基于特征隶属度的文本分类相似性度量方法

池云仙[1,2] 赵书良[2] 罗 燕[2] 赵骏鹏[2] 高 琳[2] 李 超[2]
(河北师范大学资源与环境科学学院 石家庄 050024)[1]
(河北师范大学数学与信息科学学院 石家庄 050024)[2]

摘 要 基于相似性进行文本分类是当前流行的文本处理方法。基于特征隶属度的文本分类相似性度量方法旨在利用特征与文档间的隶属关系度量文档相似性,从而实现文本分类。该方法基于特征与文档的隶属关系,对特征进行全隶属、偏隶属和无隶属词集划分,并基于3种隶属词集定义隶属度函数。全隶属词集隶属于两篇文档,隶属度随权差增大而降低;偏隶属词集仅隶属于其中某一篇文档,隶属度为一个定值;无隶属词集与两篇文档无隶属关系,隶属度为零。在度量相似性时,偏隶属关系高于全隶属关系。由于同类文档词集相近,异类文档词集差异明显,因此,基于特征与文档的隶属度进行相似性度量,可清晰界定词集与类别的隶属关系,提升分类精度。最后,采用数据集20-Newsgroups和Reuters-21578对分类有效性进行验证,结果表明基于特征隶属度的相似性度量方法的性能优于目前流行的相似性度量方法。

关键词 数据挖掘,文本分类,相似性度量,隶属度
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.11.044

Similarity Measure for Text Classification Based on Feature Subjection Degree

CHI Yun-xian[1,2] ZHAO Shu-liang[2] LUO Yan[2] ZHAO Jun-peng[2] GAO Lin[2] LI Chao[2]
(College of Resources and Environment Science, Hebei Normal University, Shijiazhuang 050024, China)[1]
(College of Mathematic and Information Science, Hebei Normal University, Shijiazhuang 050024, China)[2]

Abstract It is a fashionable method to do text classification based on similarity. Algorithm similarity measure for text classification based on feature subjection degree (SMTCFSD) aims at measuring similarity of documents through subjection relationship between feature words and documents. Algorithm SMTCFSD divides words into total subjection word sets, partial subjection word sets and none subjection word sets based on the subjection relationship, and defines subjection function based on three subjection word sets. Total subjection word sets subject to two documents, and subjection degree will decrease when the differences between two total subjection words increase. The words that only belong to one of the two documents are subsumed into partial subjection word sets, in which subjection degree is a definite value. Subjection degree of none subjection word sets is zero, because the words subject to neither of two documents. Total subjection relationship is more important than partial subjection relationship for similarity measure. Due to word sets of documents in the same category is similar to each other, while the ones in different categories have great distinction, classification accuracy will be promoted obviously based on different values of feature words, which are decided by subjection degree. Algorithm SMTCFSD is superior to the widely used similarity measure methods through experimental results on data sets from Reuters-21578 and 20-Newsgroups.

Keywords Data mining, Text classification, Similarity measure, Subjection degree

---

到稿日期:2016-07-19 返修日期:2016-10-04 本文受国家自然科学资金项目(71271067),国家社科基金重大项目(13&ZD091),河北省高等学校科学技术研究项目(QN2014196),河北师范大学硕士基金(xj2015003)资助。

池云仙(1987-),女,博士生,主要研究领域为数据挖掘、智能信息处理,E-mail:lovebeyond.630@163.com;赵书良(1967-),男,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(通信作者);罗 燕(1993-),女,硕士生,主要研究领域为数据挖掘、智能信息处理,E-mail:luoyan_work@163.com;赵骏鹏(1990-),女,硕士生,主要研究领域为数据挖掘、智能信息处理,E-mail:zhaojunpeng115@126.com;高 琳(1992-),女,硕士生,主要研究领域为数据挖掘、智能信息处理,E-mail:gaol520smile@163.com;李 超(1991-),男,硕士生,主要研究领域为数据挖掘、智能信息处理,E-mail:hbsd_lichao@sina.com。

1 引言

在大数据时代，我们只有具备足够的信息甄别能力，才能从浩瀚的信息海洋中择取有价值的类别。文本分类作为数据挖掘、Web 智能搜索等领域的焦点之一，得到了各界学者的关注[1]。Liu L 等人[2]提出了一种基于聚类的文本分类方法，在构造分类器过程中利用聚类提取可信反例来提升分类精度。Xia R 等人提出了一种评论分类方法 DSA，利用评论的两面性进行双重情绪分析[3]。Yu Z 等人[4]提出了一种通过语意压缩和散列来对短文本进行分类的方法，从知识库中为短文本中的每个词语选取概念和共现词语，进而扩充短文本。Azam N 等人[5]提出了一种基于文本分类中的特征选择度量标准来比较词频和文档频率的方法。Duan J 等人[6]针对多标记分类任务，提出了一种基于粗糙集的文本分类特征选择算法。Tang B 等人[7-8]提出了利用朴素贝叶斯进行文本分类和特征选择的方法。Cheng 等人[9]提出了一种概率性观点挖掘模型 PAMM，用于识别与类标签相关的观点，该模型的独特之处是每次执行所寻找的关联观点时仅涉及一个类别，而不是同时涉及所有类的观点。Tang 等人[10]提出了一种指数嵌入式族的参数化分类器规则 EEF，该方法基于参数分布为每类构造相应的规则，可处理基于数据驱动和模型驱动两种不同方式的分类问题。Liu 等人[11]提出了一种自动式情感主题分类模型，用于解决由于微博情感多样性导致的分类器迁移适应性差的问题。Sun 等人提出了一种通过协同进化多标签超网络进行多标签分类的方法，采用协同进化学习算法训练适用于所有标签的分类模型，利用标签关联性进行多标签分类[12]。

基于相似性实现文本分类是当前流行的文本处理方法。Zhang 等人[13]利用余弦相似性在低维空间计算两个文档的相似度，并在一个相关的相似性度量空间中执行文档聚类。Mori U 等人[14]提出了一种用于时间序列数据库聚类的相似性度量算法。Kang Y B 等人[15]提出了一种基于图的分类算法 TaxoFinder，利用语句间相似性和空间距离度量概念的关联强度对图中的关联度强度进行量化。Wang Q 等人[16]将语义感知用于实体联系模型，将文本和语义特征统一于 ER 阻塞过程，探究相似性度量对 ER 阻塞过程的影响。

为进一步扩充相似性度量在文本分类领域的应用，提出基于特征隶属度的文本分类相似性度量算法，该方法从特征与文档的隶属关系角度出发，依据隶属关系的不同，定义不同的隶属度函数取值，与其他相似性度量方法相比，所提方法能更好地对文档类别进行区分。

本文第 2 节描述基本定义；第 3 节描述数学模型；第 4 节介绍基于特征隶属度的文本分类相似性度量算法；第 5 节介绍基于特征隶属度的文档集合相似性度量方法；第 6 节介绍实验及其结果；最后总结全文。

