基于模糊粒度计算的K-means文本聚类算法研究

张 霞[1,2] 王素贞[1] 尹怡欣[2] 赵海龙[2]
(河北经贸大学计算机中心 石家庄 050061)[1] (北京科技大学信息工程学院 北京 100083)[2]

摘 要 传统的K-means算法对初始聚类中心非常敏感,聚类结果随不同的初始输入而波动,算法的稳定性下降。针对这个问题,提出了一种优化初始聚类中心的新算法:在数据对象的模糊粒度空间上给定一个归一化的距离函数,用此函数对所有距离小于粒度 d_λ 的数据对象进行初始聚类,对初始聚类簇计算其中心,得到一组优化的聚类初始值。实验对比证明,新算法有效地消除了传统K-means算法对初始输入的敏感性,提高了算法的稳定性和准确率。

关键词 模糊,粒度,K-means,文本聚类,归一化距离函数
中图法分类号 TP391 **文献标识码** A

Research of Text Clustering Based on Fuzzy Granular Computing

ZHANG Xia[1,2] WANG Su-zhen[1] YIN Yi-xin[2] ZHAO Hai-long[2]
(Computer Center, Hebei University of Economics and Business, Shijiazhuang 050061, China)[1]
(School of Information Engineering, University of Science and Technology Beijing, Beijing 100083, China)[2]

Abstract The traditional K-means is very sensitive to initial clustering centers and the clustering result will wave with the different initial input. To remove this sensitivity, a new method was proposed to get initial clustering centers. This method is as follows: provide a normalized distance function in the fuzzy granularity space of data objects, then use the function to do a initial clustering work to these data objects who has a less distance than granularity d_λ, then get the initial clustering centers. The test shows this method has such advantages on increasing the rate of accuracy and reducing the program times.

Keywords Fuzzy, Granular computing, K-means, Text cluster, Normalized distance function

1 引言

聚类分析是数据挖掘的一项基本任务,它的目标是在没有任何先验知识的前提下,将数据聚集成不同的类,通过聚类,人们能够发现全局的分布模式以及数据属性之间有趣的相互关系。K-means算法是聚类算法的主要算法之一,它是一种基于划分的聚类算法,即采用迭代更新的方法,其最终结果是促使目标函数值取得极小值,达到较优的聚类效果。传统的K-means算法是由随机选取的初始聚类中心出发逐步逼近给定目标。由于启发式算法的特点以及局部极小值的存在,算法对初始聚类中心点敏感,初始值若选择不当便会使聚类结果陷入局部极小点,从而导致不同的初始值得到不同的聚类结果,并且这些结果一般都不是全局最优解。在实际应用中,由于初始中心的不同而导致聚类结果的波动,会严重影响其稳定性和可操作性。因此,K-means算法初始聚类中心的优化对于算法的研究和具体的应用都具有非常重要的意义。

粒度计算作为一种新兴的方法被应用于知识发现过程中。美国著名数学家Zadeh给出了粒度计算的许多描述,他认为粒度计算像一把大伞,覆盖了所有有关粒度的理论、方法论、技术和工具的研究。其基本思想出现在许多领域,如粗糙集(Rough Set)、模糊集(Fuzzy Set)、分治法(Divide and Conquer)、区间分析(Interval Analysis)、机器学习(Machine Learning)、数据挖掘(Data Mining)等,并且已得到广泛而又有效的应用。张钹院士和张铃教授又将模糊集合论引入到粒度计算中,利用模糊等价关系实现了商空间模型的推广,模糊粒度计算通过对粒度计算方法学、数学框架、信息粒化算法、不同模型下的粒度计算模型的研究和应用,在模糊数学和粒度计算的领域取得了一些卓有成效的研究结果。

本文根据模糊粒度的理论基础,定义一个归一化的距离函数 $d(x_i,x_j)$,它唯一确定一个模糊等价关系 $\underset{\sim}{R}$,通过控制粒度 d_λ 得到等价关系 $\underset{\sim}{R}$,产生K-means算法的初始聚类,并通过距离的计算得到一组初始聚类中心。通过实验比较证明,基于模糊粒度计算优化初始聚类中心的方法有效地消除了K-means算法对初始值的敏感,并且在准确率和迭代时间上都远远优越于传统的K-means算法。

