一种结合上下文语义的短文本聚类算法

张 群 王红军 王伦文
(电子工程学院 合肥 230037)

摘 要 短文本因具有特征信息不足且高维稀疏等特点,使得传统文本聚类算法应用于短文本聚类任务时性能有限。针对上述情况,提出一种结合上下文语义的短文本聚类算法。首先借鉴社会网络分析领域的中心性和权威性思想设计了一种结合上下文语义的特征词权重计算方法,在此基础上构建词条-文本矩阵;然后对该矩阵进行奇异值分解,进一步将原始特征词空间映射到低维的潜在语义空间;最后通过改进的 K-means 聚类算法在低维潜在语义空间完成短文本聚类。实验结果表明,与传统的基于词频及逆向文档频权重的文本聚类算法相比,该算法能有效改善短文本特征不足及高维稀疏性,提高了短文的本聚类效果。

关键词 短文本聚类,上下文语义,奇异值分解,K 均值算法
中图法分类号 TP181 **文献标识码** A

Short Text Clustering Algorithm Combined with Context Semantic Information

ZHANG Qun WANG Hong-jun WANG Lun-wen
(Electronic Engineering Institute, Hefei 230037, China)

Abstract Because short text faces the challenges of information insufficiency, high dimensions and feature sparsity, conventional text clustering method has limited effect when applied to short text. In view of above, this paper proposed a novel short text clustering algorithm combined with the context semantic information. Firstly, drawing lessons from the idea of centrality and prestige in the field of social network analysis, the algorithm improved conventional feature weight calculation by considering the semantic information in the context. And on this basis, it constructs the term-document matrix and then carried out the singular value decomposition on the matrix to map the original high dimensional term vector space to the lower dimensional latent semantic space. Finally it clusters the short text on the lower dimensional latent semantic space by the improved K-means clustering algorithm. Experimental results show that using our scheme can effectively improve the characteristics of information insufficiency, high dimensions and feature sparsity of short text compared to the traditional text clustering method, and greatly improve the evaluation indicators of short text clustering.

Keywords Short text clustering, Context semantic information, Singular value decomposition, K-means clustering algorithm

1 引言

智能移动终端的普及使得移动互联网成为内容发布与共享的主要平台。由于移动终端屏幕相对较小,移动互联网中的内容更多以短文本形式呈现。如何从海量短文本数据中自动分析提取有价值的信息成为亟待解决的问题。文本聚类技术通过将相似的文本内容聚为一类来发现文本内容之间的内在联系,从而挖掘其中潜在有用的信息[1]。相比长文本,短文本具有特征信息不足、特征高维且更加稀疏等特点[2-4],这使得传统的文本聚类算法应用于短文本时面临严峻的挑战,因此有必要研究改进传统文本聚类算法使之更适用于短文本聚类任务。

通过有效的特征提取方法构建合理文本表示模型是提升文本聚类效果的关键。传统的文本聚类算法通常采用向量空间模型[5,6],通过特征词及其权重构成的向量表示文本。权重计算方面最常用的方法是基于词频及逆向文档频(Term Frequency-Inverse Document Frequency, TFIDF)[7,8]。向量空间模型的缺点在于其独立性假设,即每一个词独立表示一个语义概念。这一方面忽略了词与词之间丰富的语义关系,使模型不能深层次地体现文本的主题信息;另一方面导致了文本特征向量的高维稀疏性。尤其是移动互联网短文本具有长度短小、信息量不足且数量巨大的特点,传统的向量空间模型在表示短文本时体现出更加严重的特征不足且高维稀疏问题,使得模型表征短文本主题的能力变差,加大了后续短文本聚类的难度。为了弥补短文本特征不足及稀疏性问题,一些学者利用外部知识库(如知网、维基百科等语义词典)引入词语语义关系对短文本进行特征扩展[9-12],但这种方法严重依赖于知识库的质量,计算量大、计算复杂度高,同时不可避免