2 定义

定义 1（全隶属，Total Subjection） 设特征词 Fw 隶属于两篇文本文档 doc_1，doc_2，则称 Fw 为全隶属特征词，表示为 Fw_{Tot}。全隶属特征词集为：

$$FWSet\_TotSub=\{Fw|Fw\in doc_1\cap Fw\in doc_2\} \quad (1)$$

Fw_{Tot} 与 doc_1，doc_2 为全隶属关系，记作 $TotSub(Fw_{Tot}, doc_1, doc_2)$。

定义 2（偏隶属，Partial Subjection） 设有两篇文档 doc_1，doc_2，当且仅当特征词 Fw 隶属于 doc_1 且非隶属于 doc_2，或者 Fw 隶属于 doc_2 且非隶属于 doc_1 时，称 Fw 为偏隶属特征词，记为 Fw_{Par}。偏隶属特征词集为：

$$FWSet\_ParSub=\{Fw|Fw\in doc_1\cup Fw\in doc_2-Fw\in doc_1\cap Fw\in doc_2\} \quad (2)$$

Fw_{Par} 与 doc_1，doc_2 为偏隶属关系，记作 $ParSub(Fw_{Pat}, doc_1, doc_2)$。

定义 3（无隶属，None Subjection） 设有两篇文档 doc_1，doc_2，当且仅当特征词 Fw 既不隶属于 doc_1 又不隶属于 doc_2 时，称 Fw 为无隶属特征词，记为 Fw_{Non}。无隶属特征词集为：

$$FWSet\_NonSub=\{Fw|Fw\notin doc_1\cap Fw\notin doc_2\} \quad (3)$$

Fw_{Non} 与 doc_1，doc_2 为无隶属关系，记作 $NonSub(Fw_{Non}, doc_1, doc_2)$。

3 数学模型

设文档集合 DOC 中包括 N 个文档，对任意文档 $doc_i\in DOC$，$i=1,2,\cdots,N$，利用向量空间模型表示为 $doc_i=(T(doc_i),\omega(doc_i))$，其中 $T(doc_i)=\{Fw_{i1},Fw_{i2},\cdots,Fw_{ik}\}$ 为特征词文本信息，$\omega(doc_i)=\{\omega(Fw_{i1}),\omega(Fw_{i2}),\cdots,\omega(Fw_{ik})\}$ 为特征词对应权重。$\omega(Fw_{ij})(j=1,2,\cdots,k)$ 反映了 Fw_{ij} 与 doc_i 的隶属度。$\omega(Fw_{ij})>0$ 表明 Fw_{ij} 隶属于 doc_i，即 $Fw_{ij}\in doc_i$；$\omega(Fw_{ij})=0$ 表明 Fw_{ij} 非隶属于 doc_i，即 $Fw_{ij}\notin doc_i$。

基于特征隶属度的文本分类相似性度量方法（Similarity Measure for Text Classification based on Feature Subjection Degree，SMTCFSD）用特征词 $Fw_{ij}(i=1,2)$ 与文档 doc_1，doc_2 间 3 种不同的隶属关系来度量 doc_1 和 doc_2 间的相似度 $SD_{SMTCFSD}(doc_1,doc_2)$。

（1）全隶属关系：若 $Fw_{ij}\in FWSet\_TotSub$，即 $Fw_{ij}\in doc_1$ & $Fw_{ij}\in doc_2$，$\omega(Fw_{1j})>0$，$\omega(Fw_{2j})>0$，则基于 Fw_{ij}，$SD_{SMTCFSD}(doc_1,doc_2)$ 随 $|\omega(Fw_{1j})-\omega(Fw_{2j})|$ 的增大而增大。

（2）偏隶属关系：若 $Fw_{ij}\in FWSet\_ParSub$，即 $Fw_{ij}\in doc_1$ & $Fw_{ij}\notin doc_2$ 或 $Fw_{ij}\notin doc_1$ & $Fw_{ij}\in doc_2$，那么 $\omega(Fw_{1j})\omega(Fw_{2j})=0$ & $\omega(Fw_{1j})+\omega(Fw_{2j})>0$，则 $SD_{SMTCFSD}(doc_1,doc_2)$ 基于 Fw_{ij} 为某一定值。

（3）无隶属关系：若 $Fw_{ij}\in FWSet\_NonSub$，即 $Fw_{ij}\notin doc_1$ & $Fw_{ij}\notin doc_2$，$\omega(Fw_{1j})=0$，$\omega(Fw_{2j})=0$，则 $SD_{SMTCFSD}(doc_1,doc_2)$ 基于 Fw_{ij} 为零。

SMTCFSD 算法中，偏隶属关系高于全隶属关系。因为同类文档特征词集具有较高的共性；反之，异类文档特征词具有明显的差异。基于特征隶属度来度量类别间的共性或差异程度可有效提高分类精度。

4 基于特征隶属度的文本分类相似性度量算法

4.1 算法模型

定义 4 文档 doc_1，doc_2 的隶属度（Subjection Degree）函数为：

$$SubDeg(doc_1,doc_2)=\frac{\sum_{j=1}^{k}Dis(\omega(Fw_{1j}),\omega(Fw_{2j}))}{\sum_{j=1}^{k}Nor(\omega(Fw_{1j}),\omega(Fw_{2j}))} \quad (4)$$

其中,互异度(Distinction)函数为:

$$Dis(\omega(Fw_{1j}),\omega(Fw_{2j}))=\begin{cases}\exp\{-(\omega(Fw_{1j})-\omega(Fw_{2j}))^2/(1/\pi)\}, & Fw_{1j}\in FWSet\_TotSub\\ -\varepsilon, & Fw_{1j}\in FWSet\_ParSub\\ 0, & Fw_{1j}\in FWSet\_NonSub\end{cases} \quad (5)$$

规范化(Normalization)函数表示文档中非零特征的个数:

$$Nor(\omega(Fw_{1j}),\omega(Fw_{2j}))=\begin{cases}1, & Fw_{1j}\in FWSet\_TotSub\cup FWSet\_ParSub\\ 0, & Fw_{1j}\in FWSet\_NonSub\end{cases} \quad (6)$$

基于特征隶属度的文档相似度函数为:

$$SD_{SMTCFSD}(doc_1,doc_2)=\frac{SubDeg(doc_1,doc_2)+\varepsilon}{1+\varepsilon} \quad (7)$$

4.2 算法性质

为验证下述性质,假设在 doc_1,doc_2 中,除讨论的特征词外,其他特征词的相应权值均相同。

性质 1　偏隶属关系高于全隶属关系。

设有两个文档 doc_1,doc_2,以及偏隶属特征词 $Fw_{Par}\in FWSet\_ParSub$ 和全隶属特征词 $Fw_{Tot}\in FWSet\_TotSub$。假设偏隶属特征词 Fw_{Par} 隶属于 doc_1 非隶属于 doc_2,则 Fw_{Par} 与 doc_1 有关联,与 doc_2 没有关联,即基于 Fw_{Par} 而言,doc_1 和 doc_2 是不相似性的。全隶属特征词 Fw_{Tot} 同时隶属于 doc_1 和 doc_2,与二者均有一定关联,即基于 Fw_{Tot} 而言,doc_1 和 doc_2 是相似性的。即:

$$\frac{P+\exp\{-(\omega(Fw_{Tot})_{1p}-\omega(Fw_{Tot})_{2p})^2/(1/\pi)\}}{Q}>\frac{P-\varepsilon}{Q}$$

