余弦度量和适应度函数改进的聚类方法

施侃晟[1]，刘海涛[1]，白英彩[1]，宋文涛[1]，洪亮亮[2]

(1. 上海交通大学电子与电气工程系 上海 徐汇区 200030; 2. 中国孵化中心 杭州 310053)

【摘要】K-均值算法因其简单和高效性，在文本聚类中占有重要地位。针对传统的K-均值算法对初始点敏感、易陷入局部最优的问题，结合遗传算法已经成为一种趋势。在充分发挥K-均值算法的高效性的同时，该文利用遗传算法的全局自适应优化特点克服了对初始点敏感的问题。同时，以余弦度量评价对象间的相似性并以此构造新的遗传算法适应度函数、收敛准则以及遗传算法种群更新方式，提高了K-均值和遗传算法这种结合方式的聚类精度，并增强了该结合算法的稳定性。

关 键 词 遗传算法; 适应度函数; K-均值算法; 相似性度量; 文本聚类

中图分类号 TP18 文献标志码 A doi:10.3969/j.issn.1001-0548.2013.04.017

Text Clustering Method with Improved Fitness Function and Cosine Similarity Measure

SHI Kan-sheng[1], LIU Hai-tao[1], BAI Yin-cai[1], SONG Wen-tao[1], and HONG Liang-liang[2]

(1. College of Electronic and Electric Engineering, Shanghai Jiaotong University Xuhui Shanghai 200030;

2. China Incubating Center Hangzhou 310053)

Abstract The traditional K-means algorithm is widely used because of its simplicity and efficiency. However, it is sensitive to the initial point and easy to fall into local optimum. In this paper, we use cosine measure to evaluate the similarity between objects and construct a new fitness function of genetic algorithm and the new convergence criterion for K-means algorithm. Experimental results show that the new method enhances the clustering accuracy and stability for the combination of K-means and genetic algorithm.

Key words genetic algorithm; fitness function; K-means algorithm; similarity measurement; text clustering

文本聚类作为一种无监督的机器学习方法，由于不需要训练过程及预先对文档手工标注类别，因此具有一定的灵活性，已成为对中文文本信息进行有效地组织、摘要和导航的重要手段，为越来越多的研究人员所关注[1]。典型的文本聚类方法有多种，其中K-均值算法因其简单和高效性，在文本聚类中占有重要地位[2]，但它对聚类初始中心点的选取比较敏感且易陷入局部最优，文献[3]提出了用语义信息改善该问题的方法。目前，有研究者将遗传算法和K-means算法相结合克服初始点敏感问题[4-9]。遗传算法是一种通过模拟自然进化过程搜索最优解的方法，它只需检测少量结构就可反映搜索空间较大的区域，便于实时处理，同时具有较强的稳健性可避免陷入局部最优。所以，K-均值与遗传算法的结合是一种趋势。

本文进一步以余弦度量评价对象间的相似性，并以此构造遗传算法的适应度函数、收敛准则来更新遗传算法种群，提高了K-均值与遗传算法这种结合方式的聚类精度和稳定性。

1 改进的文本聚类算法

针对K-均值与遗传算法相结合的趋势，给出新的提高该种结合方式的聚类精度和稳定性的算法设计和实际操作步骤。

1.1 相似性度量设计

聚类过程中，两个对象间的相似性计算是非常重要的，相似性度量准则的优劣很大程度上影响了聚类的效果。在向量空间模型下，可以借助向量之间的某种距离表示文本间的相似度。目前研究者已提出了许多方法来评价同一个特征空间中的两个对象间的距离，然而并非所有的度量在各种情况下都是适用的，如对象的数据类型是分类的和连续的情

收稿日期：2011－08－29; 修回日期：2012－04－18

基金项目：国家自然科学基金(61073150)

作者简介：施侃晟(1966－)，男，教授，主要从事信息挖掘、云计算和物联网方面的研究.