2 模糊粒度聚类基础

定义1 设 $\underset{\sim}{R}$ 为 X 上的一个模糊关系,若 $\underset{\sim}{R}$ 满足下列条件:

1) 自反性:$\underset{\sim}{R}(x,x)=1, \forall x\in X$

到稿日期:2009-03-09 返修日期:2009-07-20 本文受国家自然科学基金项目(60374032),河北省教育厅科研计划项目(2009116)资助。
张 霞(1975—),女,博士,讲师,主要研究方向为人工智能、数据挖掘,E-mail:zhang_xia_xia@yahoo.cn;王素贞(1964—),女,博士,教授,主要研究方向为移动计算、网络安全;尹怡欣 男,教授,博士生导师,主要研究方向为人工智能、人工生命;赵海龙 男,博士,主要研究方向为人工智能等。

2) 对称性：$\underset{\sim}{R}(x,y)=\underset{\sim}{R}(y.x),\forall x,y\in X$

3) 传递性：$\underset{\sim}{R}^2\subseteq\underset{\sim}{R}$

则称R是X上的一个模糊等价关系。

定理1 R是X上的模糊等价关系的充分必要条件是：对任意的$\lambda\in[0,1]$，截关系R_λ是X上的等价关系。

有关证明请参考文献[2]。

定理2 设在X的某一商空间$[X]$上给定一个归一化的距离函数$d(\cdot,\cdot)$，又设

$B_\lambda=\{(x,y)\mid d(x,y)\leqslant\lambda,\lambda\geqslant 0\}$

D_λ是以B_λ为基的等价关系，令$R_\lambda=D_{1-\lambda}$，则$\{R_\lambda\mid 0\leqslant\lambda\leqslant 1\}$唯一确定$X$上的一个模糊等价关系$\underset{\sim}{R}$，它以$R_\lambda$为截关系。

证明：令D_λ对应的商空间为$S(\lambda)$，由D_λ的定义易得，当$0\leqslant\lambda_1<\lambda_2\leqslant 1$时，有$S(\lambda_2)$是$S(\lambda_1)$的商空间，故$\{S(\lambda)\mid 0\leqslant\lambda\leqslant 1\}$在商空间包含关系下构成了一个有序链。所以$R_\lambda$满足：当$0\leqslant\lambda_1<\lambda_2\leqslant 1$时，$R_{\lambda_1}<R_{\lambda_2}$，故由$\{R_\lambda\mid 0\leqslant\lambda\leqslant 1\}$唯一确定$X$上的一个模糊等价关系$\underset{\sim}{R}$，它以$R_\lambda$为截关系。证毕。

以上两个命题得出在X的某一商空间$[X]$上给定一个归一化的距离函数$d(\cdot,\cdot)$，设$\{d(x,y)\leqslant\lambda,\lambda\geqslant 0\}$，则对任意的$\lambda\in[0,1]$，截关系$R_\lambda$是$X$上的等价关系。根据不同的粒度值$\lambda$，得到不同的聚类。

3 基于模糊粒度计算的K-means文本聚类

3.1 文本的预处理

3.1.1 特征项的抽取

特征项的抽取是指从一特征项集合中选出一部分最有代表性的特征，从而降低向量空间维数，简化计算，防止过分拟合。本实验中我们选择信息增益(Information Gain)的特征项抽取方法。信息增益是其包含信息量的度量。对没有先验知识的待聚类文本进行特征项抽取，词条t_k对待聚类文本的信息增益$IG(t_k)$为：

$$IG(t_k)=H(D)-H(D|t_k)$$

其中，D表示文本集合，文本$d_i\in D$。D的信息熵为：

$$H(D)=-\sum_{d_i\in D}(P(d_i)\times\log_2(p(d_i)))$$

词条t_k的条件熵为：

$$H(D|t_k)=-\sum_{d_i\in D}(P(d_i|t_k)\times log_2(p(d_i|t_k)))$$

$IG(t_k)$反映了t_k所包含的信息量。将$IG(t_k)$由大到小排序，由排序的结果可以设置阈值，做截断处理，排在前面的对应的上下文候选特征词最终可取得列选特征的资格。

