基于遗传算法的自适应文本模糊聚类研究*)

戴文华[1] 焦翠珍[1] 何婷婷[2]

(咸宁学院计算机系 咸宁 437100)[1] (华中师范大学计算机科学系 武汉 430079)[2]

摘 要 针对FCM聚类算法对初始聚类中心的选择敏感,以及聚类数C难以确定的问题,提出一种基于遗传算法的自适应文本模糊聚类方法。该方法首先将文档集合表示成向量空间模型,并采用一种新型的可变长染色体编码方案,随机选择文本向量作为初始聚类中心形成染色体,然后结合FCM算法的高效性和遗传算法的全局优化能力,通过遗传进化,有效地避免了局部最优解的出现,同时得到了优化的聚类数目和聚类结果。实验表明该算法是一种精确高效的文本聚类方法。

关键词 遗传算法,FCM聚类,文本聚类,模糊聚类

Research on Adaptive Text Fuzzy Clustering Based on Genetic Algorithm

DAI Wen-hua[1] JIAO Cui-zheng[1] HE Ting-ting[2]

(Department of Computer, Xianning College, Xianning 437100, China)[1]

(Department of Computer Science, Central China Normal University, Wuhan 430079, China)[2]

Abstract As Fuzzy C-means Clustering Algorithm was sensitive to the choice of the initial cluster centers and it's difficult to determine the cluster number, we proposed an Adaptive Text Fuzzy Clustering Method Based on Genetic Algorithm. According to the principle of Vector Space Model, documents were represented as vectors. Then we adopted a new strategy of variable-length chromosome encoding and randomly chose initial clustering centers to form chromosomes among document vectors. Combining the efficiency of Fuzzy C-means Algorithm with the global optimization ability of Genetic Algorithm, the local optimal solution was avoided and the optimum number and the optimum result of cluster were obtained by means of genetic evolution. Experiments indicated that this algorithm was efficient and accurate.

Keywords Genetic algorithm, Fuzzy C-means clustering, Text clustering, Fuzzy clustering

1 引言

文本聚类是一种有效的文本挖掘方法。典型的文本聚类方法有多种,其中FCM算法[1,2]因其简单和高效性,且具有模糊分类功能,在文本聚类中占有重要地位。由于FCM算法在聚类中心和模糊分类矩阵的计算过程中采用了启发式方法,因而有效地降低了算法复杂度,提高了运算速度。也正是因为这样的原因,使得该算法对初始聚类中心的选择较为敏感,易于陷入局部最优解。

同时,传统的FCM算法是在聚类数C确定的前提下进行的,然而实际聚类问题中C值的确定往往非常困难,只能根据经验大致确定。这种估值方法必将带来算法精确度的下降。

为了避免文本聚类对初始聚类中心选择的敏感性和聚类数C难于确定的问题,我们提出一种基于遗传算法的自适应文本模糊聚类方法。通过该方法,我们能在达到文本聚类目的的同时得到经过优化的聚类数目,因此聚类的精确度也将得到极大改善。

2 文本表示

在文本聚类之前,首先应将文本转换为易被计算机理解的形式,然后通过判断文本间的相似性,将文档集合划分为不同的簇。

文本聚类问题中常采用向量空间模型(Vector Space Model,VSM)[3]进行文本表示。在该模型中,文本空间被看作一组正交特征向量组成的向量空间,每个文档d_i均被映射成文本特征的权重向量:

$$v(d_i)=(w_1(d_i),w_2(d_i),\cdots,w_n(d_i)) \quad (1)$$

其中n表示文本特征抽取时所选用的特征数,$w_j(d_i)$表示第j个文本特征在文档d_i中的权重。在向量空间模型中,文本特征权重的计算一般采用$tf.idf$方法[4]:

$$w_j(d_i)=tf_{ij}*\log_2(N/N_j+0.01) \quad (2)$$

其中tf_{ij}为第j个文本特征在文档d_i中出现的频次,N为文档集合中的总文档数,N_j为文档集合中出现第j个文本特征的文档数。为了减小文档长度差异对文本相似度计算的影响,通常将每个向量归一化到单位向量,最后得到文本特征权重的计算公式如下:

$$w_j(d_i)=\frac{tf_{ij}*\log_2(N/N_j+0.01)}{\sqrt{\sum_{k=1}^{n}(tf_{ik})^2*[\log_2(N/N_k+0.01)]^2}} \quad (3)$$