况。文本属于连续型数据，目前有许多方法如Pearson相关、Jaccard系数[10]以及欧几里得距离在文本聚类中非常有效，这是普遍接受的。但是，在基于文本的环境以及采用TF-IDF加权策略下，用余弦相似性度量判断文本间的相似性其性能最优[11]。

VSM(vector space model)[12]用于表征文本，每个文本 D_i 均被映射成文本特征的权重向量。文本的每个特征项均被赋予一个权重 w_k，以表示该特征项在该文本中的重要程度。文本可表示为：

$$D = D(t_1, w_1; t_2, w_2; \cdots; t_n, w_n) \tag{1}$$

其中特征项 t_k 的权重为 w_k，$1 \leqslant k \leqslant n$。为了简化分析，不考虑 t_k 在文本中的先后次序，要求 t_k 不重复。文本 D_1 和 D_2 的余弦相似度定义为与之对应的两个VSM，即 V_{D_1} 和 V_{D_2} 间的余弦角为：

$$S(V_{D_1}, V_{D_2}) = \frac{V_{D_1} \cdot V_{D_2}}{\left\| V_{D_1} \right\| \cdot \left\| V_{D_2} \right\|} \tag{2}$$

式(2)着重从形状考虑 D_1 和 D_2 之间的关系。当两个向量方向相近时，夹角余弦值较大；反之则较小。

1.2 编码方案设计

一方面，采用浮点数编码，能克服二进制编码计算量大的问题，能相对加快求解时间；另一方面，基于VSM的文本聚类是实数域的求解问题。基于此，本文采用浮点数编码。以 $\text{chrom}_i = (c_1, c_2, \cdots, c_K)$ $(1 \leqslant i \leqslant p$，$p$ 表示种群大小，文中设 p 为偶数) 表示每条染色体。给定 n 个文本，对每个文本取特征权重最大的前num个词作为各对应文本的特征项。从这些文本提取的特征项中，不同的特征项个数之和标记为 m，对每个文本采用这些不同的特征项构造VSM。假如该文本中没有这些特征项，那么VSM中对应元素值为0；否则按照TF-IDF方法计算该文本中这些特征项的权值。$\boldsymbol{c}_j$ $(1 \leqslant j \leqslant K)$则是与VSM长度相同的向量，表示该文本集的第 j 个聚类中心。由于对基于VSM的权值进行了归一化处理，所以编码时向量 $\boldsymbol{c}_j$ $(1 \leqslant j \leqslant K)$中的每个元素都是0和1之间的实数。染色体长度 l 为 $K \times m$。

1.3 适应度函数设计

采用余弦度量评价文本间的相似度，余弦值越大，对象间的非相似度越小；余弦值越小，对象间的非相似度越大。由1.2节的编码方案，对染色体中的 K 个中心点，按照余弦度量依次计算各个样本与 K 个中心点之间的相似度，将各样本归入与之相似度最大的那个中心点所代表的类。若一个中心点与离它最近的那些对象的相似性越大，即非相似性程度越小，那么其适应度值就应该越大。本文设计适应度函数为：

$$f = \frac{1}{1 + \sum_{j=1}^{K} \sum_{x_i \in \boldsymbol{C}_j} (1 - \cos(x_i, \boldsymbol{c}_j))} \tag{3}$$

1.4 遗传算子

1) 选择算子。

本文采用赌轮方式。首先根据式(3)选取适应度值最大的selectnum，再按赌轮方式将个体 chrom_i 以概率：

$$p_s(\text{chrom}_i) = \frac{f(\text{chrom}_i)}{\sum_{i=1}^{p} f(\text{chrom}_i)} \tag{4}$$

选取，直到种群中有 p 个染色体为止。

2) 交叉算子。

本文采用均匀交叉。均匀交叉通常优于单点交叉，因均匀交叉可能得到更多的图式结果。

交叉过程如下：从当前种群中依次选取两条染色体(如第1条和第2条、第3条和第4条……依此类推)，每次选好两条染色体后，都随机产生一个0和1之间的实数cr。若 $\text{cr} < p_c$ 则按照式(5)执行交叉操作；否则不变。

$$\begin{aligned} x' &= \text{cr} \cdot x + (1 - \text{cr}) \cdot y \\ y' &= \text{cr} \cdot y + (1 - \text{cr}) \cdot x \end{aligned} \tag{5}$$