其中,$Q=\sum_{j=1}^{k}Nor(\omega(Fw_{1j}),\omega(Fw_{2j}))$,$P=\sum_{j=1,j\neq p}^{k}Dis(\omega(Fw_{1j}),\omega(Fw_{2j}))$。因此 $SubDeg(doc_1,doc_2)_{Tot}>SubDeg(doc_1,doc_2)_{Par}$,故,$SD_{SMTCFSD}(doc_1,doc_2)_{Tot}>SD_{SMTCFSD}(doc_1,doc_2)_{Par}$。

由此可知,基于偏隶属特征词导致的相似度下降程度高于基于全隶属特征词权差导致的相似度下降程度。偏隶属关系高于全隶属关系。

性质 2　基于全隶属特征词,相似度随特征权差的增大而降低。

设有文档 doc_1,doc_2,以及全隶属特征词 $Fw_{Tot1}\in FWSet\_TotSub$。若 Fw_{Tot1} 的权差增大,即 $|\omega(Fw_{Tot})_{1j}-\omega(Fw_{Tot})_{2j}|$ 增大,从而导致 $\exp\{-(\omega(Fw_{Tot})_{1j}-\omega(Fw_{Tot})_{2j})^2/(1/\pi)\}$ 减小,那么 $\frac{P+\exp\{-(\omega(Fw_{Tot})_{1j}-\omega(Fw_{Tot})_{2j})^2/(1/\pi)\}}{Q}$ 减小,即隶属度函数 $SubDeg(doc_1,doc_2)$ 减小,从而相似性度量函数 $SD_{SMTCFSD}(doc_1,doc_2)$ 减小。由此可知,相似度随全隶属特征词权差的增大而减小。

性质 3　相似度随偏隶属特征词数的增多而降低。

设有有 3 个文档 doc_1,doc_2,doc_3,以及偏隶属特征词 $Fw_{Par}\in FWSet\_ParSub$。假设 doc_1,doc_2,doc_3 共有 m 个全隶属特征词,doc_1,doc_2 之间有 a 个偏隶属特征词,doc_1,doc_3 之间有 b 个偏隶属特征词,并且 doc_1,doc_2 之间有 a 个偏隶属特征词也隶属于 doc_1,doc_3,即 $a\leqslant b$。于是有:$SubDeg(doc_1,doc_2)=\frac{m-\varepsilon a}{m+a}$,$SubDeg(doc_1,doc_3)=\frac{m-\varepsilon b}{m+b}$。

因为 $m-\varepsilon a\geqslant m-\varepsilon b$,$m+a\leqslant m+b$,所以 $SubDeg(doc_1,doc_2)\geqslant SubDeg(doc_1,doc_3)$。因此,$SD_{SMTCFSD}(doc_1,doc_2)\geqslant SD_{SMTCFSD}(doc_1,doc_3)$。

由此可知,相似度随偏隶属特征词数的增多而降低。

性质 4　相似度值域为[0,1]。

由性质 3 可得相似度随偏隶属特征词数的增多而降低,那么当全部非零特征均为偏隶属特征词时,相似度最小。

假设有两个文档 doc_1,doc_2,以及偏隶属特征词 $Fw_{Par}\in FWSet\_ParSub$。设 doc_1 中有 p 个非零特征词 $doc_1=\langle Fw_1,\cdots,Fw_p,0,\cdots,0\rangle$,$doc_2$ 中有 q 个非零特征词,且这些非零特征词均为偏隶属特征词,即 $doc_1=\langle Fw_1,\cdots,Fw_p,0,\cdots,0\rangle$,$doc_2=\langle 0,\cdots,0,Fw_{p+1},\cdots,Fw_{p+q},0,\cdots,0\rangle$。那么 $SubDeg(doc_1,doc_2)=\frac{-\varepsilon p-\varepsilon q}{p+q}=-\varepsilon$,此时 $\min\_SD_{SMTCFSD}(doc_1,doc_2)=0$。

由性质 2 可得相似度随全隶属特征词权差的减小而增大。由此可推出,若全部非零特征均为全隶属特征词且权值均相等,则相似度最大。

设文档 doc_1,doc_2 中有 p 个权重相等的全隶属特征词 $Fw_{Toti}\in FWSet\_TotSub(i=1,2,\cdots,p)$,且无其他非零特征词,即 $doc_1=\langle Fw_1,\cdots,Fw_p,0,\cdots,0\rangle$,$doc_2=\langle Fw_1,\cdots,Fw_p,0,\cdots,0\rangle$,则 $SubDeg(doc_1,doc_2)=p/p=1$,此时 $\max\_SD_{SMTCFSD}(doc_1,doc_2)=1$。

由此可知,相似度值域为[0,1]。

性质 5　相似度是对称的。

因为隶属度函数的定义与特征词顺序无关,即 $Dis(\omega(Fw_{1j}),\omega(Fw_{2j}))=Dis(\omega(Fw_{2j}),\omega(Fw_{1j}))$,$Nor(\omega(Fw_{1j}),\omega(Fw_{2j}))=Nor(\omega(Fw_{2j}),\omega(Fw_{1j}))$,所以 $SubDeg(doc_1,doc_2)=SubDeg(doc_2,doc_1)$,$SD_{SMTCFSD}(doc_1,doc_2)=SD_{SMTCFSD}(doc_2,doc_1)$。

由此可知,相似度是对称的。

4.3 算法伪代码

算法 1　基于特征隶属度的相似性度量方法 SMTCFSD

```
INPUT:doc1={ω(Fw11),ω(Fw12),…,ω(Fw1k)}
      doc2={ω(Fw21),ω(Fw22),…,ω(Fw2k)}
      ε:常数
OUTPUT:相似度 SDSMTCFSD(doc1,doc2)
METHOD:
1. SDSMTCFSD(doc1,doc2)=0
2. SubDeg(doc1,doc2)=0
3. Dis(ω(Fw1j),ω(Fw2j))=0
4. Nor(ω(Fw1j),ω(Fw2j))=0
5. FOR j=1 TO k /*求取隶属度*/
6.    IF Fw1j∈doc1 &Fw2j∈doc2
7.       IF ω(Fw1j)ω(Fw2j)>0 /*全隶属*/
8.          Dis(ω(Fw1j),ω(Fw2j))+=exp{-(ω(Fw1j)-ω(Fw2j))2/
            (1/π)}
```