3 FCM聚类方法

对于N个n维数据样本$X=\{x_1,x_2,\cdots,x_N\}$,如果要将它们划分成$C(2\leqslant C\leqslant N)$个类别,我们可以采用FCM聚类方法进行聚类,具体步骤如下:

① 选定C个初始聚类中心$Z=\{z_1,z_2,\cdots,z_C\}$,设置终止

*)国家自然科学基金(No.60442005,No.60673040);国家社会科学基金(No.06BYY029);教育部重点研究项目(No.105117);湖北省教育厅科研重点项目(No.D200728002)。**戴文华** 副教授,硕士。

迭代误差 ε，初始化迭代次数 $g=1$；

② 计算模糊隶属度矩阵 $U_{N\times C}$。当 $Y_i=\{j\mid 1\leqslant j\leqslant C, \| x_i-z_j \|=0\}=\Phi$ 时，矩阵各元素按公式(4)计算；设 $\overline{Y}_i=\{1,2,\cdots,C\}-Y_i$，当 $Y_i\neq\Phi$ 时，矩阵各元素按公式(5)计算；

$$u_{ij}=\frac{\| x_i-z_j \|^{-2}}{\sum_{r=1}^{C}\| x_i-z_r \|^{-2}} \tag{4}$$

$$u_{ij}=\begin{cases}0 & (\forall j\in\overline{Y}_i)\\ 1 & (\forall j\in Y_i)\end{cases} \tag{5}$$

③ 更新聚类中心 Z，聚类中心各元素按公式(6)计算：

$$z_j=\frac{\sum_{i=1}^{N}u_{ij}^2x_i}{\sum_{i=1}^{N}u_{ij}^2} \tag{6}$$

④ 如果 $\| Z^{(g)}-Z^{(g-1)} \|<\varepsilon$，即上一次迭代聚类中心与本次迭代聚类中心基本不变时停止迭代并转步骤⑤，否则令 $g=g+1$，转步骤①。

⑤ 如果要求将样本精确划分到一个类别，则按照模糊隶属度矩阵将样本按最大隶属度划分到不同类别；如果样本可属于多个类别，则设定分类阈值 η，按模糊隶属度矩阵将样本划分到不同类别。

4 基于遗传算法的自适应文本模糊聚类方法

由美国 Michigan 大学的 J. H. Holland 教授于 20 世纪 60 年代提出的遗传算法(Genetic Algorithm，GA)[5,6]，是模拟自然界生物进化机制的随机化搜索算法，适用于处理传统搜索方法难于解决的复杂优化问题。

在文本聚类问题中，如果采用遗传算法对 FCM 算法中的不同初始聚类中心进行优化选择，必定能找到最优初始聚类中心，从而得到最优的聚类结果。同时，如果我们在遗传进化过程中，采用可变长染色体编码方案动态选择聚类数目，聚类精确度将会得到极大提高。

通过以上分析，我们将 FCM 算法和遗传算法相结合，提出一种基于遗传算法的自适应文本模糊聚类方法。该方法能充分利用 FCM 算法的高效性和遗传算法的全局搜索能力，有效地保证文本聚类的效率和精度。具体算法如下：

① 设置交叉概率 Pc、变异概率 Pm 和最大遗传代数 $Gnum$；

② 产生 Gsize 条染色体，形成初始种群；

③ 根据每条染色体表示的初始聚类中心，将文本进行 FCM 聚类，然后根据聚类结果计算各染色体的适应度；

④ 对种群进行选择、交叉和变异操作；

⑤ 判断是否满足遗传终止条件，如果满足则退出遗传并转⑥，否则转③；

⑥ 将适应度最高的染色体作为初始聚类中心，将该初始聚类中心对应的 FCM 聚类结果作为最终聚类结果。

将遗传算法应用于文本聚类问题时，必须考虑到在算法的实现过程中，编码方案、适应度函数、遗传算子、种群的初始化和停止标准等都是影响算法效率的非常关键的因素。下面将就这些问题进行讨论。