3.1.2 文本特征表示

在向量空间模型(VSM)中，每一个文本d都表示为空间内的一个向量或者空间点，一般表示为：

$$V(d)=(w(t_1),w(t_2),\cdots,w(t_n))$$

其中，n为文本空间的维数；$w(\cdot)$是词条t_k在文本向量中的权重。设tf_{ij}表示词i在第j篇文本中出现的频度，df_i为文本集中包含词i的文本数目，$ndocs$为文本集文本总数。考虑到文本长度对权值的影响，将各项的权值规范到$[0,1]$之间。

$$w_{ij}=\frac{tf_{ij}\times\log(\frac{ndocs}{df_i}+0.01)}{\sqrt{\sum_{i=1}^{n}[tf_{ij}\times\log(\frac{ndocs}{df_i}+0.01)]^2}}$$

3.2 模糊粒度计算在文本聚类中的应用

3.2.1 归一化距离函数的定义

基于模糊粒度计算的文本聚类中采用如下的距离函数：

$$d(d_i,d_j)=1-\frac{\sum_{k=1}^{n}(w_{ik}\wedge w_{jk})}{\frac{1}{2}\sum_{k=1}^{n}(w_{ik}+w_{jk})}$$

其中，$w_{ik}\wedge w_{jk}=\min(w_{ik},w_{jk})$，$d_i\in D,d_j\in D$。

这样的距离函数满足下面的条件：

(1) 正定性：$d(d_i,d_j)\geqslant 0,\forall d_i,d_j\in D$，有$d(d_i,d_j)\in[0,1]$，是一个严格单调增加的函数，并且当$\|(d_i-d_j)\|\to\infty$时，$d(d_i,d_j)\to 1$；

(2) $d(d_i,d_j)=0$当且仅当$d_i=d_j$；

(3) 对称性：$d(d_i,d_j)=d(d_j,d_i)$；

(4) 归一化：$d(d_i,d_j)\in[0,1]$。

所以$d(d_i,d_j)$是归一化的距离函数。

3.2.2 模糊粒度计算文本聚类算法

基于模糊粒度计算的K-means聚类算法如图1所示。本算法将模糊粒度计算聚类化为一个在距离空间(X,d_λ)上的计算，通过归一化的距离函数，唯一确定一个模糊等价关系R。将$d(x_i,x_j)\leqslant d_\lambda$的归为一类，得出普通等价关系$R$，粒度值$d_\lambda$由粗变细的过程，产生了动态聚类的结果。通过对粒度$d_\lambda$的控制，得到一个初始聚类结果。对初始聚类结果计算其中心，得到一组优化的聚类初始值。

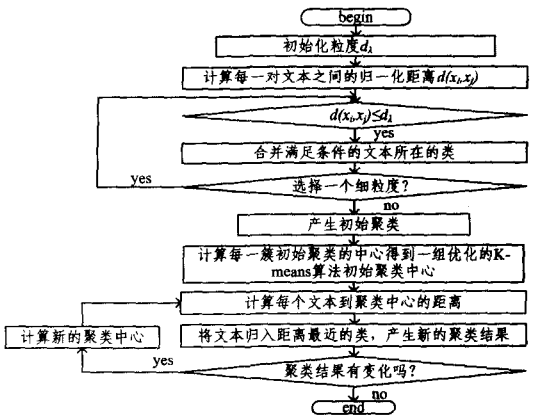

图1 基于模糊粒度计算的文本聚类算法

3.3 实验结果

从搜狗实验室的文本语料库中选取财经类、IT类、军事类各10篇共30篇文本，记为D，其中财经类$D1=\{d_1,d_2,d_3,d_4,d_5,d_6,d_7,d_8,d_9,d_{10}\}$，军事类$D2=\{d_{11},d_{12},d_{13},d_{14},d_{15},d_{16},d_{17},d_{18},d_{19},d_{20}\}$，IT类$D3=\{d_{21},d_{22},d_{23},d_{24},d_{25},d_{26},d_{27},d_{28},d_{29},d_{30}\}$。