9. Nor($\omega(Fw_{1j})$,$\omega(Fw_{2j})$)+=1
10. ELSE IF $\omega(Fw_{1j})=0$&$\omega(Fw_{2j})=0$ /*无隶属*/
11. Dis($\omega(Fw_{1j})$,$\omega(Fw_{2j})$)+=0
12. Nor($\omega(Fw_{1j})$,$\omega(Fw_{2j})$)+=0
13. ELSE /*偏隶属*/
14. Dis($\omega(Fw_{1j})$,$\omega(Fw_{2j})$)+=$-\varepsilon$
15. Nor($\omega(Fw_{1j})$,$\omega(Fw_{2j})$)+=1
16. END IF
17. END IF
18. END IF
19. END FOR

20. $\mathrm{SubDeg}(doc_1,doc_2)=\dfrac{\sum_{j=1}^{k}\mathrm{Dis}(\omega(Fw_{1j}),\omega(Fw_{2j}))}{\sum_{j=1}^{k}\mathrm{Nor}(\omega(Fw_{1j}),\omega(Fw_{2j}))}$

21. $\mathrm{SD}_{\mathrm{SMTCFSD}}(doc_1,doc_2)=\dfrac{\mathrm{SubDeg}(doc_1,doc_2)+\varepsilon}{1+\varepsilon}$

22. RETURN $\mathrm{SD}_{\mathrm{SMTCFSD}}(doc_1,doc_2)$

5 基于特征隶属度的文档集合相似性度量方法

定义 5 集合 DS_1,DS_2 分别包含 h_1,h_2 个文档,$DS_1=\{doc_{11},doc_{12},\cdots,doc_{1h_1}\}$,$DS_2=\{doc_{21},doc_{22},\cdots,doc_{2h_2}\}$,其中,$doc_{ij}=\{\omega(Fw_{ij1}),\omega(Fw_{ij2}),\cdots,\omega(Fw_{ijp})\}$。文档集合 DS_1,DS_2 的隶属度函数为:

$$\begin{aligned}&SubDeg(DS_1,DS_2)\\&=\frac{\sum_{i=1}^{h_1}\sum_{j=1}^{h_2}\sum_{k=1}^{p}Dis(\omega(Fw_{1ik}),\omega(Fw_{2jk}))}{\sum_{i=1}^{h_1}\sum_{j=1}^{h_2}\sum_{k=1}^{m}Nor(\omega(Fw_{1ik}),\omega(Fw_{2jk}))}\\&=\frac{\sum_{k=1}^{p}\sum_{i=1}^{h_1}\sum_{j=1}^{h_2}Dis(\omega(Fw_{1ik}),\omega(Fw_{2jk}))}{\sum_{k=1}^{p}\sum_{i=1}^{h_1}\sum_{j=1}^{h_2}Nor(\omega(Fw_{1ik}),\omega(Fw_{2jk}))}\end{aligned}\tag{8}$$

文档集合相似度为:

$$SDS_{SMTCFSD}(DS_1,DS_2)=\frac{SubDeg(DS_1,DS_2)+\varepsilon}{1+\varepsilon}\tag{9}$$

对于某特征词 Fw_k:

$$\begin{aligned}&\sum_{i=1}^{h_1}\sum_{j=1}^{h_2}Dis(\omega(Fw_{1ik}),\omega(Fw_{2jk}))\\&=\sum_{\omega(Fw_{1ik})>0}\sum_{\omega(Fw_{2jk})>0}Dis(\omega(Fw_{1ik}),\omega(Fw_{2jk}))+\\&\quad\sum_{\omega(Fw_{1ik})=0}\sum_{\omega(Fw_{2jk})>0}Dis(\omega(Fw_{1ik}),\omega(Fw_{2jk}))+\\&\quad\sum_{\omega(Fw_{1ik})>0}\sum_{\omega(Fw_{2jk})=0}Dis(\omega(Fw_{1ik}),\omega(Fw_{2jk}))+\\&\quad\sum_{\omega(Fw_{1ik})=0}\sum_{\omega(Fw_{2jk})=0}Dis(\omega(Fw_{1ik}),\omega(Fw_{2jk}))\\&=\sum_{\omega(Fw_{1ik})>0}\sum_{w(p_{jk}^2)>0}\exp\{-(\omega(Fw_{1j})-\omega(Fw_{2j}))^2/(1/\\&\quad\pi)\}+(-\varepsilon)\times(h_1-\sum_{i=1}^{h_1}\mathrm{sgn}(\omega(Fw_{1ik})))\sum_{j=1}^{h_2}\mathrm{sgn}(\omega\\&\quad(Fw_{2jk}))+(-\varepsilon)\times(\sum_{i=1}^{h_1}\mathrm{sgn}(\omega(Fw_{1ik})))(h_2-\sum_{j=1}^{h_2}\\&\quad\mathrm{sgn}(\omega(Fw_{2jk})))\end{aligned}\tag{10}$$

$$\begin{aligned}&\sum_{i=1}^{h_1}\sum_{j=1}^{h_2}Nor(\omega(Fw_{1ik}),\omega(Fw_{2jk}))\\&=\sum_{i=1}^{h_1}\mathrm{sgn}(\omega(Fw_{1ik}))\sum_{j=1}^{h_2}\mathrm{sgn}(\omega(Fw_{2jk}))+(h_1-\sum_{i=1}^{h_1}\mathrm{sgn}(\omega\\&\quad(Fw_{1ik})))\sum_{j=1}^{h_2}\mathrm{sgn}(\omega(Fw_{2jk}))+(\sum_{i=1}^{h_1}\mathrm{sgn}(\omega(Fw_{1ik})))\\&\quad(h_2-\sum_{j=1}^{h_2}\mathrm{sgn}(\omega(Fw_{2jk})))\end{aligned}\tag{11}$$

其中,

$$\mathrm{sgn}(a)=\begin{cases}1, & a>0\\0, & a\leqslant 0\end{cases}\tag{12}$$

当 $h_1=h_2=1$ 时,式(8)等价于式(4)。当 $h_1=1$,$h_2>1$ 时,式(8)度量某文档与各类集合的相似度,可用于文档分类。

6 实验结果

6.1 数据集

文本挖掘常用的公共语料库为 20-Newsgroups 和 Reuters-21578。

数据集 20-Newsgroups 包含 4 个大类:comp,rec,sci 和 talk,分别包含的文档数目为 1162 篇、1190 篇、1183 篇和 975 篇。通常采用一对一分类方法。

数据集 Reuters-21578 通常选取 65 个主题中的前 8 个类别:acq,crude,earn,grain(包含 corn 和 wheat),interest,money,ship,trade,分别包含的文档数目为 1659 篇、405 篇、2775 篇、773 篇、335 篇、502 篇、200 篇、340 篇。因此,Reuters-21578 常被称为"R8"[17]。

