ISSN 1673-9418 CODEN JKYTA8
Journal of Computer Science and Frontiers
1673-9418/2007/01(03)-0331-09

E-mail:fcst@public2.bta.net.cn
http://www.ceaj.org
Tel:+86-10-51616056

极大熵球面K均值文本聚类分析*

修 宇[1,3],王士同[1,2+],朱 林[1],宗成庆[2]
XIU Yu[1,3], WANG Shitong[1,2+], ZHU Lin[1], ZONG Chengqing[2]

1.江南大学 信息工程学院,江苏 无锡 214036
2.中科院自动化研究所 模式识别国家重点实验室,北京 100080
3.安徽工程科技学院 计算机科学与工程系,安徽 芜湖 241000
1.School of Information Engineering, Jiangnan University, Wuxi, Jiangsu 214036, China
2.National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China
3.Department of Computer Science and Engineering, Anhui University of Technology and Science, Wuhu, Anhui 241000, China
+Corresponding author: E-mail: wxwangst@yahoo.com.cn

XIU Yu, WANG Shitong, ZHU Lin, et al. Maximum-entropy sphere K-means document clustering analysis. Journal of Computer Science and Frontiers, 2007,1(3):331-339.

Abstract: A maximum-entropy version of the spherical K-means document clustering algorithm ME-SPKM is presented based on the well-known maximum-entropy. The proposed method uses the cosine similarity which is adopted by the typical text clustering algorithm SPKmeans, then constructs a maximum-entropy-based objective function. Experimental results demonstrate that the maximum-entropy spherical K-means ME-SPKM can achieve better clustering results than traditional clustering approaches in text clustering.
Key words: maximum-entropy; document clustering; spherical K-means

摘 要: 提出了一种基于极大熵理论的球面K均值文本聚类算法ME-SPKM。该算法利用了传统文本聚类算法SPKmeans中使用的余弦相似度度量,进而引入极大熵理论构造了适合文本聚类的极大熵目标函数。对文本数据的实验证明了极大熵球面K均值文本聚类算法取得了比传统文本聚类算法更好的聚类效果。
关键词: 极大熵;文本聚类;球面K均值
文献标识码: A **中图分类号:** TP18

1 引言

万维网(World Wide Web,WWW)是一个庞大而又充满着混沌的网络。随着万维网以及各种文本资源的不断增长,人们对快速、准确而全面获取信息

* the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences(中科院自动化研究所模式识别国家重点实验室基金资助).
Received 2007-08, Accepted 2007-10.

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

的渴望日益增加，文本聚类技术作为非监督聚类技术得到了越来越多的重视与研究。目前文本聚类技术已经成为自动文本归类、自动摘录、快速信息检索以及信息过滤的关键技术[1]。一个搜索引擎会在一个搜索命令下返回成千上万的页面，以至于用户很难找到相关的有用信息，而通过聚类技术可以预先将相关信息归类，从而最大可能地提供用户有用的信息。此外大型的数据库也通过预先聚类数据使得系统可以更快地响应用户的检索。因此如何对海量的网络信息、电子数据库、新闻等数据进行有效的聚类成为近年来的研究热点。

在一定的向量空间模型下通过适当的预处理，文本可以按词项表示为高维空间的一个向量，该向量具有稀疏性和非负性[1]。近年来研究表明文本数据还具有方向性[2]。比如文本向量的度量只需要考虑向量数据的方向，数据本身无大小之分，只有方向或相对位置不同。这种特征使得聚类处理前一般将文本向量数据做归一化处理，而后再进行聚类分析，相应的相似性度量一般采用向量的内积或夹角余弦。文献[3]提出了球面聚类算法 SPKmeans。SPKmeans 算法合理地采用了余弦相似度来度量文本向量的相关性。实验表明对于超大数量文本集聚类问题 SPKmeans 算法具有较好的综合性能，但算法的聚类性能仍有待提高。文章在对球面聚类算法 SPKmeans 及极大熵聚类算法研究的基础上提出了一种熵意义下的模糊球面聚类算法，进而达到提高文本算法的聚类性能的目标。

1990 年 Rose 等人首次提出基于熵的确定性退火聚类算法[4]，受到它的启发，一种新的模糊 K 均值方法被提出，其在 K 均值算法的目标函数中引入熵的概念从而得到熵意义下的模糊聚类算法，该方法被称作是极大熵方法 MEC。Karayiannis (1994)[5]、Li 与 Mukaidono[6]分别发表了极大熵聚类算法的研究结果，文献[7]研究了极大熵聚类算法的收敛性及其聚类模型。这些研究构造新颖，为克服传统聚类算法存在的问题提供了一个新的思路。近几年将极大熵原理也被广泛应用于自然语言处理[8]及文本分类中[9]。A. Banerjee 在文献[10]中指出基于混合 vMF 密度模型的文本聚类算法 movMF 在文本聚类过程中隐变量熵的变化具有自退火的特征，Shi Zhong[11,12]近一步提出了可以通过确定性退火技术来改善 movMF 算法对文本的聚类性能，这些研究为在传统文本聚类算法中引入极大熵原理提供了依据。