4.1 可变长染色体编码方案

在聚类问题中，由于聚类中心数难以确定，只能凭经验设置，这种凭经验确定的聚类中心数会对聚类结果产生偏差，因此我们采用遗传算法以动态方式来确定聚类中心数，相应的染色体采用可变长染色体编码方案。

可变长染色体有两种编码方式：第一种方式，染色体基因由初始聚类中心对应的文本在文档集中的编号表示；第二种方式，染色体基因由初始聚类中心对应的文本向量直接表示。随着文本数量的增加，染色体长度逐渐加长，第一种染色体编码方式占用的内存相对较小，具有更大的优势，因此本文将采用第一种染色体编码方案。具体编码形式为

$$CH=\{ch_1,ch_2,\cdots,ch_t\} \tag{7}$$

其中 t 为某条染色体的编码长度，对不同的染色体，t 的值是变化的。$ch_i(i=1,2,\cdots,t)$ 为第 i 个聚类中心对应的文本在文档集中的编号，为一个[1，N]之间的自然数(N 为待聚类文本数)。

4.2 适应度函数

由于染色体采用可变长编码，因此聚类中心的个数并不固定，所以适应度函数与定长编码时的适应度函数有所区别。具体定义如下：

$$Fit(Ind)=\frac{1}{1+\sum_{j=1}^{Len(Ind)}\sum_{i=1}^{N}u_{ij}^2\| x_i-z_j \|^2} \tag{8}$$

其中 $Len(Ind)$ 为个体 Ind 的染色体长度。

文本间的距离采用公式(9)进行计算：

$$Dis(d_i,d_j)=\| d_i-d_j \|=\sqrt{\sum_{k=1}^{n}(x_{ik}-x_{jk})^2} \tag{9}$$

其中 x_{ik}，x_{jk} 分别表示文档 d_i 和 d_j 的第 k 个文本特征的权重，n 为特征词个数。

4.3 选择算子

选择策略对算法性能的影响起到举足轻重的作用。在种群进化过程中，我们采取精英保留策略，首先保留遗传过程中的精英个体，让它们不参与交叉和变异操作而直接进入下一代群体。然后采用轮盘赌法，由适应度函数对应的概率分布确定把当前群体中的其它个体按选择概率 Ps 抽出，并进行交叉和变异，以提高群体的平均适应度。选择概率 Ps 由公式(10)进行计算：