选取文档总数的 70%作为训练集,剩余文档用作测试集。

6.2 评价标准

(1)基本评价指标

分类准确率:

$$Acc=\frac{TP+TN}{TP+FN+FP+TN}\tag{13}$$

精确率:

$$P=\frac{TP}{TP+FP}\tag{14}$$

召回率:

$$R=\frac{TP}{TP+FN}\tag{15}$$

F_1 度量值:

$$F_1=\frac{2PR}{P+R}\tag{16}$$

其中,TP 为文档被正确判为类 C 的数目;FP 为文档被错判为类 C 的数目;FN 为文档被错判为非类 C 的数目;TN 为文档被正确判为非类 C 的数目。

(2)宏平均值

宏平均准确率:

$$MacroAcc=\sum_{i=1}^{|C|}Acc_i/|C|\tag{17}$$

宏平均精确率:

$$MacroP=\sum_{i=1}^{|C|}P_i/|C|\tag{18}$$

宏平均召回率:

$$MacroR=\sum_{i=1}^{|C|}R_i/|C|\tag{19}$$

宏平均 F_1 度量值:

$$MacroF_1=\frac{2\cdot MacroP\cdot MacroR}{MacroP+MacroR}\tag{20}$$

6.3　相似性度量方法

为了验证相似性度量方法 SMTCFSD 的有效性，选取目前被广泛应用于文本分类的经典相似性度量方法进行比较：

欧氏距离[18]：

$$DIS_{Euc}(doc_1,doc_2)=[(doc_1-doc_2)\cdot(doc_1-doc_2)]^{\frac{1}{2}} \tag{21}$$

杰卡德相似度[19]：

$$S_{Jac}(doc_1,doc_2)=\frac{doc_1\cdot doc_2}{doc_1\cdot doc_1+doc_2\cdot doc_2-doc_1\cdot doc_2} \tag{22}$$

余弦相似性[20]：

$$S_{Cos}(doc_1,doc_2)=\frac{doc_1\cdot doc_2}{(doc_1\cdot doc_1)^{1/2}(doc_2\cdot doc_2)^{1/2}} \tag{23}$$

6.4　实验结果

6.4.1　分类方法中参数 ε 的选取

在数据集 20-Newsgroups 和 Reuters-21578 上，应用 KNN 算法进行相似性度量，K 分别取 1，3，5，7，9，11。设置 ε=0.001，0.01，0.05，0.1，0.2，0.4，0.6，0.8，1.0。图 1 和图 2 给出了分类精度变化趋势。随着 ε 取值的增大，算法 SMTCFSD 的分类精度逐渐增高。对于同类文档，特征词隶属于两篇文档的情况多于仅隶属于某一篇文档的情况；但对于异类文档，多数情况下特征词语仅隶属于某一篇文档。因此，当 ε 取较大值时，同类文档中偏隶属词语较少，相似度受偏隶属词集的影响较小，相似度较高；异类文档中偏隶属词语较多，ε 取较大值时会显著降低文档间的相似度，从而可有效地区分不同类别。因此，ε 值较大时更利于分类，设置 ε=1.0。

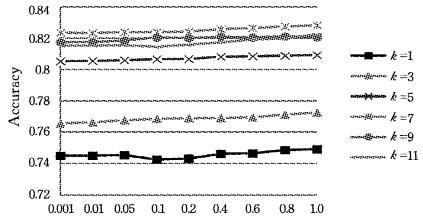

图 1　在数据集 20-Newsgroups 上 SMTCFSD 选取不同 ε 值时所得的分类精度

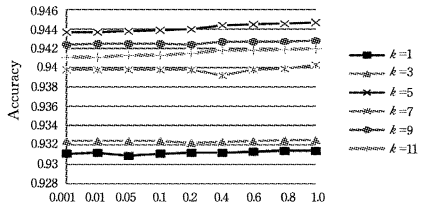

图 2　在数据集 Reuters-21578 上 SMTCFSD 选取不同 ε 值时所得的分类精度

6.4.2　与经典相似性度量算法的分类性能比较

6.4.2.1　分类性能

表 1 和表 2 列出了在数据集 20-Newsgroups 上 SMTCFSD 算法与经典相似性度量算法采用 χ^2 和 MI 特征选择方法所得的分类性能的对比结果。

表 1　在 20-Newsgroups 数据集上 SMTCFSD 算法与经典相似性度量算法采用 χ^2 特征选择方法的分类性能/%

20-Newsgroups Data Sets	Model	Acc	*PR*	F_1
comp vs rec	SMTCFSD	**89.37**	**90.57/87.61**	**89.07**
	Euclidean	78.19	79.00/76.08	77.51
	Jaccard	82.44	83.47/80.38	81.90
	Cosine	80.82	81.71/78.83	80.24
comp vs sci	SMTCFSD	**78.46**	**79.57/76.08**	**77.79**
	Euclidean	67.21	67.40/65.49	66.43
	Jaccard	70.96	71.15/69.62	70.38
	Cosine	69.30	69.42/67.99	68.70
comp vs talk	SMTCFSD	**96.16**	**97.20/95.70**	**96.44**
	Euclidean	85.40	87.48/85.37	86.41
	Jaccard	90.22	91.91/89.93	90.91
	Cosine	87.79	90.04/87.18	88.59
rec vs sci	SMTCFSD	**77.67**	**80.11/73.78**	**76.81**
	Euclidean	66.79	67.36/65.55	66.44
	Jaccard	71.47	71.94/70.67	71.30
	Cosine	68.94	69.38/68.15	68.76
rec vs talk	SMTCFSD	**82.77**	**86.25/81.68**	**83.90**
	Euclidean	75.80	79.68/75.13	77.34
	Jaccard	80.51	84.16/79.50	81.76
	Cosine	78.38	82.18/76.72	79.36
sci vs talk	SMTCFSD	**77.20**	**80.99/76.33**	**78.59**
	Euclidean	66.91	71.18/66.61	68.82
	Jaccard	71.46	75.65/70.67	73.08
	Cosine	69.14	73.18/68.98	71.02
宏平均	SMTCFSD	**83.61**	**85.78/81.86**	**83.77**
	Euclidean	73.38	75.35/72.37	73.83
	Jaccard	77.84	79.71/76.80	78.23
	Cosine	75.73	77.65/74.64	76.12

