初始中心优化的 K-Means 聚类算法*)

K-Means Clustering Algorithm with Refined Initial Center

李 飞 薛 彬 黄亚楼

(南开大学计算机科学与技术系 天津300071)

Abstract As one of the most popular clustering techniques, K-Means algorithm usually obtains locally optimal solutions due to its sensitivity to initial starting center. To overcome this problem, a genetic algorithm is used to search the initial center for K-Means algorithm. A concept of "Gene difference" is introduced to control the crossover operator and mutation operator in genetic algorithm. Experiments on standard database of UCI show that the proposed method can efficiently improve the clustering result.

Keywords K-Means algorithm, Genetic algorithm, Gene difference

1. 引言

聚类分析(clustering)是人工智能研究的重要领域。聚类方法被广泛研究并应用于机器学习、统计分析、模式识别以及数据库数据挖掘与知识发现等不同的领域。

各种聚类方法中,基于目标函数的 K-Means 聚类方法应用极为广泛,根据聚类结果的表达方式又可分为硬 K-Means (HCM)算法、模糊 K-Means 算法(FCM)和概率 K-Means 算法(PCM)[1]。各种 K-Means 算法都以确定的目标函数来测度聚类的效果,最佳的聚类效果对应于目标函数的极值点。由于目标函数局部极小值点的存在[2]以及算法的贪心性,导致聚类结果对初始中心敏感,往往达不到全局最优。

针对 K-Means 算法对初始聚类中心的敏感问题,可以选取不同的初始值多次执行该算法,然后选取最好的结果。显然,如果初始中心选取次数较少不能保证得到最优解,选取次数较多则会大大增加计算量。因此,许多研究者针对这个问题提出了改进的算法。P. S. Bradley 等人提出一种从聚类数据集随机选取多个子样本集的方法[3],对所选取的子样本集重复执行 K-Means 算法得到优化的初始聚类中心,然后再对整个聚类数据集采用 K-Means 方法聚类。但是,这种算法得到的只是一种"次优"的聚类结果,而且受子样本集选取方式的影响。其他研究者把随机算法的思想引入到初始聚类中心的选取中,如 CLARANS 算法[4],随机搜索多个初始聚类中心对应的局部极小值,比较得到全局最优解;最近,有学者将全局优化方法中的模拟退火技术应用于聚类分析,提出一种基于模拟退火算法的动态聚类算法[5,6]。

遗传算法是一种高效率的随机全局优化搜索算法,广泛应用于各种优化问题的求解,包括应用于聚类分析[7,8]。本文应用遗传算法的全局高效搜索能力来解决 K-Means 聚类算法对初始中心的敏感性问题。同时,针对标准的遗传算法本身的早熟收敛、误差平分等缺陷[9],根据所要解决问题的先验知识,引进基因差异度对标准遗传算法的交叉、变异算子进行修正,限制适应度差的个体的生成,以期提高算法的寻优能力。

2. 优化算法原理

传统的 K-Means 聚类算法,是一种已知聚类类别数的"无监督学习"算法。指定类别数为 C,对样本集合进行聚类,聚类的结果由 C 个聚类中心来表达。基于给定的聚类目标函数(或者说是聚类效果判别准则),算法采用迭代更新的方法,每一次迭代过程都是向目标函数值减少的方向进行,最终的聚类结果使目标函数值取得极小值,达到较优的聚类效果。

设聚类的样本集为:$X=\{x_i \mid x_i \in R^p, i=1,2,\cdots,N\}$,得到的 C 个聚类中心为 $z_1,z_2,\cdots,z_c$。令 $w_j(j=1,2,\cdots,C)$表示聚类的 C 个类别,则:

$$z_j=\frac{1}{N_j}\sum_{x\in w_j} x \tag{1}$$

定义目标函数:

$$J=\sum_{i=1}^{C}\sum_{j=1}^{n_i} d_{ij}(x_j,z_i) \tag{2}$$

其中 n_i 表示 w_i 类包含的样本个数,一般采用欧氏距离 $d_{ij}(x_j,z_i)=\sqrt{(x_j-z_i)^T(x_j-z_i)}$作为样本间的距离。欧氏距离适合于类内样本数据为超球形分布的情况。目标函数 J 为每个样本数据点到相应聚类中心的距离平方和,即聚类的最小均方误差。

传统的 K-Means 聚类算法(HCM 算法)流程如下:

(1)随机指定 C 个样本点 $z_1(1),z_2(1),\cdots,z_c(1)$为初始聚类中心;

(2)按照距离最近的原则,对样本集合聚类,确定每个样本的类属关系;