作者将极大熵原理引入到球面 K 均值文本聚类算法中，提出了熵意义下的模糊球面聚类算法即极大熵球面 K 均值聚类算法 MESPKM，并通过大量实验证明了该算法聚类性能优于其他文本聚类算法。

文章结构如下：第 2 章简要介绍极大熵聚类算法和球面 K 均值聚类算法；第 3 章提出极大熵球面 K 均值聚类算法；第 4 章首先说明实验采用的基准数据集及度量算法性能标准和实验参数设置，然后给出实验结果及分析；第 5 章总结全文并给出了下一步研究的方向。

2 极大熵聚类算法 MEC 和球面 K 均值聚类算法 SPKmeans

2.1 确定性退火技术与极大熵聚类算法

利用统计物理的退化过程，Rose 博士提出了确定性退火技术[4]，它是按自然法则计算的一个重要分支。它根据退火过程，将求解优化问题的最优点转化成求解一系列随温度变化的物理系统的自由能函数极小。Karayianni[5]等人将确定性退火技术引入到聚类算法中，提出了极大熵聚类算法，其本质仍是利用确定性退火技术求得聚类的目标函数最小。极大熵聚类算法 MEC 是引入极大熵方法中的一个典型代表。在多种版本的极大熵聚类算法 MEC 中，虽然描述各不相同，但只是形式上的差别。这里仅以文献[6]中的描述对极大熵聚类算法 MEC 进行介绍。

对于数据集 $X=\{x_1,\cdots,x_N\}\subset R^d$，$V=\{v_1,v_2,\cdots,v_K\}$ 是 K 个聚类中心，$v_i\in R^d$，$2\leq K<N$，$U=\{u_{i,j}\}_{K\times N}$ 是一个隶属度矩阵，$u_{i,j}$ 为每个样本属于类中心的概率，且满足

$$u_{i,j}\in[0,1],1\leq i\leq K,1\leq j\leq n$$
$$\sum_{i=1}^{K}u_{i,j}=1 \quad (1)$$

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

极大熵模糊聚类算法 MEC 把 N 个向量 x_i (i=1,2,…,N)分成 K 个簇 G_i (i=1,2,…,K),并求得每个簇的聚类中心,以下目标函数达到最小

$$J_T(U,V)=\sum_{i=1}^{K}\sum_{j=1}^{N}u_{i,j}\|x_j-v_i\|^2+T\sum_{i=1}^{K}\sum_{j=1}^{N}u_{i,j}\ln u_{i,j} \quad (2)$$

这里 $\|x_j-v_i\|^2=(x_j-v_i)^T(x_j-v_i)$,T 是 Lagrange 乘子。上式也可表示成

$$J_T=J_c(U,V)-TH(u) \quad (3)$$

其中 $J_c(U,V)=\sum_{i=1}^{K}\sum_{j=1}^{N}u_{i,j}\|x_j-v_i\|^2$。对于大的 T,主要是试图最大化熵 H(u),系统维持在较高温度,随着 T 的降低,以熵换取失真的减小,当 T 趋于零,最小 $J_c(U,V)$ 直接获得一个非随机(硬)解。因此这里的 Lagrange 乘子 T 等价于确定性退火技术的温度系数也称为退火系数。极大熵聚类算法 MEC 的基本步骤如下:

(1)初始化:给出初始聚类中心 $V^{(0)}=\{v_1^{(0)},v_2^{(0)},\cdots,v_K^{(0)}\}$,及模糊划分矩阵 $U=\{u_{i,j}\}_{K\times N}$。l=0,l 为迭代次数,最大迭代次数 M,设定退火系数 T,最低退火系数 MinT 阈值 ε。

(2)用下列公式更新 $u_{i,j}^{l+1}$

$$u_{i,j}=\frac{\exp\left(-\frac{\|x_j-v_i\|^2}{T}\right)}{\sum_{h=1}^{K}\exp\left(-\frac{\|x_j-v_h\|^2}{T}\right)} \quad (4)$$

i=1,2,…,N,j=1,2,…,N

(3)用下列公式更新 v_i^{l+1}

$$v_i=\frac{\sum_{j=1}^{N}u_{i,j}x_j}{\sum_{j=1}^{N}u_{i,j}} \quad i=1,2,\cdots,K \quad (5)$$

如果 T 达到最小,则停止;否则如果 $\max_i\|v_i^{(l+1)}-v_i^{(l)}\|<\varepsilon$ 或者 l>M,调整退火系数 T=T-ΔT,转至(2)。

MEC 算法能避开局部极小而得到全局极小,因而得到了广泛的应用。但是 MEC 算法其中的一个缺陷是使用欧式度量,而对于高维向量化的文本数据,文本向量的方向特征远比其大小特征更为重要,故 MEC 不适合对文本数据进行聚类分析。