表 2　在 20-Newsgroups 数据集上 SMTCFSD 算法与经典相似性度量算法采用 MI 特征选择方法的分类性能/%

20-Newsgroups Data Sets	Model	Acc	*PR*	F_1
comp vs rec	SMTCFSD	**88.14**	**89.38/86.23**	**87.78**
	Euclidean	75.58	77.30/72.38	74.76
	Jaccard	80.14	81.91/76.76	79.25
	Cosine	78.49	80.09/75.13	77.53
comp vs sci	SMTCFSD	**75.95**	**76.32/74.61**	**75.46**
	Euclidean	66.44	66.44/64.63	65.62
	Jaccard	70.70	71.07/68.93	69.98
	Cosine	69.96	69.27/67.13	68.18
comp vs talk	SMTCFSD	**95.09**	**96.82/94.32**	**95.55**
	Euclidean	83.20	85.75/82.87	84.29
	Jaccard	88.02	90.23/87.44	88.81
	Cosine	86.10	88.44/85.63	87.01
rec vs sci	SMTCFSD	**73.91**	**76.22/69.75**	**72.84**
	Euclidean	65.86	67.15/62.52	64.75
	Jaccard	70.54	71.94/67.65	69.73
	Cosine	68.48	69.56/66.05	67.76
rec vs talk	SMTCFSD	**81.80**	**85.28/80.84**	**83.00**
	Euclidean	72.98	76.56/73.28	74.88
	Jaccard	77.51	80.91/77.31	79.07
	Cosine	75.19	78.97/74.79	76.82
sci vs talk	SMTCFSD	**74.79**	**78.55/74.30**	**76.37**
	Euclidean	65.29	69.20/66.10	67.61
	Jaccard	70.76	74.47/71.01	72.70
	Cosine	68.03	71.83/68.55	70.15
宏平均	SMTCFSD	**81.61**	**83.76/80.01**	**81.84**
	Euclidean	71.56	73.77/70.30	71.86
	Jaccard	76.28	78.42/74.85	76.59
	Cosine	74.38	76.36/72.88	74.58

由表可知，在各子类中 SMTCFSD 的分类性能均优于目前广泛采用的经典相似性度量方法。从宏平均值可得出，相较于性能最好的 Jaccard 方法，SMTCFSD 采用 χ^2 特征选择方法时，Acc、PR 值、F_1 度量值分别高出 5.77%，6.07%/5.06%，5.54%；采用 MI 特征选择方法时，Acc、PR 值、F_1 度量值分别高出 5.33%，5.34%/5.16%，5.25%。

表 3 和表 4 列出了在数据集 Reuters-21578 上 SMTCFSD 算法与经典相似性度量算法采用 χ^2 和 MI 特征选择方法所得的分类性能的对比结果。由表可知，仅在子类 crude 中 Jaccard 采用 χ^2 方法的分类性能优于 SMTCFSD，这是由于 crude 类中特征词语也隶属于其他类别的非关联文档，因此无法发挥 SMTCFSD 模式的优势。除此之外，SMTCFSD 的分类性能均优于目前广泛采用的经典相似性度量方法。从宏平均值得出，相较性能最好的 Jaccard，SMTCFSD 采用 χ^2 特征选择方法时，Acc、PR 值、F_1 度量值分别高出 1.94%，9.82%/8.99%，9.39%；采用 MI 特征选择方法时，Acc、PR 值、F_1 度量值分别高出 1.2%，6.49%/8.47%，7.63%。

表 3 在 Reuters-21578 数据集上 SMTCFSD 算法与经典相似性度量算法采用 χ^2 特征选择方法的分类性能/%

Reuters-21578 Data Sets	Model	Acc	*PR*	F_1
acq	SMTCFSD	**91.94**	**83.83/81.83**	**82.82**
	Euclidean	79.97	58.01/56.62	57.31
	Jaccard	88.67	77.56/73.52	75.49
	Cosine	87.03	74.32/69.30	71.72
crude	SMTCFSD	94.25	50.31/47.40	48.81
	Euclidean	92.95	38.82/38.15	38.48
	Jaccard	**95.12**	**58.82/52.02**	**55.21**
	Cosine	94.48	52.60/46.82	49.54
earn	SMTCFSD	**89.67**	**88.13/85.53**	**86.81**
	Euclidean	75.43	69.47/68.12	69.79
	Jaccard	83.99	83.43/74.52	78.72
	Cosine	83.38	82.52/73.84	77.94
grain	SMTCFSD	**93.85**	**74.58/66.87**	**70.51**
	Euclidean	89.60	52.90/49.85	51.33
	Jaccard	92.91	68.59/65.05	66.77
	Cosine	92.44	66.14/64.13	65.12
interest	SMTCFSD	**95.19**	**49.64/47.55**	**48.57**
	Euclidean	93.25	29.08/28.67	28.87
	Jaccard	93.65	32.56/30.77	31.65
	Cosine	93.55	31.34/29.37	30.32
money	SMTCFSD	**94.52**	**62.69/58.60**	**60.58**
	Euclidean	91.34	39.11/36.74	37.89
	Jaccard	92.28	46.33/46.98	46.65
	Cosine	91.94	43.81/42.79	43.29
ship	SMTCFSD	**97.26**	**52.56/47.67**	**50.00**
	Euclidean	95.75	25.88/25.58	25.73
	Jaccard	96.12	32.14/31.40	31.77
	Cosine	96.09	30.38/27.91	29.09
trade	SMTCFSD	**95.32**	**51.94/45.89**	**48.73**
	Euclidean	93.11	29.17/28.77	28.97
	Jaccard	93.75	35.66/34.93	35.29
	Cosine	93.35	31.47/30.82	31.14
宏平均	SMTCFSD	**94.00**	**64.21/60.14**	**62.11**
	Euclidean	88.93	42.81/41.56	42.18
	Jaccard	92.06	54.39/51.15	52.72
	Cosine	91.53	51.57/48.12	49.79

表 4 在 Reuters-21578 数据集上 SMTCFSD 算法与经典相似性度量算法采用 MI 特征选择方法的分类性能/%

Reuters-21578 Data Sets	Model	Acc	*PR*	F_1
acq	SMTCFSD	**88.30**	**76.32/73.52**	**74.86**
	Euclidean	77.10	51.84/49.58	50.71
	Jaccard	87.23	74.26/70.70	72.44
	Cosine	85.22	69.82/66.48	68.11
crude	SMTCFSD	**94.52**	**52.83/48.55**	**50.60**
	Euclidean	91.94	30.00/29.48	29.74
	Jaccard	93.92	47.13/42.77	44.84
	Cosine	93.41	42.59/39.88	41.19
earn	SMTCFSD	**86.86**	**84.49/82.00**	**83.23**
	Euclidean	72.52	66.11/63.33	64.69
	Jaccard	84.19	84.23/74.10	78.84
	Cosine	82.58	80.42/74.26	77.22
gain	SMTCFSD	**93.75**	**72.90/68.69**	**70.73**
	Euclidean	86.26	37.35/36.79	37.07
	Jaccard	92.11	66.20/57.75	61.69
	Cosine	90.77	58.63/54.71	56.60
interest	SMTCFSD	**94.62**	**43.57/42.66**	**43.11**
	Euclidean	93.18	28.37/27.97	28.17
	Jaccard	93.98	34.45/28.67	31.30
	Cosine	93.28	28.99/27.97	28.47
money	SMTCFSD	**93.35**	**53.81/52.56**	**53.18**
	Euclidean	91.54	40.95/40.00	40.47
	Jaccard	91.84	42.93/40.93	41.91
	Cosine	91.64	41.63/40.47	41.04
ship	SMTCFSD	**96.76**	**42.86/38.37**	**40.49**
	Euclidean	95.89	28.24/27.91	28.07
	Jaccard	96.19	33.33/32.56	32.94
	Cosine	96.32	34.62/31.40	32.93
trade	SMTCFSD	**94.55**	**44.06/43.15**	**43.60**
	Euclidean	93.21	30.07/29.45	29.76
	Jaccard	93.68	34.97/34.25	34.61
	Cosine	93.25	30.56/30.14	30.35
宏平均	SMTCFSD	**92.84**	**58.68/56.19**	**57.49**
	Euclidean	87.71	39.11/38.06	38.58
	Jaccard	91.64	52.19/47.72	49.86
	Cosine	90.81	48.41/45.66	46.99

