANG H，FENG C. Query Expansion Based on a Feedback Concept Model for Microblog Retrieval[C]// International Conference on World Wide Web. 2017：559-568.

[10] 崔婉秋，杜军平，寇菲菲，等. 面向微博短文本的社交与概念化语义扩展搜索方法[J]. 计算机研究与发展，2018，55(8)：1641-1652.

[11] 吕超镇，姬东鸿，吴飞飞. 基于 LDA 特征扩展的短文本分类[J]. 计算机工程与应用，2015，51(4)：123-127.

[12] XU K，FENG Y，HUANG S，et al. Semantic Relation Classification via Convolutional Neural Networks with Simple Negative Sampling[J]. Computer Science，2015，71(7)：941-949.

[13] SRIRAM B，FUHRY D，DEMIR E，et al. Short text classification in twitter to improve information filtering[C]// International ACM SIGIR Conference on Research and Development in Information Retrieval. 2010：841-842.

[14] ZHANG W，XU W，CHEN G，et al. A Feature Extraction Method Based on Word Embedding for Word Similarity Computing [J]. Communications in Computer & Information Science，2014，496：160-167.

[15] 袁满，欧阳元新，熊璋，等. 一种基于频繁词集的短文本特征扩展方法[J]. 东南大学学报(自然科学版)，2014，44(2)：256-260.

[16] 郭永辉. 面向短文本分类的特征扩展方法[D]. 哈尔滨：哈尔滨工业大学，2013.

[17] MENDES E. Introduction to Bayesian Networks[J]. Medical Imaging Technology，2014，21(2)：1-5.

[18] PEARL J. Probabilistic Reasoning in Intelligent Systems[M]. Morgan Kaufmann Publishers，1988：1022-1027.

[19] YI Z H，WEI W L，XI C Y，et al. Research Progress of Probabilistic Graphical Models：A Survey[J]. Journal of Software，2013，24(11)：2476-2497.

[20] TANG B，KAY S，HE H. Toward Optimal Feature Selection in Naive Bayes for Text Categorization[J]. IEEE Transactions on Knowledge & Data Engineering，2016，28(9)：2508-2521.

[21] 陈为，朱标，张宏鑫. BN-Mapping：基于贝叶斯网络的地理空间数据可视分析[J]. 计算机学报，2016(7)：1281-1293.

[22] 王双成，高瑞，杜瑞杰. 具有超父结点时间序列贝叶斯网络集成回归模型[J]. 计算机学报，2017，40(12)：2748-2761.

[23] HECKERMAN D，DAN G，CHICKERING D M. Learning Bayesian networks：The combination of knowledge and statistical data[J]. Machine Learning，1995，20(3)：197-243.

[24] BLITZER J，DREDZE M，PEREIRA F. Biographies，Bollywood，Boom-boxes and Blenders：Domain Adaptation for Sentiment Classification[C]// Proceedings of ACL'07. 2007.

(上接第 65 页)

[10] 陈卫华，徐国祥. 基于深度学习和股票论坛数据的股市波动率预测精度研究[J]. 管理世界，2018(1)：180-181.

[11] 王国栋，韩斌，孙文赟. 基于 LSTM 的舰船运动姿态短期预测[J]. 舰船科学技术，2017(7)：69-72.

[12] 陆泽楠，商玉林. 基于 LSTM 神经网络模型的钢铁价格预测[J]. 科技视界，2017(13)：116-117.

[13] CAO C Y，LV Q. Using Bidirectional LSTM Deep Neural Network for Protein Residue Contact Prediction [J]. Journal of Chinese Computer Systems，2017(3)：531-535.

[14] KONG D J，TANG S L，WU Fei. Location Prediction via Generative Adversarial Network with Spatial Temporal Embedding [J]. Pattern Recognition and Artificial Intelligence，2018(1)：49-60.

[15] NEIL D，PFEIFFER M，LIU S C. Phased LSTM：Accelerating Recurrent Network Training for Long or Event-based Sequences [C]// Advances in Neural Information Processing Systems 29 (NIPS 2016). 2016.

[16] HUANG Y S，CHOU S Y，YANG Y H. Pop Music Highlighter：Marking the Emotion Keypoints[J]. Transactions of the International Society for Music Information Retrieval，2018，1(1)：68-78.

[17] PERSIO L D，HONCHAR O. Analysis of recurrent neural networks for short-term energy load forecasting[C]// American Institute of Physics Conference Series. 2017.

[18] MIRSAMADI S，BARSOUM E，ZHANG C. Automatic Speech Emotion Recognition Using Recurrent Neural Networks with Local Attention[C]// ICASSP. IEEE，2017.

[19] HOCHREITER S，SCHMIDHUBER J. Long Short-term Memory [J]. Neural Computation，1997，9(8)：1735-1780.