(3)使用公式(1),计算新的聚类中心 $z_1(k),z_2(k),\cdots,z_c(k)$($k$ 表示迭代次数);

(4)重复执行(2)—(4),直到聚类中心稳定为止。

显然,算法受初始值的影响很大,聚类的结果往往只是局部最优。即使对不同的初始值多次执行该算法,也只是在庞大的初值空间里简单地进行搜索,其结果也很难达到全局最优。

解决算法对于初始值的敏感性问题,相当于一个全局优化搜索问题。基于遗传算法的高效全局优化搜索能力以及内在的随机性和隐含的并行性,本文提出使用遗传算法来优化 K-Means 聚类算法的初始聚类中心。算法流程改变了传统 K-Means 算法随机选取中心的方法,即修改了上述 K-Means 算法流程的第(1)步,采用遗传算法来选择初始的聚类中心。

*)本文得到天津市自然科学基金项目资助(003600311)。李 飞 博士研究生,研究方向为智能信息处理、数据挖掘和数据仓库技术。黄亚楼教授,博士生导师,主要研究方向为智能机器人,智能信息处理,数据挖掘。

优化算法的流程为：

(1)采用遗传算法选择初始聚类中心，算法中的每个个体为K个样本；

(2)基于上述初始中心，采用K-Means算法得到聚类结果。

优化算法步骤(2)即是传统K-Means算法步骤(2)—(4)，可以使用HCM算法也可以使用FCM算法。优化算法的步骤(1)将使用一种修正的遗传算法，利用已知问题的先验知识，来克服标准遗传算法的早熟收敛等缺陷。下一节将详细描述本文采用遗传算法流程以及对算子所做的修正。

3. 遗传优化初始聚类中心算法

遗传算法是借鉴生物进化规律的随机搜索算法，对由多个个体组成的群体进行操作。与其他的搜索算法相比，具有更好的全局寻优能力。但是，标准遗传算法存在早熟收敛和误差平分等缺陷。同时，由于聚类问题的编码特征(染色体由一组聚类中心构成，基因之间无差别)，简单地采用标准遗传算法，容易造成算法的收敛性和解的全局最优性较差。针对问题的编码特征，本文引入一个基因差异度的概念，算法的执行过程中，动态地计算基因差异度值，交叉和变异算子使用该值以限制适应度差的个体产生，从而优化了遗传算法的执行性能。优化算法的流程如下：

(1)初始化：确定一个较小的正常数ε，交叉概率 P_c、变异概率 P_m 以及最大遗传代数 G_{max}。设遗传代数 $t=1$，随机产生 n 个个体，形成初始种群 $P(t)=\{V_k^{(t)}|k=1,2,\cdots,n\}$；

(2)计算种群中每个个体的目标函数 $J(k)$、适应度 $f(k)$ 和基因差异度 $d(k)$；

(3)计算 $\bar{J}(k)=\frac{1}{n}\sum_{k=1}^{n}J(k)$，如果 $|\bar{J}(t)-\bar{J}(t-1)|<\varepsilon$ 或者 $t=G_{max}$，选择最好的个体作为算法的结果，结束。否则 $t=t+1$；

(4)对 $t-1$ 代种群进行选择操作，形成种群 $P'(t-1)$；

对种群 $P'(t-1)$ 的个体以交叉概率 P_c 并根据所选个体的基因差异度进行交叉操作，形成种群 $P''(t-1)$；

对种群 $P''(t-1)$ 的个体以变异概率 P_m 并根据所选个体的基因差异度进行变异操作，形成第 t 代种群 $P(t)$，返回(2)。

对算法流程的解释如下：

·优化算法的编码方式是对聚类中心进行编码，一个聚类中心表示为个体上的一个基因，一组聚类中心组成一个个体。对每个基因仍采用浮点数表示，一个个体由 $C\times p$(p 为样本空间的维数)个浮点数组成。这种编码方式避免了二进制编码运算时的复杂的编码、解码操作以及二进制字符串有限长度对运算精度的影响。

·种群中每个个体的目标函数 $J(k)$ 由公式(2)计算，适应度函数定义为 $f(k)=1/(1+J(k))$，$J(k)$ 值小的个体，相应的适应度就高。

·个体的基因差异度 $d(k)$ 是本文根据聚类问题的特点提出的一个概念。定义个体 $V=\{z_1,z_2,\cdots,z_c\}$ 的基因差异度为：

$$d(V)=q\cdot\mathrm{Min}\{\|z_i-z_j\|i,j=1,2,\cdots,C,i\neq j\} \quad (4)$$

其中 q 为常数，且 $0<q<1$。个体内每个基因代表一个聚类中心。基因差异度表达了个体内聚类中心之间的差异程度。如果个体是一个优质个体，则其聚类中心之间的距离较远，基因差异度值 $d(V)$ 值较大；如果个体是一个劣质个体，存在距离非常接近聚类中心，则其基因差异度值 $d(V)$ 较小。因此，在执行交叉或变异操作改变个体的某个基因时，可以根据原来个体的基因差异度来决定是否接受新的个体，即子代个体基因差异度必须大于原来个体的基因差异度。