目前，文本领域的一些常用的衡量分类算法效果的评价标准主是要查准率(Presicion)和查全率(Recall)。查准率也称准确率，是指文档集中正确归类的文档数占所有被分入该类的文档总数的百分比；查全率也称召回率，是指文档集中正确归类的文档数占该类文档总数的百分比。

如表1所列，首先使用传统K-means的方法随机选取聚类中心，随着初始值输入的不同，聚类结果产生波动，准确率差距很大，迭代次数总体比较高。如果选取离这3个簇中心最近的d_8,d_{16},d_{25}点作为初始中心输入，即表中的第3组，K-means聚类结果的正确率和迭代次数均有比较好的效果，但是在未知聚类结果的前提下，一般很难达到。第4组初始聚类中心为已知的3个簇的中心点，其聚类结果为最优效果。

表1 聚类结果对比表

中心点	类别类	查准率	查全率	迭代次数

随机初始聚类中心 d_6,d_9,d_{22}	D1	85.71%	60%	2
	D2	0%	0%	
	D3	52.63%	100%	
随机初始聚类中心 d_4,d_6,d_8	D1	100%	100%	6
	D2	80%	40%	
	D3	60%	90%	
离簇中心最近的点 d_8,d_{16},d_{25}	D1	100%	100%	2
	D2	81.82%	90%	
	D3	88.89%	80%	
三个簇的中心点	D1	100%	100%	2
	D2	83.33%	100%	
	D3	100%	80%	
基于模糊粒度计算的K-means算法初始聚类中心	D1	100%	100%	2
	D2	83.33%	100%	
	D3	100%	80%	

实验中，当粒度 $d_\lambda=0.05\sim0.04$ 之间时，得到初始聚类，计算其中心得到优化的 K-means 初始值。由表 1 可知，基于模糊粒度计算的方法得到的 K-means 初始聚类中心，正确率高，迭代次数少。其聚类结果与理想的最优效果基本一致。

结束语 1）通过实验对比证明，基于模糊粒度计算的初始聚类中心有效地消除了 K-means 算法对于初始输入的敏感性，提高了算法的稳定性和准确率。2）本算法利用模糊粒度计算的理论基础，定义归一化距离函数，通过粒度的控制得到优化的初始聚类中心，方法简单，便于在实际应用中操作；并且由文本间归一化距离产生的相似矩阵与文本的属性维数无关，算法的时间复杂度和空间复杂度都较低，算法的实用性得到了保证。

参 考 文 献

[1] Zadeh L A. The key roles of information granulation and fuzzy logic in human reasoning[C]//Proc of the Fifth IEEE International Conference on Fuzzy Systems. 1996(1):8-11

[2] 肖位枢. 模糊数学基础及应用[M]. 北京：航空工业出版社，1992:50-54

[3] Yao Yiyu. Granular computing: Basic issues and possible solutions[C]//Proc of the 5th Joint Conference on Information Sciences. Atlantic(NJ,USA). 2000:186-189

[4] 张铃，张钹. 问题求解理论及应用(第二版)[M]. 北京：清华大学出版社，2007:66-80

[5] 张铃，张钹. 模糊商空间理论(方法)[J]. 软件学报，2003,14(4):770-776

(上接第 188 页)