式中，x 和 y 为从当前群体中按上述方法选取的两条染色体；x' 和 y' 为交叉后得到的新的染色体。

3) 变异算子。

为结合实际问题，本文采用单点变异。

变异方法如下：首先将[0,1]划分成 l 个等宽度的子区间，然后依次选择一个染色体(如第1个、第2个……依此类推)，每选择一个染色体均随机产生一个0和1之间的实数mr，若该mr落在第mloc $(1 \leqslant \text{mloc} \leqslant l)$个子区间且 $\text{mr} < p_m$ (p_m 为变异概率)，那么用一个[0,1]之间的随机数取代该染色体的第mloc位的值；否则该染色体保持不变。

1.5 算法收敛准则设计

K-means算法收敛准则：

$$\begin{aligned} J &= \sum_{j=1}^{K} \sum_{x_i \in \boldsymbol{C}_j} (1 - \cos(x_i, \boldsymbol{c}_j)) \\ j &= 1, 2, \cdots, K \end{aligned} \tag{6}$$

聚类中心为：

$$\boldsymbol{c}_j' = \frac{1}{|\boldsymbol{C}_j|} \sum_{\boldsymbol{X} \in \boldsymbol{C}_j} \boldsymbol{X} \tag{7}$$

设给定样本集X={ $x_1, x_2, \cdots, x_n$ }，给定聚类中心个数K，并设K个聚类中心分别为$c_1, c_2, \cdots, c_K$。聚类可描述为：对于给定数据集的n个点$x_1, x_2, \cdots, x_n$，按照它们之间相似性程度将其划分为K个簇$C_1, C_2, \cdots, C_K$，将样本集X中各个样本根据最大最小距离原则分配给簇C_i。

1.6　算法终止条件

实际应用中，采用如下终止条件：

1) 固定最大迭代次数iternum。当算法执行了iternum次遗传进化后停止。

2) 根据算法的收敛程度。假如群体的最大适应度值连续几代不发生变化或者没有明显的变化，那么遗传算法停止。

1.7　算法实际操作步骤

输入：样本集U (样本数为n，样本维数为m)，聚类数K，种群大小p，交叉概率p_c，变异概率p_m，终止条件。

输出：聚类结果。

1) 文本预处理。

2) 文本表示。采用以词作为特征[7]、以TF-IDF作为加权策略的VSM，并进行特征空间降维。

3) 编码。

4) 初始化种群。随机产生浮点数作为染色体基因，种群可以表示为$P = (\text{chrom}_1, \text{chrom}_2, \cdots, \text{chrom}_p)$。

5) 计算各个体适应度函数值。

6) 选择、交叉和变异。

7) 对变异后得到的种群中的每个染色体执行一次K-均值算法，这里的K-均值中采用余弦度量评价样本间的相似性，并以式(6)作为收敛准则，采用式(7)更新遗传算法变异后的群体作为下一次迭代的种群。

8) 判断是否满足算法终止条件，是则进入步骤9)；否则转步骤5)。

9) 输出聚类结果。

2　改进算法的仿真与分析

算法实测中采用Iris[14]数据集进行仿真。

2.1　实验参数设置

遗传算法的种群大小$p \in [20,150]$，本文取p=100，交叉概率$p_c \in [0.6, 0.9]$，$p_c = 0.86$，变异概率$p_m \in [0.001,\ 0.05]$，$p_m = 0.02$，精英个体数selectnum = 3，最大进化代数iternum = 100，聚类数$K = 3$。