本文受国家自然科学基金(61273302)资助。
张 群(1992-),男,硕士生,主要研究方向为智能信息处理、Web 数据挖掘,E-mail:zhangqunbit@163.com;王红军(1968-),男,博士,副教授,主要研究方向为无线通信网、认知电子战;王伦文(1966-),男,博士,副教授,主要研究方向为数据挖掘、智能信息处理。

会引入噪声，对解决“维数灾难”问题没有帮助。潜在语义分析(Latent Semantic Analysis，LSA)[13-15]通过奇异值分解(Singular Value Decomposition，SVD)将原始特征词空间映射到潜在语义空间，在扩展语义信息的同时达到降维去噪的目的，从而提高文本聚类效果。但传统的潜在语义分析方法通常基于 TFIDF 权重，而且潜在语义空间不能包含所有的语义，当应用于特征严重稀疏的短文本聚类任务时效果并不理想。

在研究传统算法的基础上，本文提出一种结合上下文语义的短文本聚类算法。该算法通过发现词语间共现关系引入上下文语义信息，改进了传统 TFIDF 权重计算方法，然后进一步结合潜在语义分析方法降维到潜在语义空间完成短文本聚类，有效拓展了短文本特征信息，强化了潜在语义空间的语义含义，改善了短文本特征不足及特征高维稀疏问题。该算法流程如图 1 所示。

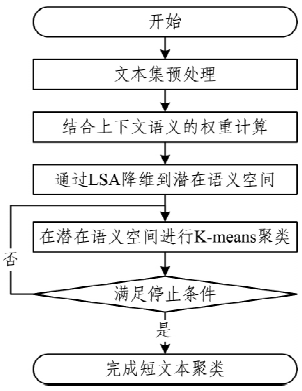

图 1　结合上下文语义的短文本聚类算法流程图

2　结合上下文语义的权重计算

2.1　算法框架描述

向量空间模型(Vector Space Model，VSM)是最常用的文本表示模型，其用文本分词后的词条作为文本特征，并赋予特征词一定的权重系数以反映特征词对于文本的重要程度。定义文本集词汇表 $V=\{t_i \mid i=1,2,\cdots,M\}$，$M$ 表示词汇表大小；定义文本集 $D=\{d_j \mid j=1,2,\cdots,N\}$，$N$ 表示文本集中文本总数，则文本集 D 中第 j 个文本的向量表示如式(1)所示。

$$d_j=\{(t_{1j},w_{1j}),(t_{2j},w_{2j}),\cdots,(t_{Mj},w_{Mj})\} \tag{1}$$

其中，每一个特征词 t_i 对应词汇表 V 构成的词空间的一个维度，$w=(w_1,w_2,\cdots,w_M)$ 表示每个特征词对应的权重，即词空间中每个维度的坐标。

针对特征词的权重分配，常用的方法是基于词频及逆向文档频(TFIDF)，TFIDF 权重的计算公式如下。

令 f_{ij} 为特征词 t_i 出现在文本 d_j 中的次数，df_i 为文本集 D 中出现特征词 t_i 的文本数，则文本 d_j 中特征词 t_i 的标准化词频(记为 itf_{ij})如式(2)所示。

$$tf_{ij}=\frac{f_{ij}}{\max\{f_{1j},f_{2j},\cdots,f_{Mj}\}} \tag{2}$$

特征词 t_i 的逆向文档频(记为 idf_i)如式(3)所示。

$$idf_i=\log\frac{N}{df_i} \tag{3}$$

最终文本 d_j 中特征词 t_i 的 TFIDF 权重等于其词频因子与逆向文档频因子的乘积，如式(4)所示。

$$w_{ij}=tf_{ij}\cdot idf_i \tag{4}$$

式(4)所示的 TFIDF 权重表达式的基本思想是认为特征词表征文本主题的能力由两方面来决定：一方面是特征词在文本中出现的次数，即词频因子；另一方面是出现在文本集里许多文本中的特征词对区分各文本没有判别力，即逆向文档频因子表达的含义。