·选择算子采用转轮法。计算当前种群全部个体的适应度之和 $f_s=\sum_k f(k)$，定义 $p_k=\frac{f(k)}{f_s}(k=1,2,\cdots,n)$。令 $q_i=\sum_{j=1}^{i}p_j(i=1,2,\cdots,n)$，设 $q_0=0$，随机产生一个在区间[0,1]均匀分布的随机数 r，如果 r 满足条件：

$$q_{i-1}<r\leqslant q_i \quad (5)$$

则第 i 个个体选中，放入种群 $P'(t-1)$，重复 $n-1$ 次操作得到种群 $P'(t-1)$ 的 $n-1$ 个个体，种群 $P(t-1)$ 的适应度最优的个体直接放入种群 $P'(t-1)$。

·编码方式决定了基因之间没有顺序关系，因此在交叉和变异操作时，可能会出现相似基因重复出现的现象，具有相似基因的个体显然具有很差的适应度，采用基因差异度对交叉、变异加以控制，限制了不良个体的产生，避免出现振荡现象。

在交叉操作时，如果交叉得到的新个体的基因差异度小于原来个体基因差异度，则保留原来的个体；如果新个体的基因差异度不小于原来个体的基因差异度，则保留新的个体。

在变异操作时，如果变异得到的新个体的基因差异度小于原来个体的基因差异度，则重新执行变异操作，如果在指定的变异次数(可以设置最大变异执行次数为一常数)之内仍然得不到大于原来个体基因差异度的变异个体，则保留原来的个体，否则把新的个体放入下一代种群。

·交叉概率 P_c、变异概率 P_m 参数可设置为稍大一些的值，因为在交叉和变异算子中，减去了某些可能的交叉和变异操作。

4. 实验结果与分析

4.1 实验描述

为验证上述算法，采用的数据集controlprocess是UCI的一个标准数据库，描述一个进程控制系统的状态。该数据库有600条记录，每条记录有60个属性。

对3种算法的结果进行比较，这3种算法是：

(1)用本文所介绍的遗传算法获得初始中心，再采用K-Means算法进行聚类；

(2)不采用文中所介绍的个体的基因差异度d(k)，使用遗传算法获得初始中心，再采用K-Means算法进行聚类；

(3)直接采用K-Means算法进行聚类，因为初始中心的随机性会导致聚类结果的不同，为便于比较，本实验中聚类的初始中心采用遗传算法初始种群中最好的个体。

4.2 实验结果

按照上述实验方法，采用下列参数值：交叉概率 $P_c=0.3$，变异概率 $P_m=0.05$，初始种群中个体数 $n=7$，反复执行算法10次，得到表1的结果：

表1 三类算法聚类结果

	错分率结果	遗传代数	K-Means算法的迭代次数
本文提出的遗传算法	8.2%	13～29	6～9
不用s(k)的遗传算法	11.5%	12～42	7～15
直接采用K-Means算法	19.4%	无	17～28

从以上结果看出，使用本文介绍的遗传算法在不降低执

行效率的前提下，能明显提高聚类的效果。

结论 本文针对K-Means算法对初始中心敏感的问题，利用遗传算法的全局搜索能力，提出了一种将遗传算法用于K-Means算法的方法，以期得到全局最初的初始中心。本文提出了基因差异度的概念来控制遗传算法的交叉与变异，从而避免传统遗传算法所存在的早熟收敛和误差平分等缺陷。对标准数据库的测试的结果表明，这种方法能比较显著地改善聚类结果。

参考文献

1 Bezdek J C, et al. Multiple-Prototype Classifier Design. IEEE Trans Syst Man Cybern, 1998, 24(9):67～79

2 Selim S Z, Ismail M A. K-Means-Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality. IEEE Trans Pattern Analysis and Machine Intelligence, 1984, PAMI-6(1): 81～87

3 Bradley P S, Fayyad U M. Refining Initial Points for K-Means Clustering. Advances in Knowledge Discovery and Data Mining. MIT Press, 1996

4 Raymond T. Ng, Han Jiawei. Efficient and Effective Clustering Methods for Spatial Data Mining. In: Proc. of the 20th VLDB Conf. Santiago, Chile, 1994