$$Ps(i)=\frac{Fit(i)}{\sum_{j=1}^{Gsize}Fit(j)} \tag{10}$$

其中 Gsize 为种群大小，$Fit(i)$ 为第 i 号染色体的适应度。

4.4 插入删除交叉算子

针对可变长染色体编码，我们特意设计了插入删除交叉算子，以适应遗传进化过程中染色体长度的变化。

插入删除交叉算子的主要思想是：将一个染色体的一段基因删除，并将这段基因插入另一个染色体的某一位置。插入删除交叉算子操作过程可用图 1 表示。

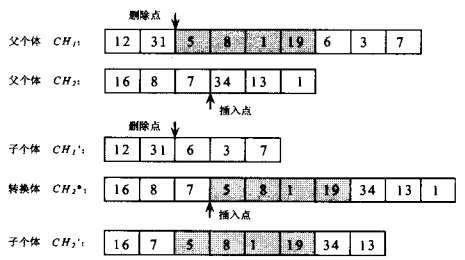

图 1 插入删除交叉操作示意图

具体操作步骤如下：

① 以 CH_1 作为被删除染色体，以 CH_2 作为被插入染色体，计算染色体 CH_1 和 CH_2 的长度 t_1 和 t_2；

② 如果 $t_2=\lceil\sqrt{N}\rceil$，则重新选择染色体 CH_2，直到 $t_2<\lceil\sqrt{N}\rceil$；

其中 N 为待聚类文本数，$\lceil\sqrt{N}\rceil$ 为聚类个数的经验值[7]。其实，设置该值的目的只是为了加快算法速度，如果对算法精度

要求较高，则可适当放宽该值的尺度。要求 $t_2<\lceil\sqrt{N}\rceil$ 是为了防止插入操作后染色体 CH_2 的基因由于超长截断而无变化。

③ 随机生成插入点位置 Ins、删除点位置 Del 和插入(或删除)长度 DLen；

其中插入长度与删除长度相等，均为 DLen。要求满足如下条件：

$0\leqslant Del<t_1, 0\leqslant Ins\leqslant t_2$ 且 $Len<t_1$

④ 将染色体 CH_1 从删除点开始，将长度 DLen 的基因段删除，并将该基因段插入染色体 CH_2 中；

⑤ 将染色体 CH_2 中的重复基因去除；

⑥ 如果染色体 CH_2 的长度超长，则对其进行截尾操作。

从图 1 可以看出，经过插入删除操作后，染色体长度分别由原来的 9 和 6 变成了 5 和 8，这种变化显然保证了进化过程中染色体的多样性，同时聚类数目也在进化过程中得到了动态改变，有利于遗传算法的优化和搜索。

针对染色体长度的动态变化，以及插入删除操作的频繁性，我们采用动态链表的方式存储染色体。这种存储方式插入删除操作快速，且具有链表长度的可动态变化性。

4.5 变异算子

染色体变异操作步骤如下：

① 计算染色体长度 *Len*；

② 随机产生一个[1，*Len*]之间的自然数 *C*，作为变异点个数；

③ $c=1$；

④ 随机产生一个不重复的[1，*Len*]之间的自然数，作为变异点；

⑤ 随机产生一个[0，1]之间的数 *r*，如果 $r\leqslant Pm$（*Pm* 为变异概率）转⑥，否则转⑦；

⑥ 随机产生一个[1，*N*]之间的在染色体中不存在的自然数，将父个体在变异点处的基因用这个自然数取代；

⑦ $c=c+1$；

⑧ 如果 $c>C$，退出变异，否则转④。

4.6 种群初始化

由于染色体长度可变，因此种群初始化方法具有自身的一些特点。具体步骤如下：

① 设置种群规模 *Gsize*；

② $I=1$；

③ 如果 $I\leqslant Gsize$，则转④，否则结束初始化；

④ 随机设置染色体长度 $Len(Len\leqslant\lceil\sqrt{N}\rceil)$；

⑤ 随机产生 *Len* 个[1，*N*]之间的不重复的自然数，形成一条染色体 *Ind*；

⑥ 判断染色体 *Ind* 是否已经在种群中存在，如果存在则转④，否则转⑦；

⑦ $I=I+1$；

⑧ 转③。

4.7 算法停止标准

在实际系统中，我们采取如下停止标准：进化代数超过最大遗传代数 Gnum 或种群平均适应度值连续多代遗传后仍无明显变化，遗传算法停止。

5 实验及结果分析

为了验证本文所提出的基于遗传算法的自适应文本模糊聚类方法(标记为 ATFCGA)的实际性能，我们进行了下述实验。实验中各参数：种群大小 $Gsize=100$，最大进化代数 $Gnum=100$ 代，交叉概率 $Pc=0.86$，变异概率 $Pm=0.02$，精英个体数 $Elite=4$。

实验一　聚类划分的类别数正确率测试

为了验证本文所提出的基于遗传算法的自适应文本模糊聚类方法(标记为 ATFCGA)形成的聚类数目的有效性，我们在国家语委现代汉语语料库中抽取不同数量文本，组成 500 个文档集合。使用本文提出的 ATFCGA 文本聚类方法对上述文档集合进行聚类，聚类划分的类别数正确率达到 96.8%。该项实验说明本文提出的文本聚类方法能通过可变长染色体编码方案的遗传算法，动态优化文本聚类数目，从而获取较为精确的聚类数，有效地解决了文本聚类方法对聚类数 *C* 难于确定的问题。

实验二　文本聚类准确率测试

为了验证 ATFCGA 文本聚类算法的聚类准确性，我们在国家语委现代汉语语料库中抽取 350 篇文本组成文档集，其中政治类 125 篇、军事类 105 篇、经济类 120 篇，使用 ATFCGA 文本聚类方法对上述文档集进行聚类。实验中各参数同实验一，同时设定分类阈值 $\eta=0.30$。聚类结果如表 1 所示。