VSM 模型的独立性假设实际很难满足，特征词之间并非相互独立，而是有很强的语义相关性，而 TFIDF 权重没有考虑到特征词的语义信息对其文本主题表征能力的加成；另一方面 TFIDF 权重应用于短文本时表现出严重的稀疏性。因此对 TFIDF 权重进行改进，提出结合上下文语义的权重计算方法，一方面提升模型表征文本主题的能力，另一方面缓解因短文本特征不足带来的严重稀疏性问题。

该算法的基本思想是通过特征词在文本上下文的共现关系引入语义信息，即认为一个特征词的出现总是伴随着上下文中其他特征的同时出现，而这种共现关系体现了特征词之间的语义相关性。如在一篇有关“计算机”主题的文本中，“计算机”这个词通常会伴随上下文中“软件”、“硬件”、“网络”等词的同时出现，而如果“计算机”这个词仅仅是单独出现在一篇文本中，该文本通常并非“计算机”主题，因此“计算机”在该文本中的权重也应相应减弱。

结合上下文语义的权重计算方法描述如算法 1 所示。

算法 1　结合上下文语义的权重计算方法。

输入：短文本集

输出：结合上下文语义的文本向量

步骤 1　计算 TFIDF 权重，构建文本向量空间模型。

步骤 2　找出文本集中所有上下文共现词对。

步骤 3　根据上下文共现词对计算特征词的语义权重(Term Semantics Weight，TSW)。

步骤 4　综合 TFIDF 权重与语义权重，得到最终特征词权重表达式。

步骤 2 中共现词对的发现方法以及步骤 3 中语义权重的计算方法在 2.2 节详细描述。

2.2　基于上下文共现关系的语义权重计算

若两个特征词同时出现在文本集中的次数大于或等于 2，则称这两个特征词为一组共现词对，将共现词对在文本集中出现的次数定义为该共现词对的共现度。定义详细描述如下。

将文本集中每个文本看作由其所包含词条构成的集合，忽略词条出现次数及顺序，则每个文本 $d_j\in D$ 是文本集词汇表 V 的一个子集，即 $d_j\subseteq V$。定义一个包含两个词条的集合 $T_{ij}=\{t_i,t_j \mid t_i,t_j\in V$ 且 $i\neq j\}$，则 $T_{ij}\subseteq V$。若词条集 T_{ij} 是文本 $d_j\in D$ 的一个子集，即 $T_{ij}\subseteq d_j$，则称文本 d_j 包含词条集 T_{ij}。

定义 1　若文本集 D 中包含词条集 T_{ij} 的文本数大于或等于 2，则称词条集 T_{ij} 中的两个词条 t_i 与 t_j 为一组共现词对，同时将包含 T_{ij} 的文本数称为该共现词对的共现度，记为 $\sup(T_{ij})$。

共现词对的发现方法如算法 2 所示。

算法 2　共现词对的发现方法

输入：短文本集 $D=\{d_j \mid i=1,2,\cdots,N\}$

输出：共现词对集 $T=\{T_{ij}\}$ 及每个共现词对的共现度 $\sup(T_{ij})$

1. 构建短文本集词汇表 $V=\{t_i \mid i=1,2,\cdots,M\}$，并将每一篇短文本 $d_j\in D$ 集合化
2. for each t_i from V：
3.　计算词条的 t_i 文档频 df_i
4.　if $df_i<2$：
5.　　delete t_i from V

6. end if
7. end for
8. 将 V 中词条两两合并，产生共现词对集 T={T_{ij}}
9. for each T_{ij} from T:
10. for each d_j in D:
11. if $T_{ij}\subseteq d_j$:
12. sup(T_{ij})++
13. end if
14. end for
15. if sup(T_{ij})<2:
16. sup(T_{ij})=0
17. end if
18. end for
19. return T={T_{ij}},sup(T_{ij})

算法 2 中第 15，16 行，若 T_{ij} 出现次数少于 2，为方便后面计算，不将其从共现词对集删除，而是将其共现度 $\sup(T_{ij})$ 置为 0，以表示 T_{ij} 中的两个词没有共现关系。