2.2 球面 K 均值聚类算法

对于高维向量化的文本数据,文献[13]指出使用欧式度量比使用余弦相似性度量的聚类效果差。这说明了文本向量的方向特性比向量的大小特征更重要。Dhillon I.S 在文献[3]中提出了基于余弦相似度量的球面 K 均值聚类方法,并证明了算法的有效性。文献[10]提出了基于混合方向分布 vMF 的文本聚类方法 movMF,同时指出球面 K 均值文本聚类算法是 movMF 的一种特殊情况,进而证实了文本数据的方向性和使用余弦相似度来度量文本数据的必要性。以下给出球面 K 均值聚类算法的简单描述。

对于样本集 $X=\{x_1,\cdots,x_N\}\subset R^d$ 且有 $x_i^Tx_i=1$(1≤i≤N),$V=\{v_1,v_2,\cdots,v_K\}$ 是 K 个聚类中心,且有 $v_i^Tv_i=1$(1≤i≤N),2≤K<N,$U=\{u_{i,j}\}_{K\times N}$ 是一个隶属度矩阵,则球面 K 均值文本聚类算法的目标是最大化全局代价函数

$$J=\sum_{i}^{N}\sum_{j}^{K}x_i^Tv_j \quad (6)$$

其中 $x_i^Tv_j$ 为向量 x_j 和向量 v_j 的内积,即 2 个向量的余弦相似度。球面 K 均值聚类算法的基本步骤如下:

(1)初始化:给出初始聚类中心 $V^{(0)}=\{v_1^{(0)},v_2^{(0)},\cdots,v_K^{(0)}\}$,且满足 $v_i^Tv_i=1$(1≤i≤N),l=0,l 为迭代次数,最大迭代次数 M,阈值 ε。

(2)用下列公式更新:

$$u_{i,j}^{(l+1)}=\begin{cases}1, & \text{如果 } i=\arg\max_K\{x_j^Tv_i^{(l)}\}\\ 0, & \text{否则}\end{cases}$$

(3)用下列公式更新 $v_i^{(l+1)}$

$$v_j=\frac{\sum_{i}^{K}u_{i,j}^{(l+1)}x_i}{\sqrt{\left(\sum_{i}^{K}u_{i,j}^{(l+1)}x_i\right)^T\left(\sum_{i}^{K}u_{i,j}^{(l+1)}x_i\right)}} \quad (7)$$

(4)如果 $\max_i\|v_i^{(l+1)}-v_i^{(l)}\|<\varepsilon$,或者 l>M,则停止;

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

否则 l=l+1,转至(2)。

球面 K 均值算法思想简单,实现容易,运行速度快,内存消耗小,因为采用了余弦度量,能有效地处理大文本数据集[3]。但球面 K 均值算法也有不少缺点,如采用的是硬划分,隶属度要么是 1 要么是 0,不能反映数据点与类中心的实际关系;此外对初始化敏感,极易陷入局部最小点,这些缺点都有待于进一步改进。

作者注意到,当样本向量和中心向量都为单位向量时(‖x‖=‖y‖=1),则有

$$\sqrt{\|x-y\|^2}=\sqrt{\|x\|^2+\|y\|^2-2x^{\mathrm{T}}y}=\sqrt{2(1-x^{\mathrm{T}}y)} \tag{8}$$

即

$$\mathrm{Euclidean}(x,y)=\sqrt{2}\sqrt{1-\cos(x,y)} \tag{9}$$

注意式(9)揭示了归一化后的向量使用欧式距离和余弦相似之间的联系,最大化向量的余弦聚类即最小化向量的欧式距离,这也是将球面 K 均值聚类算法视为一种特殊的 K 均值聚类算法的原因。然而,近年来研究表明二者的内在概率却有本质区别,文献[11,14]指出 K 均值聚类算法的内在概率为高斯分布,文献[10]推导出球面均值聚类算法实际上是混合方向分布模型聚类算法 movMF 的一种特殊情况,即证实了球面 K 均值文本聚类算法的内在概率分布是方向性分布而不是高斯分布。这同时也说明为什么对于高维向量化的文本数据,使用欧式度量的聚类效果远差于使用余弦相似性度量的聚类效果。