表 1　文本聚类结果

结果对比		文本模糊聚类结果						
		只属于政治类	只属于军事类	只属于经济类	同时属于政治类和军事类	同时属于政治类和经济类	同时属于军事类和政治类	同时属于政治类、军事类和经济类
原分类结果	政治类(125 篇)	58	2	3	19	38	2	3
	军事类(105 篇)	2	70	1	18	2	11	1
	经济类(120 篇)	2	1	63	3	32	17	2

从表 1 可以看出，抽取的 350 篇文本中，通过文本模糊聚类后，只有 18 篇文本聚类错误，聚类准确率达到 94.86%，具有较高的精确度。这正是由于本文所提出的聚类方法通过进化计算，动态挖掘文本聚类数目，同时优化了初始聚类中心的选择，因此在避免聚类数盲目估计的同时能获取最优初始聚类中心，从而保证了聚类的精确度。

此外，本文提出的文本模糊聚类方法能将待聚类文本合理地聚类，使得同一篇文本同时属于多个类别，从而使聚类结果更能客观地反映文本聚类的正确划分。

结束语　本文通过提出一种基于遗传算法的自适应文本模糊聚类方法，克服了传统 FCM 算法对初始聚类中心选择的敏感，以及聚类数 *C* 难以确定的问题，充分发挥了 FCM 算法的高效性和遗传算法的全局优化能力。有效地均衡了算法对聚类空间的探索和开发能力，实验证明该算法是一种高精度的文本聚类方法。

本文所提出的文本聚类方法，有效地利用了遗传算法的全局优化能力，但是没有充分利用遗传算法的并行性。下一步我们将研究并使用并行遗传算法对文本聚类方法进行改进，以充分利用遗传算法的并行性，增强算法的效率和精确度。

参 考 文 献

[1]　Zhao Y，Karypis G. Criterion Functions for Document Clustering

Experiments and Analysis[R]. Department of Comp. Sci. & Eng University of Minnesota, 2001: 1-40

[2] Steinbach M, Karypis G, Kumar V. A comparison of Document Clustering Techniques[R]. Department of Comp. Sci. & Eng University of Minnesota, 2000: 1-20

[3] Salton G, Wang A, Yang C S. A vector space model for automatic indexing[J]. Communication of the ACM, 1975, 18(11): 613-620

[4] Luo Xiao, Sun Maosong, Tsou B K. Covering Ambiguity Resolution in Chinese Word Segmentation Based on Contextual Information[A]// Proceedings of the 19th COLING. 2002: 598-604

[5] Li J, Kwan RSK. A fuzzy genetic algorithm for driver scheduling [J]. European Journal of Operational Research, ELSEVIER, 2003, 147(2): 334-344

[6] Andre J, Siarry P, Dognon T. An improvement of the standard genetic algorithm fighting premature convergence in continuous optimization[J]. Advances in Engineering Software, 2001, 32: 49-60

[7] Ramze R M, Lelieveldt B PF, Reiber J H C. A new cluster validity indexes for the fuzzy c-mean[J]. Pattern Recognition Letters, 1998, 19: 237-246

[8] Makoto I, Takenobu T. Hierarchical Bayesian clustering for automatic text classification[R]. Department of Computer Science Tokyo Institute of Technology. Tech Rep, TR95-0015. 1995

(上接第 111 页)

[9] Mousseau V, Slowinski R. Inferring an ELECTRE-TRI model fromassignment examples. Jounal of Global Optimization, 1998, 12(2): 157-174

[10] Mousseau V, Slowinski R, Zielniewicz P. A user-oriented implementation of the ELECTRE-TRI method integrating preference elicitation support. Computers and Operations Research, 2000, 27(7/8): 757-777

[11] Greco S, Matarazzo B, Slowinski R. Rough Approximation of Preference Relation by Dominace relations. European Journal of Operational Research, 1999, 117: 63-68