在算法 2 中发现的共现词对的基础上，借鉴社会网络分析领域的中心性和权威性思想，计算特征词的语义权重。社会网络分析中通过构造网络图研究网络中参与者的属性以及参与者之间的关系，认为一个参与者在网络中的重要性不仅与其本身属性有关，还与网络中其他与其有关系的参与者有关，即与其有关系的参与者在网络中的重要程度也体现了其本身的重要程度。因此，将文本集看作一个网络，表示为无向图模型 $G=(V,T)$，V 表示网络中所有参与者集合，即文本集词汇表，T 表示网络中所有参与者关系集合，这里仅考虑特征词的上下文共现关系，即 $T=\{T_{ij}\}$。则特征词 t_i 的语义权重为：

$$TSW_i=\sum_{j=1}^{M}TFIDF_j\cdot\frac{sup(T_{ij})}{N} \tag{5}$$

其中，TSW_i 为特征词 t_i 的语义权重，$TFIDF_j$ 为特征词 t_j 的 TFIDF 权重，$sup(T_{ij})$ 为词对 t_i 与 t_j 的共现度，N 为文本集中文本数。式(5)表明：一方面，与 t_i 有上下文共现关系的词越多并且之间共现度越高，则 t_i 语义权重越大；另一方面用 TFIDF 权重衡量与 t_i 有共现关系的词的权威度，权威度越大，则 t_i 的语义权重越大。

综合 TFIDF 权重与语义权重，得特征词最终结合上下文语义的权重表达式为：

$$WEIGHT_i=\alpha\cdot TFIDF_i+\beta\cdot TSW_i \tag{6}$$

其中，$\alpha+\beta=1$，本文取 $\alpha=\beta=0.5$。

由式(6)可以看出，即使一个特征词在一篇文本中没有出现，即其 TFIDF 权重为 0，但上下文中出现的与其语义相关的其他特征词会为之贡献一定的语义权重，从而使该特征项总权重不为 0 值，改善短文本向量的稀疏性。

3 基于潜在语义分析的 K-means 聚类算法

在上文提出的结合上下文语义的权重计算的基础上，通过潜在语义分析方法进一步将原始词空间映射到潜在语义空间，利用 K-means 聚类算法完成短文本聚类。

潜在语义分析(Latent Semantic Analysis，LSA)也称为潜在语义索引(Latent Semantic Index，LSI)。LSA 模型首先构造一个 $M\times N$ 的特征词-文本矩阵 A，A 的行向量为特征词向量，列向量为文本向量，A 的各个项值 A_{ij} 表示文本 d_j 中特征词 t_i 的权重。一般情况下采用 TF 或者 TFIDF 权重，但由于短文本的特点使得用 TF 或者 TFIDF 权重构建的矩阵 A 表现出严重的高维稀疏性。因此采用第 1 节中提出的结合上下文语义的权重构建矩阵 A，以缓解矩阵 A 的稀疏性。

特征词-文本矩阵 A 构造后，通过奇异值分解(Singular Value Decomposition，SVD)方法将原始矩阵 A 分解为 3 个矩阵，如式(7)所示。

$$A=U_{M\times r}\sum_{r\times r}V^{T}_{r\times N} \tag{7}$$

其中，U 矩阵为 $M\times r$ 的左奇异向量矩阵；V 矩阵为 $N\times r$ 的右奇异矩阵；Σ 是一个 $r\times r$ 的对角阵，对角向量为 $\Sigma=\mathrm{diag}(\sigma_1,\sigma_2,\cdots,\sigma_r)$，$\sigma_1,\sigma_2,\cdots,\sigma_r$ 称为奇异值，按降序排列，即 $\sigma_1\geqslant\sigma_2\geqslant\cdots\geqslant\sigma_r\geqslant0$。$M$ 为矩阵 A 的行数，即文本集特征词数目；N 为矩阵 A 的列数，即文本数；r 为矩阵 A 的秩，即转换后的潜在语义空间的维数，$r\leqslant\min(M,N)$。