6.4.2.2 PR 曲线

图 3 和图 4 为数据集 20-Newsgroups 上的 PR 曲线。

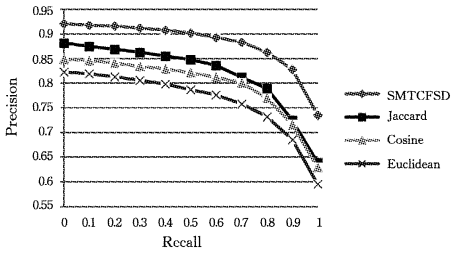

图 3 在数据集 20-Newsgroups 上 χ^2 特征选择方法用于 SMTCFSD 算法时与经典相似性度量算法所得的 PR 曲线对比

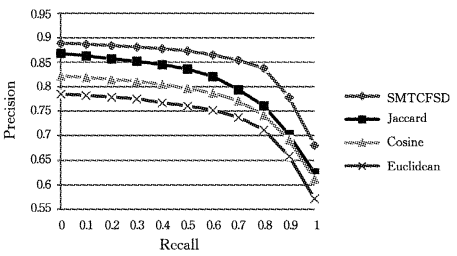

图 4 在数据集 20-Newsgroups 上 MI 特征选择方法用于 SMTCFSD 算法时与经典相似性度量算法所得的 PR 曲线对比

图 5 和图 6 为数据集 Reuters-21578 上的 PR 曲线。由图可知,相比于其他基于经典相似性度量的分类方法,SMTCFSD 分类方法的分类效果明显增高。

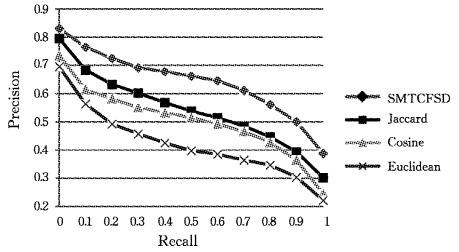

图 5　在 Reuters-21578 数据集上 χ^2 特征选择算法时用于 SMTCFSD 算法与经典相似性度量算法所得的 PR 曲线对比

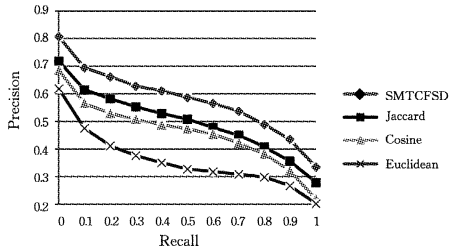

图 6　在 Reuters-21578 数据集上 MI 特征选择算法用于 SMTCFSD 算法时与经典相似性度量算法所得的 PR 曲线对比

6.4.3　与新分类方法的分类性能比较

目前,关于文本分类的新方法层出不穷。文献[7]基于信息论提出了两种面向文本分类的朴素贝叶斯最优特征选择方法 MD 和 MD-χ^2,旨在依据特征对类别的区分力进行排序和选择,以此进行文本分类。文献[8]提出了基于类依赖的文本分类方法 PPT,该方法遵循 PDF 投影定理,利用类依赖特征自动进行文本分类。本节将 SMTCFSD 与上述 3 种新的分类方法进行对比,对比结果如图 7 和图 8 所示。为了对各分类方法的分类准确率进行评估,特征的选择范围选定为[10,2000]。

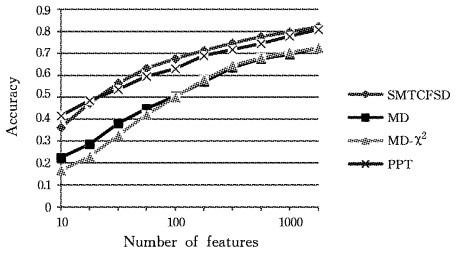

图 7　在数据集 20-Newsgroups 上 SMTCFSD 算法与 3 种新分类算法的分类准确率对比

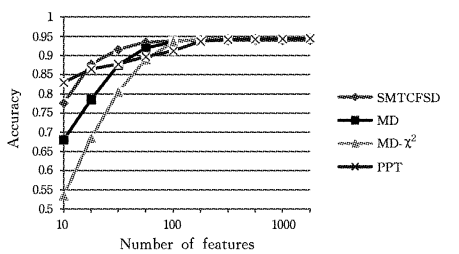

图 8　在数据集 Reuters-21578 上 SMTCFSD 算法与 3 种新分类算法的分类准确率对比

从图中可以看出,随着所选特征数目的增多,各分类方法的分类性能均逐渐提高。图 7 为在数据集 20-Newsgroups 上 SMTCFSD 算法与 3 种新分类算法的分类准确率对比,SMTCFSD 算法的分类准确率高于另外 3 种分类方法,并且 SMTCFSD 算法选择前 200 个特征时就能获得与分类方法 MD 和 MD-χ^2 选择前 1000 个特征时相近的性能;图 8 为在数据集 Reuters-21578 上 SMTCFSD 算法与 3 种新分类算法的分类准确率对比,当所选特征数目增多时,SMTCFSD 算法与另外 3 种分类方法的分类准确率相近,但 SMTCFSD 算法在所选特征数目较少时,其分类性能均高于其他 3 种分类方法,从而说明当所选特征数目受限时,基于特征与文档间的不同隶属关系赋予不同权重,可以更加有效地提升分类精度。

综上所述,SMTCFSD 算法基于特征与文档的隶属度对特征进行划分和加权,能够使词集与类别间的隶属关系更加清晰,显著提高分类性能。

6.4.4　运行效率

SMTCFSD 算法的时间复杂度为 $O(mn)$,其中,特征权重计算的时间复杂度为 $O(m)$,文档相似度计算的时间复杂度为 $O(n)$。分类方法 MD,MD-χ^2 的时间复杂度为 $O(mn)$,PPT 算法的时间复杂度为 $O(n^2)$。经典相似性度量算法 Euclidean,Jaccard 和 Cosine 的时间复杂度均为 $O(mn)$。由于 Euclidean 每轮运算仅需做一次内积,而 Jaccard 和 Cosine 需要做 3 次内积,因此 Jaccard 和 Cosine 的平均运行时间比 Euclidean 多了近 3 倍,SMTCFSD 算法的运行时间比 Euclidean 多了约 2 倍。