3 极大熵球面 K 均值聚类算法 ME-SPKM

基于第 2 章的讨论,下面将极大熵原理引入到球面均值聚类中,推导出适合文本聚类的极大熵球面 K 均值聚类算法 ME-SPKM。

对于样本集 $X=\{x_1,\cdots,x_N\}\subset R^d$,有 $x_i^{\mathrm{T}}x_i=1$ $(1\le i\le N)$,$V=\{v_1,v_2,\cdots,v_K\}$ 是 K 个聚类中心,且有 $v_i^{\mathrm{T}}v_i=1$ $(1\le i\le N)$,$2\le K<N$,$U=\{u_{i,j}\}_{K\times N}$ 是一个隶属度矩阵。$u_{i,j}$ 是样本 x_j 属于 K 个中心的隶属度,其取值不同于球面 K 均值采用的硬划分,而是在 0 到 1 之间的模糊划分,真实地反映了数据点与类中心点的实际关系,且满足 $\sum_i^K u_{i,j}=1$。此时全局最大化代价函数可以视为

$$J=\sum_i^K\sum_j^N u_{i,j}x_j^{\mathrm{T}}v_i \tag{10}$$

为求得式(10)的极大值,避开局部极小而得到全局极小,可以引入极大熵原理,此时可以定义最小化目标函数为

$$J_T(U,V)=-\sum_{i=1}^K\sum_{j=1}^N u_{i,j}x_j^{\mathrm{T}}v_i+\frac{1}{T}\sum_{i=1}^K\sum_{j=1}^N u_{i,j}\ln u_{i,j} \tag{11}$$

注意式(11)的形式类似于式(2),都引入了熵项。但两者意义却不同,式(2)使用的是欧式度量,而式(11)使用的是余弦相似度量。式(11)也可表示成

$$J_T=J_c(U,V)-\frac{1}{T}H(U) \tag{12}$$

其中 $J_c(U,V)=-\sum_{i=1}^K\sum_{j=1}^N u_{i,j}x_j^{\mathrm{T}}v_i$,T 是 Lagrange 乘子,可以根据需要取值,其值对最终聚类结果有一定的影响。H(U)为隶属度矩阵的熵,当 $\frac{1}{T}$ 值很大时,最小化 $J_T(U,V)$ 实际上需最大化熵 H(U)。随着 $\frac{1}{T}$ 值减小,最小化 $J_T(U,V)$ 转向最小化 $J_c(U,V)$,从而取得全局极小点。求 $J_T(U,V)$ 的最小值即求 $v_i^{\mathrm{T}}v_i=1$ $(i=1,2,\cdots,K)$,以及 $\sum_i^K u_{i,j}=1$ 条件限制下的目标函数的峰值,为此引入拉格朗日乘子 λ,并定义如下的拉格朗日目标函数 $L(U,v,\lambda,\gamma)$:

$$L(U,v,\lambda,\gamma)=J_T(U,V)+\lambda\sum_{i=1}^K(v_i^{\mathrm{T}}v_i-1)+\gamma\sum_{j=1}^N\left(\sum_{i=1}^K u_{ij}-1\right) \tag{13}$$

对 $L(U,v,\lambda,r)$ 中的每个中心向量 v_i 求偏导有

$$\frac{\partial L(U,v,\lambda,r)}{\partial v_i}=-\sum_{j=1}^N u_{i,j}x_j^{\mathrm{T}}+2\lambda v_i \tag{14}$$

令式(14)等于 0,则有

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$v_i=\frac{\sum_{j=1}^{N}u_{i,j}x_j}{2\lambda} \tag{5}$$

因 $v_i^T v_i=1$,由式(5)可以进一步推出

$$v_i=\frac{\sum_{j=1}^{N}u_{i,j}x_j}{\sqrt{\left(\sum_{j=1}^{N}u_{i,j}x_j\right)^T\left(\sum_{j=1}^{N}u_{i,j}x_j\right)}} \tag{6}$$

对 $L(u,v,\lambda,r)$ 中的每个 $u_{i,j}$ 求偏导有

$$\frac{\partial L(u,v,\lambda,r)}{\partial u_{i,j}}=-x_j^T v_i+\frac{1}{T}(\ln u_{i,j}+1)+r \tag{7}$$

令式(7)等于 0,则有

$$\ln u_{i,j}=Tx_j^T v_i-(Tr+1) \tag{8}$$

因 $\sum_{i}^{K}u_{i,j}=1$,由式(8)可以进一步推出

$$u_{i,j}=\frac{e^{Tx_j^T v_i}}{\sum_{i=1}^{K}e^{Tx_j^T v_i}} \tag{9}$$

这里称式(4)、(7)循环迭代求解目标函数(9)最小的过程为极大熵球面 K 均值聚类算法(ME-SPKM),注意到此时 Lagrange 乘子 T 相当于倒置的退火系数,即当 T 值较小时系统维持在较高温度,T 增加的过程即系统退火的过程,通过一系列随温度 T 变化从而求得目标函数的极小点。

下面给出 ME-SPKM 算法的完整描述。

Algorithm:ME-SPKM(Maximum-Entropy Sphere k-means)

步骤 1 初始化,固定 K(1≤K<N),给出初始聚类中心 $V(0)=\{v_1^{(0)},v_2^{(0)},\cdots,v_K^{(0)}\}$,模糊划分矩阵 $U=\{u_{i,j}\}_{K\times N}$,l=0,l 为迭代次数,最大迭代次数 M,设定退火系数 T,最大退火系数 MaxT,阈值 ε,迭代次数 r=0;