[12] Greco S, Matarazzo B, Slowinski R. Conjoint measurement and rough sets approach for multicriteria sorting problems in presence of ordinal data// Colorni A, Paruccini M, Roy B, eds. EUR Report. Joint Research Centrem The European Commission, Ispra, 2001: 114-141

[13] Greco S, Matarazzo B, Slowinski R. Rough set approach to multi-attributechoice and ranking problems. ICS Research Report 38/95. Warsaw University of Technology. Warsaw, 1995

[14] Greco S, Matarazzo B, Slowinski R. An algorithm for induction decision rules consistent with the dominance principle// Rough Sets and Current Trends in Computing. Proceedings of the 2nd Int. Conference RSCTC2000. Banff, October, 2000. LNAI 2005, Springer, 2001: 304-313

[15] Greco S, Slowinski R, Stefanowski J. Incremental versus Non-incremental Rule Induction for Multi -criteria Classification// Peters J F, et al., eds. Transactions on Rough Sets, LNCS 3135, 2004: 33-53

[16] Stefanowski J. Rough set based rule induction techniques for classification problems// Proc. 6 th European Congress on Intelligent Techniques and Soft Computing. Aachen, Sept. 1998, 1: 109-113

[17] Grzymala-Busse J W. LERS-a system for learning from examples base on rough sets// Slowinski R, ed. Intelligent Decision Support. Handbook of Applications and Advances of the Rough Sets Theory. Kluwer Academic Publishers, Dordrecht, 1992: 3-18

[18] 李克文,等. 基于序关系的粗糙集// 2003 年人工智能会议. 邮电大学出版社, 1359-1363

[19] 吴毅民,叶东毅. 基于优势关系的粗糙集中的一种求核算法. 计算机科学, 2004, 31(10. A): 138-139

[20] 袁修久,何华灿. 优势关系下广义决策约简和上近似约简. 计算机工程与应用, 2006, 5: 4-7

[21] Greco S, Matarazzo B, Slowinski R, et al. Variable Consistency Model of Dominance-Based Rough Sets Approach// Ziarko W, Yao Y, eds. RSCTC 2000, LNAI 2005, 2001: 170-181

[22] Giove S, Greco S, Matarazzo B, et al. Variable Consistency Monotonic Decision Trees// Alpigini J J, et al., eds. RSCTC2002, LNAI2475, 2002: 247-254

[23] 胡寿松,何亚群. 粗糙决策理论与应用. 北京:北京航空航天大学出版社, 2006

[24] 王隆鑫,刘财辉,王黔英. 基于偏序关系的 Rough 集定义和 Rough格. 计算机科学, 2002, 29(9. 专刊): 110-111

[25] 徐伟华,张文修. 基于优势关系下的协调近似空间. 计算机科学, 2005, 32(9): 164-165

[26] 徐伟华,张文修. 基于优势关系下不协调目标信息系统的知识约简. 计算机科学, 2006, 33(2): 182-184

[27] Yue Chaoyuan, Yao Shengbao, Zhang Peng, et al. Rough Approximationof a Preference Relation for Stochastic Multi-attribute Decision Problems// Wang L, Jin Y, eds. FSKD 2005, LNAI 3613, 2005: 1242-1245

[28] Dembczynski K, Pindur R, Susmaga R. Generation of Exhaustive Set of Rules within Dominance-based Rough Set Approach. http://www. elsevier. nl/locate/ entcs/volume82. html

[29] Zaras K. Rough approximation of a preference relation by a multi-attributestochastic dominance for determinist and stochastic evaluation problems. European Journal of Operational Research, 2001, 130: 305-314

[30] Zaras K. Rough approximation of a preference relation by a multi-attributestochastic dominance for determinist and stochastic and fuzzy problems. European Journal of Operational Research, 2004, 159: 196-206

[31] Dembczynski K, Greco S, Slowinski R. Second-Order Rough Approximations in Multi-criteria Classification with Imprecise Evaluations and Assignments// Slezak D, et al. eds. RSFDGrC 2005, LNAI 3641, 2005: 54-63

[32] 何亚群,胡寿松. 基于粗糙集的空军航材供应点的偏好选址. 系统工程理论与实践, 2003, 23(7): 95-99