从式(7)可以看出，通过 SVD，LSA 方法将原始高维空间的词向量和文本向量映射到了低维的语义空间。在低维语义空间中，矩阵 U 中存储了语义相关的词向量，矩阵 V 中存储了主题相关的文本向量，对角矩阵 Σ 表示词语义和文本主题之间的相关性。

另外，对角矩阵 Σ 中的奇异值 $\sigma_1,\sigma_2,\cdots,\sigma_r$ 反映了低维潜在语义空间的语义特征的重要性，并且奇异值降速非常快，通常前 10% 甚至 1% 的奇异值就包含了全部奇异值信息的 99% 以上。因此可以仅保留前 k 个大奇异值所对应的重要语义特征维，删除剩下 $r-k$ 个小奇异值对应的无关紧要的噪声维，即删除对角阵 Σ 中最后的 $r-k$ 行和列，以及矩阵 U 和 V 中最后的 $r-k$ 列，如式(8)所示。

$$A_k=U_{M\times k}\sum_{k\times k}V^{T}_{k\times N} \tag{8}$$

其中，矩阵 A_k 可近似原始矩阵 A。针对 k 的取值，本文做法是保留所有奇异值 90% 的信息，即将奇异值的平方和累加到总值的 90% 为止，如式(9)所示。

$$\frac{\sum_{i=1}^{k}\sigma_i}{\sum_{i=1}^{r}\sigma_i}=90\% \tag{9}$$

通过 SVD 分解以及取前 k 个奇异值，LSA 将原始高维的特征词空间映射到低维的潜在语义空间，缓解了高维稀疏性，同时突出强化了词与词、文本与文本、词与文本间的语义关系，并消除了噪声因素。

$\sum_{k\times k}V^{T}_{k\times N}$ 表示了转变后的文本语义空间，接下来在该文本语义空间中采用改进的 K-means 聚类算法对文本向量进行聚类。

K-means 聚类算法描述如算法 3 所示。

算法 3 K-means 聚类算法

输入：语义空间中的文本向量 d_j'，j=1，2，…，N
输出：文本集合
步骤 1 指定聚类数目 K，以及 K 个初始聚类中心。
步骤 2 指定一个距离函数。
步骤 3 计算每个文本向量 d_j' 与 K 个初始聚类中心的距离，将每一个文本向量 d_j' 分配给距离最小的聚类中心。
步骤 4 重新计算新的 K 个聚类中心。
步骤 5 重复步骤 3 及步骤 4，直到满足某个终止条件。

针对算法 3 的说明及改进如下：步骤 1 中通过人工指定并反复尝试确定较为合理的聚类数目 K；认为语义空间中语义特征维数多的文本向量更具有主题代表性，因此选取语义特征维数多的文本向量作为初始聚类中心；步骤 2 中不采用

距离函数而是采用余弦相似度，这是因为在文本聚类中余弦相似度相比距离函数能有效地反映特征维度之间的差异而规避特征数值上的差异；相应地，步骤 5 中终止条件将“使距离误差平方之和最小”改为“使余弦相似度之和局部最大”，余弦相似度之和(Cosine Similarity Sum，CSS)的计算公式为：

$$CSS=\sum_{k=1}^{K}\sum_{d_j'\in C_k}\text{cosine}(d_j',c_k) \tag{10}$$

其中，K 为聚类数，c_k 为第 k 个聚类中心，C_k 为以 c_k 为聚类中心的类集合。

4 实验结果及分析

4.1 数据集

采用包含人工标注的中文微博数据集作为实验数据集(http://www.datatang.com/data/45871)。数据集共包含 10000 篇共 10 个大类的新浪微博短文本。选取其中 5 个大类共 2000 篇短文本进行实验，分别为社会类 527 篇、娱乐类 379 篇、科技类 335 篇、经济类 394 篇、体育类 365 篇。