步骤 2 按公式(9)求出更新 $u_{i,j}^{(l+1)}$;

步骤 3 按公式(6)更新中心 $v_i^{(l+1)}$;

步骤 4 如果 T 等于最小,则停止;否则如果 $\max_i\|v_i^{(l+1)}-v_i^{(l)}\|<\varepsilon$ 或者 l>M,调整退火系数 T=T-ΔT,转至步骤 2。

4 实验结果及分析

首先说明评价文本聚类性能的评价标准,然后说明实验采用的各种数据集及实验设置,最后给出并分析对各数据集的实验结果。

4.1 算法性能评价准则

对于基于目标函数的算法的性能评价标准有内部评价标准和外部评价标准,文中实验将采用内外两个标准对各算法的聚类性能评价。内部评价标准往往同目标函数值相关,也即算法优化的目标。文中采用的平均余弦相似度 ACS 值作为算法的内部评价标准,若 v_i 为归一化的类 K_i 的中心,则 ACS 定义如下

$$ACS=\frac{1}{N}\sum_{i=1}^{K}\sum_{x_j\in K_i}x_j^T v_i \tag{10}$$

其中 N 为样本个数,ACS 值越大说明数据和各中心向量总的紧密度越高。此外对于文本聚类实验而言,文本的类标往往是已知的,故算法的外部评价标准一般采用互信息量 NMI(Normalized Mutual Information)值(Strehl&Ghosh,2002)[15]。NMI 值的具体定义如下:假设 X 代表已知的文本类标随机变量,Y 代表聚类结果的类标随机变量,则

$$NMI=\frac{I(X;Y)}{\sqrt{H(X)\cdot H(Y)}} \tag{11}$$

其中 I(X;Y)为变量 X,Y 的互信息量,H(X),H(Y)为变量 X 和 Y 的熵。文中采用 NMI 值的一般形式[11,12]

$$NMI=\frac{\sum_{h,l}n_{h,l}\log\left(\frac{n\cdot n_{h,l}}{n_h n_l}\right)}{\sqrt{\left(\sum_h n_h\log\frac{n_h}{n}\right)\left(\sum_l n_l\log\frac{n_l}{n}\right)}} \tag{12}$$

其中 n 代表样本数,n_h,n_l 分别代表已知类中包含的真实样本数和聚类结果的类中包含的样本数,$n_{h,l}$ 表示已知类的样本数和实际聚类结果的样本数一致的个数。因为聚类往往不预先知道聚类数,故使用 NMI 值可以较好地评价不同聚类数时算法的性能,NMI 值越高表示聚类结果越准确,NMI 值为 1 说明聚类结果与类标完全一致。

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

4.2 实验数据集说明

实验采用 20-Newsgroups[16]的数据集及部分来自 CLUTO[17]文本聚类工具箱的 8 个数据集。数据集含有的样本数从 690 个到 19 949 个不等，数据维数最小的为 8 261 维，最大的为 43 586 维，实际聚类数最小的为 3 个，最大的为 20 个。从以上特征看出这些数据集较好地反映了文本数据集所具有的特征。其中 NG20 数据平均地选自 20 个不同新闻组，经过的 Bow toolkit[18]对 20-Newsgroups 文本进行了预处理后含有 19 949 个向量文本数据。NG17-19 是 NG20 数据的一个子集，实际分类数为 3 类，每类包括将近 1 000 个来自政治新闻的文本，并根据这些新闻的特征不同分为 3 类，以往的聚类算法对该数据集的聚类结果表明，因为类与类之间有重叠导致对该数据集的聚类难度较高。其它数据的均来自 CLUTO 工具箱[17]，这些数据集已经预处理为向量文本数据。数据集的详细说明见表 1。

需要指出的是，表 1 中的 balance 是数据的平衡系数即包含最少文本数的类与包含最多文本数的类中的文本数之比，这个值反映了数据集内类与类之间的平衡性。实验中使用的 NG20，NG17-19，sports 数据集较为平衡，即每类中包括相近的样本数，而其他数据集的平衡度很差。

4.3 实验设置

实验平台是 P4 1.8 G CPU 和 384 M 内存，软件环境为 Window 2000，所有代码用 Matlab 6.5 实现。实验中将对 3 个文本聚类算法，球面 K 均值算法 SPKmeans，自退火复合 movMF 密度算法 DA-VMFS 和 ME-SPKM 的聚类性能做比较。其中自退火混合 movMF 密度算法 DA-VMFS 是文献[11,12]提出的使用确定性退火技术的 movMF 算法，相关实验表明 DA-VMFS 算法可以明显提高文本聚类算法 movMF 的聚类性能，实验中采用的 DA-VMFS 参数同文献[11,12]。实验中极大熵球面 K 均值 ME-SPKM 算法中退火参数 T 取[20,60,100,160,200,400]。此外所有的算法均采用随机初始化的策略选取初始的聚类中心。