结束语　特征词在一定程度上包含了所属文档的特点,同类文档特征词集相似度较高,异类文档之间特征词集差异显著。为了凸显不同类别间特征的独特性,提出基于特征隶属度的文本分类相似性度量算法。基于特征与文档的隶属关系,将特征词集划分为全隶属词集、偏隶属词集和无隶属词集。相比于隶属于两篇文档的全隶属词集,仅隶属于某一篇文档的偏隶属词集在辨别异类文档时更有区分度。同类文档间全隶属词语较多,偏隶属词语较少,且相似度越高的文档间全隶属词语权重越相近;异类文档间偏隶属词语更多,全隶属词语较少,且所包含的全隶属词语权重间差异较为明显。因此,文档相似度会随着偏隶属词语数目的增多而降低,也会随全隶属词集权差的增大而降低。因此,SMTCFSD 模型在度量文档相似度时特征划分得更加合理,能够明显提升分类性能。

今后的工作将致力于探究一种适合该相似性度量模型的特征选择方法;基于特征与文档间的关系对特征词进行区别和选择,使之适用于 SMTCFSD 模型,从而更有效地区分文档类别,进一步提升分类精度。

参 考 文 献

[1] SEBASTIANI F. Machine learning in automated text categorization[J]. ACM CSUR,2002,34(1):1-47.

[2] LIU L,PENG T,ZUO W L,et al. Clustering-Based PU Active Text Clasfication Method [J]. Journal of Software,2013,24(11):2571-2583. (in Chinese)
流露,彭涛,左万利,等. 一种基于聚类的 PU 主动文本分类方法[J]. 软件学报,2013,24(11):2571-2583.

[3] XIA R,XU F,ZONG C Q,et al. Dual Sentiment Analysis:Considering Two Sides of One Review [J]. IEEE Transactions on Knowledge and Data Engineering,2015,27(8):2120-2133.

[4] YU Z,WANG H X,LIN X M,et al. Understanding Short Texts through Semantic Enrichment and Hashing [J]. IEEE Transactions on Knowledge and Data Engineering,2016,28(2):566-579.

[5] AZAM N,YAO J. Comparison of term frequency and document frequency based feature selection metrics in text categorization [J]. Expert Syst. Appl.,2012,39(5):4760-4768.

[6] DUAN J,HU Q H,ZHANG L J,et al. Feature Selection for Multi-Label Classification Based on Neighborhood Rough Set [J]. Journal of Computer Research and Development,2015,52(1):56-65. (in Chinese)
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 [J]. 计算机研究与发展,2015,52(1):56-65.

[7] 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.

[8] TANG B,HE H B,BAGGENSTOSS P M,et al. A Bayesian Classification Approach Using Class-Specific Features for Text Categorization[J]. IEEE Transactions Knowledge and Data Engineering,2016,28(6):1602-1606.

[9] CHENG V C,LEUNG C H C,LIU J M,et al. Probabilistic Aspect Mining Model for Drug Reviews [J]. IEEE Transactions on Knowledge and Data Engineering,2014,26(8):2002-2013.

[10] TANG B,HE H ,DING D,et al. A parametric classification rule based on the exponentially embedded family [J]. IEEE Transactions on Neural Networks and Learning Systems,2015,26(2):367-377.

[11] LIU S H,CHENG X Q,LI F X,et al. TASC: Topic-Adaptive Sentiment Classification on Dynamic Tweets [J]. IEEE Transactions on Knowledge and Data Engineering,2015,27(6),1696-1709.

[12] SUN K W,LEE C H,WANG J. Multilabel Classification via Co-Evolutionary Multilabel Hypernetwork [J]. IEEE Transactions on Knowledge and Data Engineering,2016,28(9):2438-2451.

[13] ZHANG T,TANG Y Y,FANG B,et al. Document clustering in correlation similarity measure space[J]. IEEE Transactions on Knowledge and Data Engineering,2012,24(6):1002-1013.

[14] MORI U,MENDIBURI A,LOZANO J A. Similarity Measure Selection for Clustering Time Series Databases[J]. IEEE Transactions on Knowledge and Data Engineering,2016,28(1):181-195.

[15] KANG Y B,HAGHIGH P D,BURSTEIN F. TaxoFinder:A Graph-Based Approach for Taxonomy Learning [J]. IEEE Transactions on Knowledge and Data Engineering,2016,28(2):524-536.

[16] WANG Q,CUI M Y,LIANG H Z. Semantic-Aware Bloking for Entiy Resolution [J]. IEEE Transactions on Knowledge and Data Engineering,2016,28(1):166-180.

[17] MCCALLUM A,NIGAM K,et al. A comparison of event models for naive bayes text classification[C]// AAAI-98 Workshop on Learning for Text Categorization. 1998:41-48.

[18] SCHOENHARL T W,MADEY G. Evaluation of measurement techniques for the validation of agent-based simulations against streaming data[C]// Proc. ICCS. Kraków,Poland,2008:6-15.

[19] STREHL A,GHOSH J. Value-based customer grouping from large retail data-sets[C]// Proc. SPIE. Orlando,FL,USA,2000:33-42.

[20] BISWAS S K,MILANFAR P. One Shot Detection with Laplacian Object and Fast Matrix Cosine Similarity [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2016,38(3):546-562.

(上接第 288 页)

[10] XIA Y M,CHENG B,CHEN J L,et al. Optimizing Services Composition Based on Improved Ant Colony Algorithm[J]. Chinese Journal of Computers,2012,35(2):270-281. (in Chinese)
夏亚梅,程渤,陈俊亮,等. 基于改进蚁群算法的服务组合优化 [J]. 计算机学报,2012,35(2):270-281.

[11] RAZAVI S,TOLSON B A. A New Formulation for Feedforward Neural Networks[J]. IEEE Transactions on Neural Networks,2011,22(10):1588-1598.

[12] WU X,WILAMOWSKI B M. Advantage Analysis of Sigmoid Based RBF Networks[C]// IEEE 17th International Conference on Intelligent Engineering Systems. Costa Rica,2013.

[13] MULLER B. A note on the generation of random normal deviates[J]. Annals of Mathematical Statistics,1958,29:610-611.

[14] SAMPAIO L H D,ADANIYA M H A C,DE PAULA MARQUES M,et al. Ant colony optimization for resource allocation and anomaly detection in communication networks[M]. INTECH Open Access Publisher,2013.

[15] KARIMI A,NOBAHARI H,SIARRY P. Continuous ant colony system and tabu search algorithms hybridized for global minimization of continuous multi-minima functions[J]. Computational Optimization and Applications,2010,45(3):639-661.

[16] HUANG C L,HUANG W C,CHANG H Y,et al. Hybridization strategies for continuous ant colony optimization and particle swarm optimization applied to data clustering[J]. Applied Soft Computing,2013,13(9):3864-3872.

[17] BACHE K,LICHMAN M. UCI machine learning repository [OL]. URL:http://archive.ics.uci.edu/ml,2013.

[18] LIAO T,DORIGO M. Ant colony optimization for mixed-variable optimization problems[J]. IEEE Transactions on Evolutionary Computation,2014,18(4):503-518.