4.2 评价指标

从算法收敛速度与聚类结果两个方面对算法进行评价。

聚类算法的收敛速度用达到最大余弦相似度之和时的迭代次数进行衡量，余弦相似度之和(CSS)计算方法如式(10)所示。

聚类结果用准确率(Precision，Pr)、召回率(recall，Re)以及 F_1 值 3 个指标来衡量，如式(11)－式(13)所示。

$$Pr=\frac{TP}{TP+FP} \tag{11}$$

$$Re=\frac{TP}{TP+FN} \tag{12}$$

$$F_1=\frac{2\cdot Pr\cdot Re}{Pr+Re} \tag{13}$$

式中各参数含义如表 1 所列。

表 1 参数含义表

	分类为 c 类	分类非 c 类
实际为 c 类	TP	FN
实际非 c 类	FP	TN

其中，准确率考察的是聚类结果的正确性，召回率考察的是聚类结果的完备性，F_1 值是二者的调和平均值。

4.3 结果及分析

实验中，分别在数据集上测试本文提出的结合上下文语义的短文本聚类算法(TSW-LSA)以及传统基于 TFIDF 权重的 LSA 聚类算法(TFIDF-LSA)。针对本文提出的算法测试过程如下：首先对短文本数据集进行预处理，包括中文分词、停用词过滤、构造文本集词汇表等操作，词汇表中词条个数为 3676 个；然后利用第 2 节提出的结合语义的权重计算方法计算特征词的权重，在此基础上构建词条-文本矩阵，该矩阵是一个 3676×2000 的矩阵；接下来通过第 3 节提到的 LSA 方法对该矩阵进行 SVD 分解，得到该矩阵的 1650 个奇异值，根据式(9)，选取前 350 个大的奇异值；最后通过改进的 K-means 聚类算法在 350 维的低维潜在语义空间完成短文本聚类。

首先比较了结合上下文语义的聚类算法与传统基于 TFIDF 权重的 LSA 聚类算法在收敛速度方面的优劣，结果如图 2 所示。

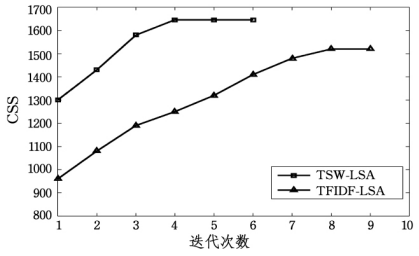

图 2 两种聚类算法余弦相似度和(CSS)随迭代次数变化曲线

从图 2 可以看出本文算法仅通过 4 次迭代就使得余弦相似度之和达到局部最大值，相比之下传统基于 TFIDF 权重的 LSA 聚类算法至少需要 9 次迭代，表明了本文算法在收敛速度方面明显优于传统算法。

聚类结果方面的比较结果如表 2 所列。

表 2 两种聚类算法聚类结果(%)

		社会类	经济类	科技类	娱乐类	体育类	平均值
TFIDF-LSA	Pr	53.2	56.4	53.8	56.2	59.4	55.8
	Re	64.7	68.7	61.0	59.3	56.2	62.0
	F_1	58.4	61.9	57.2	57.7	57.8	58.7
TSW-LSA	Pr	71.6	84.2	88.7	75.6	86.0	81.2
	Re	78.5	79.7	91.0	80.3	88.3	83.6
	F_1	74.9	81.9	89.9	77.9	87.1	82.4

表 2 显示在聚类结果方面，本文算法在聚类准确率、召回率及 F_1 值 3 个指标上均明显优于基于 TFIDF 权重的 LSA 聚类算法，表明本文算法通过引入上下文语义信息并降维到潜在语义空间进行短文本聚类，能有效改善短文本特征不足及高维稀疏性，是一种有效的短文本聚类方法。另外从表 2 可以看出本文算法在科技类和体育类上的表现明显优于社会类及娱乐类，分析原因是由于各个类别中的文本上下文语义相关性有差别，如科技类和体育类属于小范畴类别，其中文本语义相关性较强。本文算法由于引入了基于共现关系的上下文语义信息，因此受到类别中文本语义相关性的影响较大，相比之下传统算法无基于共现关系的语义信息引入，因此在各个类别上的性能表现无明显差别。