2.2　算法仿真及分析

软件环境：操作系统 Windows XP，编译软件Matlab7.0.0。硬件环境：Pentium(R)D CPU 2.80 GHz，内存2 GB。

将采用欧几里得度量的GA(K-均值算法更新变异后群体)和采用式(7)更新变异后的群体分别记为EGHCM、EGAM；将采用余弦度量的GA(采用K-均值算法更新变异后群体)记为CGM(cosine measured genetic algorithm based method)，即本文算法。各算法的收敛准则也采用对应的度量进行运算。为了测试算法的有效性，对上述算法在聚类准确率和对初始条件的敏感性等方面进行了对比实验。

采用Iris(150个对象，4个维，共3类)数据集，将数据集中各对象的值均进行归一化处理后再进行测试。本文采用Huang(1998)提出的聚类精度度量来度量聚类的准确性，其定义为：

$$r = \frac{1}{n}\sum_{l=1}^{k} a_l \tag{8}$$

式中，a_l表示同时出现在类C_l和其相应的标记类中的对象数。

每种算法运行100次。这里遗传算法的终止条件设为：若上一次迭代的最大适应度值与本次迭代的最大适应度值的差值小于0.000 001时，算法结束。各聚类算法的平均准确率如表1所示。从表1中可以看出，算法CGM的聚类精度最佳。为了更加详实地了解算法的效率，算法EGHCM、CGM和EGAM运行100次得到的聚类准确率的实际比例如表2所示(表中的顺序号是为实验过程中出现值的顺序，没有特殊含义)。

表1　聚类准确率

算法	平均准确率/(%)
EGHCM	89.31
EGAM	89.55
CGM	96.73

表2　聚类准确率对比表

ID	EGHCM	EGAM	CGM
1	0.893 3	0.893 3	0.966 7
2	0.893 3	0.893 3	0.966 7
3	0.893 3	0.893 3	0.966 7
4	0.9	0.9	0.973 3
5	0.893 3	0.893 3	0.966 7
6	0.893 3	0.893 3	0.966 7
7	0.886 7	0.9	0.9
8	0.9	0.9	0.9
9	0.666 7	0.893 3	0.893 3
10	0.893 3	0.893 3	0.893 3

将算法EGHCM、EGAM和CGM得到的适应度值进行比较，如图1所示。

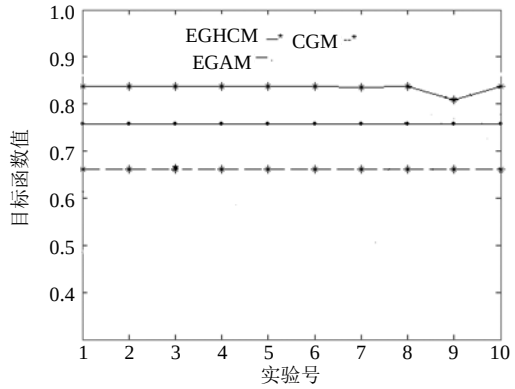

图1 对初始点的敏感性比较

从图中可以看出，本文算法CGM的稳定性高。

3 结 论

为克服传统K-均值算法对初始点敏感容易造成局部最优的问题，将K-均值算法与遗传算法相结合已经成为了一种趋势。这种结合充分发挥了K-均值算法的高效性以及遗传算法的全局优化能力。本文采用余弦度量评价对象，并由此设计新的适应度函数、收敛准则和种群更新方式，使得K-均值算法与遗传算法的结合有了更好的聚类精度，同时还得到了更高的稳定性。

本文算法已成功地应用于易合®主题情报系统以及易智童®儿童个性化早期教育系统中，并获得了发明专利[15]。在应用中，本文算法对文本形式的情报和文本形式的儿童教案进行自动的聚类，效果令人满意。

参 考 文 献

[1] 刘远超, 王晓龙, 徐志明, 等. 文档聚类综述[J]. 中文信息学报, 2006, 20(3): 55-62.
LIU Yuan-cao, WANG Xiao-long, XU Zhi-ming, et al. Survey of text clustering[J]. Journal of Chinese Information, 2006, 20(3): 55-62.