5 Selim S Z, Alsultan K. A Simulated Annealing Algorithms for the Clustering Problem. Pattern Recognition, 1991, 24(10): 1003～1008

6 杨忠明，黄道，王行愚. 基于模拟退火的动态聚类算法. 控制与决策，1997，12，Suppl. 1：520～523

7 李茂军，樊韶胜，童调生. 单亲遗传算法在模式聚类中的应用. 模式识别与人工智能，1999，12(1)

8 王涛，沈谦，朱明星，张良震. 遗传与K-Means混合算法用于聚类分析. 模式识别与人工智能，1999，12(1)

9 徐金梧，刘纪文. 基于小生境技术的遗传算法. 模式识别与人工智能，1999，12(1)

10 王实，高文. 数据挖掘中的聚类算法. 计算机科学，2000，27(4)

(上接第45页)

参考文献

1 Johnson D B, Perkins C. Mobility Support in IPv6. draft-ietf-mobileip-ipv6-14. txt(July, 2000)

2 Gustafsson E, Jonsson A, Perkins C. Mobile IPv6 Regional Registrations. draft-ietf-mobileip-reg-tunnel-04. txt(Mar., 2001)

3 Glass S, Chandra M. Registration Revocation in Mobile IP. draft-ietf-mobileip-reg-revok-01. txt(July 2001)

4 Soliman H. Hierarchical MIPv6 mobility management(HMIPv6). draft-ietf-mobileip-hmipv6-04. txt(July, 2001)

5 Malki K E. Low Latency Handoff in Mobile IPv4. draft-ietf-mobileip-lowlatency-handoffs-v4-01. txt(May, 2001)

6 Dommety G. Fast Handovers for Mobile IPv6. draft-ietf-mobileip-fast-mipv6-02txt(July, 2001)

7 Chaskar H. Requirements of a QoS Solution for Mobile IP. draft-ietf-mobileip-qos-requirements-01. txt(August, 2001)

8 Omar H, Saadawi T, Lee M. Multicast Support for Mobile-IP with the Hierarchical Local Registration Approach. In: Proc. of the third ACM intl. workshop on Wireless mobile multimedia, Boston, MA USA, 2000

9 Harrison T, et al. Mobile Multicast (MoM) protocol: Multicast support for mobile hosts. In: ACM intl. Conf. on Mobile Computing and Networking(MOBICOM) Sep. 1997

10 Acharya A, Bakre A, Badrinath B R. IP Multicast Extensions for Mobile Internetworking: [Rutgers DCS TechReport]

11 Lin C R, Wang Kai-min. Mobile multicast support in IP networks. IEEE InfoCom 2000. 1664～1672

12 Wang Yu, Chen Weidong. Supporting IP Multicast for Mobiie Hosts. Mobile Networks and Appliations, 2001(6): 57～66

13 Choi M, et al. An efficient multicast routing protocol for mobile hosts. IEEE2001. 226～231

14 Lin C R, Chung C-J. Mobile Reliable Multicast Support in IP networks. IEEE2000. 1421～1425

15 Ke C-A, Liao Wanjiun. Reliable mobile multicast protocol: A Reliable Multicast protocol for Mobile IP networks. IEEE2000. 1488～1491

16 Liao Wanjiun, et al. Reliable multicast with host mobility. IEEE2000. 1692～1696

中国计算机学会全国办公自动化技术及应用学术会议

征文通知

中国计算机学会办公自动化专业委员会拟于2003年春在扬州召开办公自动化(OA)技术及应用学术会议，同时召开全体委员会议。现将有关事项通知如下：

一、征文范围 所有涉及OA技术和应用的动向、研究、开发和应用成果的论文，主要包括：OA现状及技术发展趋势；网络技术与分布式系统；OA中的信息安全技术；OA开发工具和OA语言；工作流/数据库、数据仓库和数据挖掘技术；OA中的多媒体技术；电子商务/电子政务；人工智能与决策分析在OA中的应用；其它。

二、来稿要求

1. 理论联系实际，具有学术和应用推广价值。未被其它会议、期刊录用或发表。

2. 来稿一般不得超过6000字(含图表)，为了便于正式出版论文集，来稿必须附中、英文摘要、关键词及主要参考文献，注明作者姓名、工作单位、详细通讯地址(包括电子邮件地址)与作者简介。

3. 欢迎电子投稿，来稿一般不退，请自留底稿。

三、来稿地址 南京东南大学计算机科学与工程系 胡苏宁 邮编：210096；
电话：(025)3793044；Email：yangming@seu.edu.cn

四、重要日期 征文截止日期：2002年10月15日
录用通知发出日期：2002年11月20日