结束语 本文核心内容在于借鉴了社会网络分析领域中心性与权威性的思想，基于文本上下文词语共现关系引入语义信息，改进了传统的 TFIDF 权重计算，并在此基础上改进了潜在语义分析方法与 K-means 聚类算法，使之适用于短文本聚类任务。实验结果表明，通过引入语义权重并映射到低维潜在语义空间进行短文本聚类，能有效缓解短文本的特征不足及高维稀疏性，提高短文本聚类效果。接下来将重点研究传统聚类算法的改进以进一步提升短文本聚类效果，如 K-means 聚类算法聚类数 K 值的自适应选取、短文本主题层次聚类等问题。

参 考 文 献

[1] 孟宪军. 互联网文本聚类与检索技术研究[D]. 哈尔滨：哈尔滨工业大学，2009

[2] 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

[3] 彭泽映，俞晓明，许洪波，等. 大规模短文本的不完全聚类[J]. 中文信息学报，2011，25(1)：54-59

[4] 程传鹏，苏安婕. 一种短文本特征词提取的方法[J]. 计算机应用与软件，2014，31(6)：162-164

(下转第 450 页)

$$RI=\frac{a+d}{n\cdot(n-1)} \tag{4}$$

然后用 Iris，Wine 和 Sonar 数据集进行 W-PAM 限制聚类实验，并与 CKM，CCL 算法进行对比，如图 1－图 3 所示。

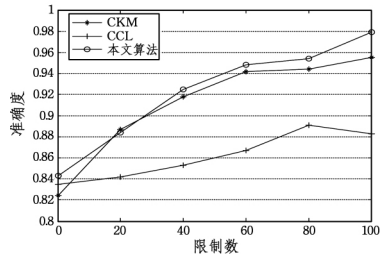

图 1　Iris 数据集

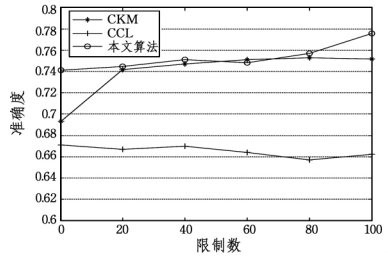

图 2　Wine 数据集

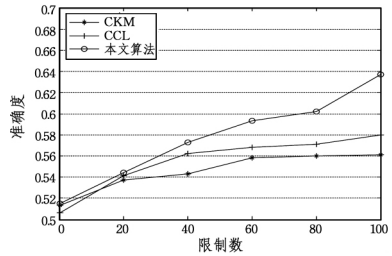

图 3　Sonar 数据集

从图 1－图 3 可以看出，W-PAM 限制聚类算法在聚类质量上有一定的改进，从总体上看，在限制数较小时聚类性能比较接近，随着限制数的增大本文算法更有优势。从 Iirs 数据集和 Sonar 数据集上看比较明显，Wine 数据集上变化不大但也有所提高。

结束语　本文通过结合传统的 PAM 聚类算法和经典的限制聚类算法，提出了一种 W-PAM 算法与关联限制相结合的限制聚类算法，实验结果表明：该算法可以初步处理一定规模的数据，在降低了 PAM 算法迭代次数的同时提高了算法的准确率。由于算法在寻找有效分隔时还存在不稳定性，在今后的学习中还需深入研究。

参 考 文 献

[1] 孙富贵，刘杰，赵连宁. 聚类算法研究[J]. 软件学报，2008，19(1)：48-61

[2] Vapnik V. Statistical Learning Theory [M]. New York：John Wiley，1998

[3] MacQueen J. Some methods for classification and analysis of multivariate observations [C]// Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley：University of California Press，1967：281-297

[4] Park H S，Jun C H. A simple and fast algorithm for K-medoids clustering [J]. Expert Systems with Applications，2009，36(2)：3336-3341