4.4 实验结果及分析

为了测试作者提出的最大熵球面 K 均值算法，分别对以上数据集固定聚类数和不同聚类数时的各算法聚类效果进行比较。实验中对每一种情况运行算法 20 次，并取其聚类结果的 NMI 平均值作为最终评价值做图，同时给出实验结果的具体 NMI 平均值和偏差表以及聚类结果的平均余弦相似度。实验表明，最大熵球面 K 均值算法对这些数据集都取得了令人满意的结果。

4.4.1 固定聚类数时各算法的聚类效果比较

从表 2 和表 3 中各算法对不同数据集的聚类结果 NMI 值可以看出，SPKMeans 算法对各数据集的聚类 NMI 值最低，其聚类效果明显低于其他 2 个聚类算法。使用 ME-SPKM 的聚类效果好于 DA-SPKM。

表 1 数据集的简要说明(n_d 代表文本数，n_w 代表词项，k 代表文本实际类数)

Table 1 Summary of text datasets (n_d is the total number of document, n_w is the total number of terms, k is the number of classes)

Data	Source	n_d	n_w	K	balance
NG20	20 Newsgroups	19 949	43 586	20	0.991
NG17-19	3 overlapping subgroups from NG20	2 998	15 810	3	0.998
reviews	San Jose Mercury (TREC)	4 069	18 483	5	0.098
sports	San Jose Mercury (TREC)	8 580	14 870	7	0.636
tr45	TREC	690	8 261	10	0.088
la1	LA Times (TREC)	3 204	31 472	6	0.290
la12	LA Times (TREC)	6 279	31 472	6	0.282
la2	LA Times (TREC)	3 075	31 472	6	0.274

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 2 数据集 NG20,NG17-19,reviews,sports,tr45 的实验 NMI 值

Table 2 NMI results on NG20,NG17-19,reviews,sports,tr45 datasets

	NG20	NG17-19	reviews	sports	tr45
K	20	3	5	7	10
SPKMeans	0.550±0.050	0.340±0.100	0.530±0.100	0.560±0.130	0.600±0.090
DA-VMFS	0.570±0.030	0.460±0.010	0.560±0.090	0.620±0.050	0.680±0.050
ME-SPKM	0.590±0.010	0.470±0.060	0.620±0.020	0.640±0.004	0.690±0.030

此外作者发现 ME-SPKMS 聚类算法的 NMI 的偏差在大多数情况下远小于 SPKMeans 和 DA-VMF 算法，这说明极大熵球面聚类的算法克服了对初始化的敏感性。在对高难度数据集 NG17-19 聚类时，作者发现算法 ME-SPKM 的 NMI 值最好可以达到 0.53，但 NMI 值偏差很大，其内在原因有待于进一步研究。

表 3 数据集 classic la1,la12,la2 datasets 的实验 NMI 值

Table 3 NMI results on classic la1,la12,la2 datasets

	classic	la1	la12	la2
K	4	6	6	6
SPKMeans	0.540±0.040	0.480±0.100	0.480±0.100	0.480±0.070
DA-VMFS	0.510±0.010	0.530±0.030	0.520±0.020	0.520±0.040
ME-SPKM	0.560±0.001	0.560±0.006	0.560±0.001	0.550±0.006

表 4 数据集 NG20,NG17-19,reviews,sports, tr45 的实验 ACS 值

Table 4 Average cosine similarity (ACS) results on NG20,NG17-19,reviews,sports,tr45 datasets

	NG20	NG17-19	reviews	sports	tr45
K	20	3	5	7	10
SPKMeans	0.158 7	0.153 1	0.213 1	0.245 0	0.378 4
DA-VMFS	0.158 3	0.153 4	0.214 0	0.249 2	0.379 8
ME-SPKM	0.160 2	0.153 5	0.214 5	0.247 6	0.384 7

表 5 数据集 classic la1,la12,la2 的实验 ACS 值

Table 5 Average cosine similarity (ACS) results on classic la1,la12,la2 datasets

	classic	la1	la12	la2
K	4	6	6	6
SPKMeans	0.150 5	0.173 8	0.174 7	0.180 3
DA-VMFS	0.152 3	0.173 0	0.175 5	0.182 8
ME-SPKM	0.152 7	0.175 8	0.176 2	0.186 8

表 4 和表 5 反映了不同聚类算法对不同数据集的聚类结果的平均余弦相似度 (ACS)，从表中可以看出 ME-SPKM 和 DA-SPKM 的 ACS 值要大于 SPKMeans 算法的 ACS 值。

4.4.2 不同聚类数时各算法的聚类效果比较

聚类算法往往不预先知道实际的聚类数，故作者对各算法在数据集取相应不同聚类类别时的聚类性能进行了比较，为保证实验的准确性，实验中对数据集取每一种聚类类别下运行算法 20 次，最后采用 20 次 NMI 的平均值作为该类的 NMI 值。图 1 至图 4 是算法对部分数据集在各种聚类数下的 NMI 值比较图，从图中可以看出 ME-SPKM 由于使用了最大熵策略来避免局部最小点，故其在不同聚类数的聚类性能都好于 SPKmeans。