opt 的总体性能要优于 ACS-3-opt 和 MMAS-3-opt。

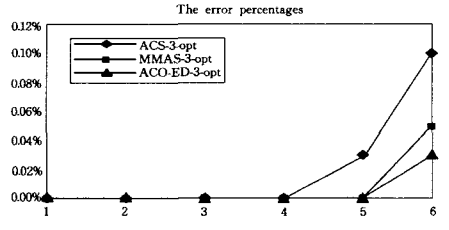

图 1 3 种算法的 Error%结果比较

图 2 是 ACO-ED-3-opt 求解 pcb442 问题的结果，图中曲线表示每次迭代中的最好解。从图上可以看出解空间没有过早收敛的情况，解的质量较高。

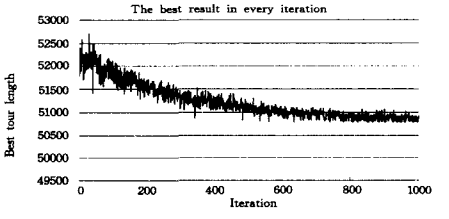

图 2 ACO-ED-3-opt 求解 pcb442 问题的结果

结束语 本文提出了一种新的融合分布估计的蚁群优化算法(ACO-ED)，用于求解组合优化问题。ACO-ED 把分布估计算法思想首次引入到蚁群优化算法中，避免了解空间过早收敛的局限性。另外，在 ACO-ED 算法中采用了局部搜索策略，进一步提高了算法性能。新算法具有结构简单、搜索效率高、求解速度快且能够获得高质量的全局近似最优解等优点。仿真实验表明该方法可行性强，性能结果令人满意。

我们进一步的工作是：(1) 把 ACO-ED 和局部搜索方法合理结合并应用到其他的组合优化问题中；(2) 用高阶概率分布模型来评估优质解的分布情况，并尝试应用到更多难解的优化问题中。

参 考 文 献

[1] Colorni A, Dorigo M, Maniezzo V. Distributed Optimization by Ant Colonies[C]//Proceedings of ECALSi - European Conference on Artificial Life. Elsevier Publishing, 1991:134-142

[2] Dorigo M, Manieszo V, Colorni A. The Ant System: Optimization by a Colony of Cooperating Agents[J]. IEEE Transacttons on Systems, Man, and Cybernetics - Part B, 1996, 26(1):29-41

[3] Dorigo M, Gambardella L M. Ant Colony System: A Cooperative Learning Approachto the Traveling Salesman Problem[R]. 96-05, INDIA. UniversitC Libre de Bruxelles, 1996

[4] Stuztle T, Hoos H. Improvements on the Ant System: Introducing MAX-MIN Ant System[M]//Smith G D, Steele N C, eds. Artificial Neural Networks and Genetic Algorithms. 1998:245-249

[5] 吴庆洪，张纪会，徐心和. 具有变异特征的蚁群算法[J]. 计算机研究与发展，1999,36(10):1240-1245

[6] Dorigo M, Di Caro G. The ant colony optimization meta-heuristic [C]//Corne D, Dorigo M, Glover F, eds. New Ideas in Optimization. London, UK: McGraw-Hill, 1999:11-32

[7] Larrañaga P, Lozano J A. Estimation of Distribution Algori - thms. A New Tool for Evolutionary Computation[M]. Boston: Kluwer Academic Publishers, 2002

[8] 周树德，孙增圻. 分布估计算法综述[J]. 自动化学报，2007,33(2):113-124

[9] Reinelt G. The Traveling Salesman: Computational Solutions for TSP Applications[C]// LNCS. volume 840, Springer Verlag, 1994

[10] Rosenkrantz D J, Stearns R E, Lewis P M. An analysis of several heuristics for the traveling salesman problem[J]. SIAM J. Comput., 1977, 6:563-581

[11] Pelikan M, Goldberg D E, Lobo F. A Survey of Optimization by Building and Using Probabilistic Models[R]. IlliGAL Report No. 99018. Urbana, Illinois: University of Illinois at Urbana-Champaign, Illi-nois Genetic Algorithms Laboratory, 1999

[12] Baluja S. Population-based Incremental Learning: A Method for Integrating Genetic Search Based Function Optimization and Competitive Learning[R]. CMU-CS-94-163. Pittsburgh, PA: Carnegie Mellon University, 1994

[13] MÄuhlenbein H, Paass G. From recombination of genes to the estimation of distributions I. Binary parameters. Parallel Problem Solving from Nature[C]//PPSN IV, Berlin, 1996:178-187

[14] Lin S. Computer solutions of the traveling salesman problem [J]. Bell Syst. J., 1965, 44:2245-2269

[15] Gambardella L, Dorigo M. Solving Symmetric and Asymmetric TSPs by Ant Colonies[C]//IEEE Conference on Evoluttonary Computation (ICEC'96). IEEE Press, 1996