[5] 何萍，徐晓华，陆林，等. 双层随机游走半监督聚类[J]. 软件学报，2014，25(5)：997-1013

[6] 马儒宁，王秀丽，丁军娣. 多层核心集凝聚算法[J]. 软件学报，2013，24(3)：490-506

[7] Zhou Y，Wang X，Wang T，et al. Fault-tolerant multi-path routing protocol for WSN based on HEED[J]. International Journal of Sensor Networks，2016，20(1)：37

[8] 陈克寒，韩盼盼，吴健. 基于用户聚类的异构社交网络推荐算法[J]. 计算机学报，2013，36(2)：349-359

[9] 刘卓，杨悦，张健沛，等. 不确定度模型下数据流自适应网格密度聚类算法[J]. 计算机研究与发展，2014，51(11)：2518-2527

[10] Wagstaff K，Cardie C. Clustering with instance-level constraints [C]// Proc of the 17th International Conference on Machine Learning (ICML-2000). 2000：1103-1110

[11] Sun Jun，Zhao Wen-bo. Xue Jiangwei et al. Clustering with feature order preferences [J]. Intelligent Data Analysis，2010，14(4)：479-495

[12] M Law，A Topchy，A Jain. Clustering with Soft and Group Constraints [C]// Proc of Joint IAPR Int'l Workshop on Structural Syntactic and Statistical Pattern Recognition. 2004：662-670

[13] Wagstaff K，Cardie C，Rogers S，et al. Constrained K-means Clustering with Background Knowledge[C]// Proc of the 18th International Conference on Machine Learning (ICML-2001). 2001：577-584

[14] Klein D，Kamvar S，Manning C. Form Instance-level Constraints to Space-Level Constraints：Making the Most of Prior Knowledge in Data Clustering[C]// Proc of the 19 International Conference on Machine Learning (ICML-2002). 2002：307-314

[15] 韩家炜，堪博著. 数据挖掘：概念与技术[M]. 范明，孟小峰，译. 北京：机械工业出版社，2007：251-267

[16] 刘正，张国印，陈志远. 基于特征加权和非负矩阵分解的多视角聚类算法[J]. 电子学报，2016，44(3)：536-540

(上接第 446 页)

[5] Jing Li-ping，Ng M K，Huang J Z. Knowledge-based vector space model for text clustering [J]. Knowledge and Information Systems，2010，25(1)：35-55

[6] 刘海峰，刘守生，张学仁. 聚类模式下一种优化的 K-means 文本特征选择[J]. 计算机科学，2011，38(1)：195-197

[7] 雷军程，黄同成，柳小文. 一种基于权重的文本特征选择方法[J]. 计算机科学，2012，39(7)：250-252，275

[8] 张保富，施化吉，马素琴. 基于 TFIDF 文本特征加权方法的改进研究[J]. 计算机应用与软件，2011，28(2)：17-20

[9] 朱征宇，孙俊华. 改进的基于知网的词汇语义相似度计算[J]. 计算机应用，2013，33(8)：2276-2279，2288

[10] 王荣波，谌志群，周建政，等. 基于 Wikipedia 的短文本语义相关度计算方法[J]. 计算机应用与软件，2015，32(1)：82-85，92

[11] 宁亚辉，樊兴华，吴渝. 基于领域词语本体的短文本分类[J]. 计算机科学，2009，36(3)：142-145

[12] Batet M. Ontology-based semantic clustering[J]. Ai Communications，2011，24(3)：291-292

[13] 强保华，李巍，邹显春，等. 基于潜在语义分析的 Deep Web 查询接口聚类研究[J]. 计算机科学，2013，40(11)：228-230，247

[14] Xia Yan，Hua Zhao. Chinese Microblog Topic Detection Based on the Latent Semantic Analysis and Structural Property [J]. Journal of Networks，2013，8(4)：917-923

[15] Dumais S T. Latent semantic analysis [J]. Annual Review of Information Science & Technology，2008，3(11)：188-230