4.4.3 各算法的聚类时间比较

表 6 是各算法对部分数据集在实际聚类数的聚类时间比较，从表中可以看出，SPKmeans 算法由于使用的是硬划分，其聚类时间远小于其他 2 个算法，ME-SPKM 运行时间稍高于 DA-VMF。基于以上各节的分析，作者认为对不同的数据聚类任务可以采用不同的聚类算法，当数据聚类任务要求时间性较强时，可以采用运算速度快的 SPKmeans；对于要求聚类精确度高的数据聚类任务可以采用聚类性能高的 ME-SPKM。

表 6 算法运行时间

Table 6 Run time results

	NG20	NG17-19	reviews	sports	classic
K	20	3			
SPKmeans/s	84.5	6.4	12.2	25.8	5.1
DA-VMFS/s	1 686.7	50.5	220.1	335.0	96.5
ME-SPKM/s	3 177.9	55.2	326.3	449.2	125.5

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

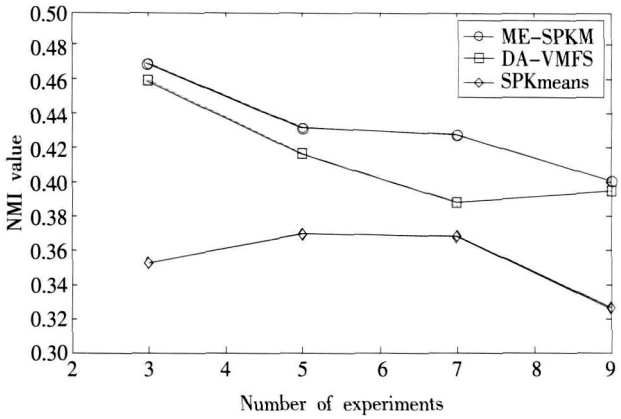

图 1 数据集 NG17-19 聚类结果

Fig.1 Results on NG17-19 dataset

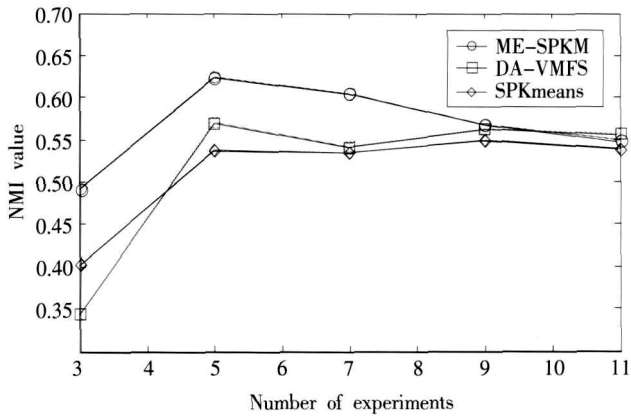

图 2 数据集 reviews 聚类结果

Fig.2 Results on reviews dataset

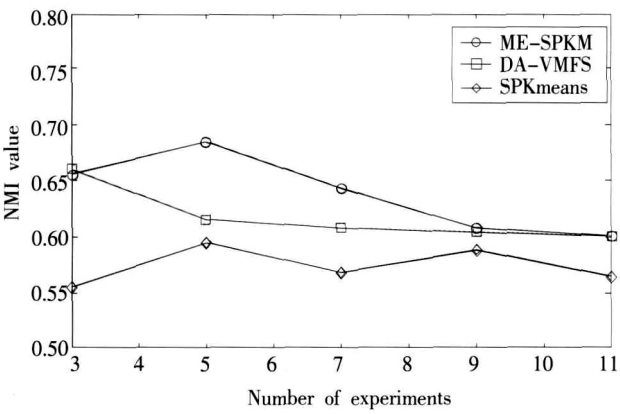

图 3 数据集 sports 聚类结果

Fig.3 Results on sports dataset

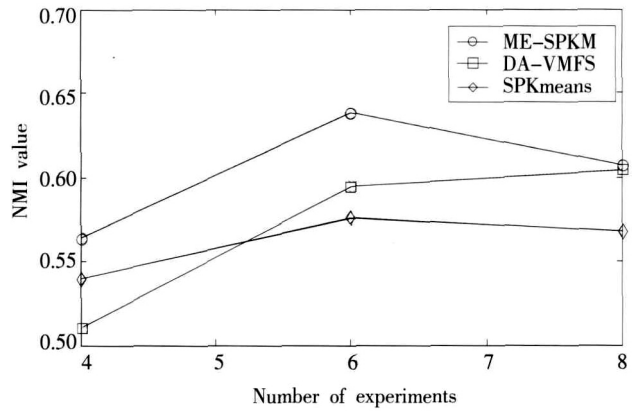

图 4 数据集 classic 聚类结果

Fig.4 Results on classic dataset

5 结束语

文章将极大熵原理运用到球面 K 均值算法的目标函数中，从而提出了适合文本聚类分析的极大熵球面 K 均值聚类算法 ME-SPKM,大量实验表明该算法可以有效地对各种文本数据集聚类，其聚类性能好于传统的一些聚类算法。此外通过实验作者也发现了一些问题：对高难度数据集聚类时,ME-SPKM 在取得较高聚类效果的同时为何其偏差较大；如何进一步在保证 ME-SPKM 聚类效果的情况下降低其聚类时间。以上问题也是作者下一步的研究目标。

References:

[1] Berry M W. Survey of text mining, clustering, classification and retrieval[M]. [S.l.]: Springer, 2003.

[2] Mardia K V, Jupp P. Directional statistics[M]. 2nd ed. [S.l.]: John Willey and Sons Ltd, 2000.

[3] Dhillon I S, Fan J, Guan Yuqiang. Efficient clustering of very large document collections[C]//Data Mining for Scientific and Engineering Applications Norwell. MA: Kluwer, 2001.

[4] Rose K, Gurewitiz E, Fox G A. Deterministic annealing approach to clustering[J]. Patter Recognition Letters, 1990, 11:589-594.

[5] Karayiannis N B. MECA: maximum entropy clustering algorithm[C]//Proc IEEE Int Conf Fuzzy Syst, Orlando, FL, 1994,1:630-635.

[6] Li R P, Mukaidono M A. Maximum entropy approach to fuzzy clustering[C]//Proc of the 4th IEEE Intern Conf on Fuzzy Systems, Yokohama, Japan, 1995,4:2227-2232.

[7] Yang Wenguang, Wang Dingxing, Zheng Weimin, et al. Research of a clustering model and algorithm by use of

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

deterministic annealing[J]. Journal of Software,1999,10 (6): 663-667.

[8] Adwait R. Maximum entropy models for natural language ambiguity resolution[D]. University of Pennsylvania, 1998.