[2] 陈浩, 何婷婷, 姬东鸿. 基于K-mean聚类的无导词义消歧[J]. 中文信息学报, 2005, 19(4): 10-16 .
CHEN Hao, HE Ting-ting, JI Dong-hong. Unsupervised K-means clustering based on word sense disambiguation on hownet[J]. Journal of Chinese Information, 2005, 19(4): 10-16.

[3] SHI Kan-sheng, LI Le-min, LIU Hai-tao, et al. A linguistic feature based K-means text clustering method[C]// Proceedings of IEEE Cloud Computing and Intelligent Systems. [S.l.]: IEEE, 2011.

[4] 何婷婷, 戴文华, 焦翠珍, 等. 基于混合并行遗传算法的文本聚类研究[J]. 中文信息学报, 2007, 21(4): 55-60 .
HE Ting-ting, DAI Wen-hua, JIAO Cui-zhen, et al. Research of text clustering based on hybrid parallel genetic algorithm[J]. Journal of Chinese Information, 2007, 21(4): 55-60.

[5] 王敞, 陈增强, 袁著祉. 基于遗传算法的K-均值聚类分析[J]. 计算机科学, 2003, 30(2): 163-164.
WANG Chang, CHEN Zen-qiang, YUAN Zhu-zhi. K-means clustering based on genetic algorithm[J]. Computer Science, 2003, 30(2): 163-164.

[6] 赖玉霞，刘建平，杨国兴. 基于遗传算法的K均值聚类分析[J]. 计算机工程, 2008, 34(20): 200-202.
LAI Yu-xia, LIU Jian-pin, YANG Guo-xin. K-means clustering based on genetic algorithm[J]. Computer Engineering, 2008, 34(20): 200-202.

[7] 胡彧, 毕晋芝. 遗传优化的K-均值聚类算法[J]. 计算机系统应用, 2010(6): 52-55.
HU Yu, BI Jin-zhi. K-means clustering algorithm based on genetic optimization[J]. Computer System Application, 2010(6): 52-55.

[8] 王康, 颜雪松, 金建, 等. 一种改进的遗传K-均值聚类算法[J]. 计算机与数字工程, 2010(1): 18-20.
WANG Kang, YAN Xue-song, JIN Jian, et al. An improved genetic K-means clustering algorithm[J]. Computer and Digital Engineering, 2010(1): 18-20.

[9] 徐家宁, 张立文, 徐素莉, 等. 改进遗传算法的K-均值聚类算法研究[J]. 微计算机应用, 2010, 31(4): 11-18.
XU Jia-ning, ZHANG Li-wen, XU Shu-li, et al. Improved genetic algorithm based K-means clustering algorithm[J]. Microcomputer Application, 2010, 31(4): 11-18.

[10] SHAMEEM M U S, FERDOUS R. An efficient K-means algorithm integrated with jaccard distance measure for document clustering[C]//First Asian Himalayas International Conference on Internet. Kathmandu: [s.n.], 2009, 1-6.

[11] JORIS D H, JORIS V, PAUZ A V, et al. Pairwise-adaptive dissimilarity measure for document clustering[J]. Information Sciences, 2010(180): 2341-2358.

[12] WEI Song, LI Cheng-hua. PARK S C. Genetic algorithm for text clustering using ontology and evaluating the validity of various semantic similarity measures[J]. Expert Systems with Applications, 2009(36): 9095-9104.

[13] SALTON G, WONG A, YANG CS. A vector space model for information retrieva[J]. Communications of the ACM, 1975, 18(11): 613-620.

[14] BLAKE C, MERZ C J. Machine learning repository [DB/OL]. [2011-06-04]. http://www.ics.uci.edu/~mlearn/MLRepository.html.

[15] 施章祖, 施侃晟. 计算机辅助报告与知识库产生的方法: 中国, ZL200810063295.1[P]. 2008-10-06.
SHI Zhang-zu, SHI Kan-sheng. Computer aided method of generating report and knowledgebase: China, ZL200810063295.1[P]. 2008-10-06

编 辑 张 俊