[9] Li Ronglu, Wang Jianhui, Chen Xiaoyun, et al. Using maximum entropy model for Chinese text categorization[J]. Journal of Computer Research and Development, 2005, 42 (1): 94-101.

[10] Banerjee A, Dhillon I S, Ghosh J, et al. Generative model based clustering of directional data[C]//Conference on Knowledge Discovery in Data, 2003: 19-28.

[11] Shi Zhong, Joydeep G J. An unified framework for model-based clustering[J]. Journal of Machine Learning Research, 2004,4 (6): 1001-1037.

[12] Shi Zhong, Ghosh J. Generative model-based document clustering: a comparative study[J]. Knowledge and Information Systems, 2005,8 (3): 374-384.

[13] Strehl A, Ghosh J, Mooney R. Impact of similarity measures on web-page clustering[C]//Proc 7th Natl Conf on Artificial Intelligence, Workshop of AI for Web Search AAAI Press, 2000: 58-64.

[14] Mitechell T. Machine learning[M]. [S.l.]: McGraw Hill, 1997.

[15] Alexander S, Joydeep G. Cluster ensembles—a knowledge reuse framework for combining partitions[J]. Journal of Machine Learning Research, 2002,3: 583-617.

[16] http://kdd.ics.uci.edu./databases/20newsgroups/20newsgroups.html.

[17] ftp://www.cs.umn.edu/~karypis/CLUTO/flies/datasets.tar.gz.

[18] Mow: a toolkit for statistical language modeling, text retrieval, classification and clustering[EB/OL]. http://www.cs.cmu.edu/mccallum/bow.

附中文参考文献：

[7] 杨文广,王鼎兴,郑纬民,等.一种利用确定性退火的聚类模型与研究算法[J].软件学报,1999,10 (6):663-667.

[9] 李荣陆,王建会,陈晓云,等.使用最大熵模型进行中文文本分类[J].计算机研究与发展,2005,42 (1):94-101.

修宇 (976-),男,安徽芜湖人,硕士研究生,主要研究方向为模式识别、数据挖掘。

XIU Yu was born in 1976. He is a master graduate. His current research interests include pattern recognition and data mining.

王士同 (1964-),男,江苏扬州人,教授,博士生导师,中国计算机学会高级会员,主要研究方向为人工智能、模式识别、数据挖掘、神经网络及生物信息学。

WANG Shitong was born in 1964. He is a professor and PhD supervisor. He is the senior member of China Computer Federation. His current research interests include artificial intelligence, pattern recognition, data mining, neural networks and bioinformatics.

朱林 (983-),男,安徽安庆人,硕士研究生,主要研究方向为图像处理、模式识别。

ZHU Lin was born in 1983. He is a master graduate. His current research interests include image processing and pattern recognition.

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net