基于遗传算法的K均值聚类分析*)

王 敞 陈增强 袁著祉
(南开大学信息技术科学学院 天津300071)

K-Means Clustering Based on Genetic Algorithm

WANG Chang CHEN Zeng-Qiang YUAN Zhu-Zhi
(College of Information Science and Technology, Nankai University, Tianjin, 300071)

Abstract This paper proposes a K-Means clustering method based on genetic algorithm. We compare our method with the traditional K-Means method and clustering method based on simple genetic algorithm. The comparison proves that our method achieves a better result than the other two. The drawback of this method is a comparably slower speed in clustering.

Keywords Data mining, Clustering, Genetic algorithm, K-Means clustering

1 前言

聚类分析就是将数据对象分组成为多个类或簇，在同一个簇中的对象之间具有较高的相似度，而不同的簇中的对象差别较大。聚类分析目前应用广泛，已经成为数据挖掘主要的研究领域。通过聚类，人们能够识别密集的和稀疏的区域，从而发现数据的整体分布模式，还能找到数据间的有趣的相互关系。关于聚类分析目前已经有K均值，CURE等很多算法，而且在实践中得到了应用。在这里，我们针对应用最为广泛的K均值方法的缺点，提出了基于遗传算法的K均值聚类分析方法。实验表明，新方法在聚类问题中得到的结果全面要优于传统K均值聚类方法，也好于单纯的遗传算法聚类。只是由于用到了遗传操作，聚类速度相对K均值方法要慢一些。

2 K均值方法的一般描述

K均值方法是基于划分的聚类方法。它在目前的聚类分析中应用最为广泛。其基本思想为：对于给定的聚类数目K，首先随机创建一个初始划分，然后采用迭代方法通过将聚类中心不断移动来尝试着改进划分。为了达到最优，这种K均值方法理论上应该穷举所有可能的划分。但实际上，这里采用了启发式方法，用每类的平均值来表示该类，这大大降低了计算的复杂性，提高了运算速度，使处理大规模数据集成为可能。但同时，这也导致了该方法受初始值影响很大，通常只能以局部最优结束。而且这种方法对于孤立点是敏感的。

3 基于单纯遗传算法的聚类

遗传算法是基于自然选择和遗传规律的搜索方法。该方法是随机选择与适者生存理论的结合。群体中的强者拥有更大的机会将其基因传给后代。我们可以简单地将一个实际问题的不同解编码成位串也就是所谓个体(Xi(t))，并评价它们的适应度(f(Xi(t)))，然后基于个体的适应度，按一定比例(Ps(Xi(t)))选择个体，进行遗传操作。遗传操作包括交叉和变异两步，这两步操作也是按一定概率(交叉概率Pc，变异概率Pm)进行的，在进行若干代遗传操作后，算法就很有希望找到最优解或近似最优解。由于遗传算法采用群体的方式进行搜索，这使得它可以同时搜索空间内的多个区域。在赋予遗传算法自组织，自适应，自学习能力的同时，优胜劣汰的自然选择和遗传操作使遗传算法具有不受其搜索空间限制性条件的约束和不需要其他辅助信息的特点。这些特点使遗传算法不仅能获得较高的效率而且具有简单，易于操作和通用的特性，这些特性使得遗传算法在各领域中应用得越发广泛。

已经有人尝试用遗传算法进行聚类，基本思想如下：将一染色体分成N个部分，每个部分对应一个数据元素的类别。例如，第10部分值为1表示10号数据元素属于类别1。在这里，适应度函数定义为数据间的欧式距离。初始群体采用随机产生的方法获得。遗传操作与传统遗传算法类似。根据遗传算法的理论基础，这种方法可以找到全局最优解，不受孤立点影响，不失为一个很好的方法。但是这种方法在数据量很大，要求聚类的类别很多时，运算时间就显得过长，而且往往效果还不如K均值方法。因此这种方法通常只适用于数量较小，类别数不大的情况。

也有人改进了遗传算法的编码方式，把各类别的聚类中心坐标编码成染色体，其他遗传操作与传统遗传算法一致。这种方法，收敛速度要好于第一种遗传算法，具有更强的适用性。不过这种方法由于单纯依靠遗传算法寻优，因此收敛依然缓慢，而且采用这种编码方式又导致了算法对孤立点变得敏感。

4 基于遗传算法的K均值聚类分析

我们的算法是在遗传算法思想与K均值算法思想的基础上提出来的。我们把K均值方法引入到遗传算法的进化中。首先，随机产生遗传算法的第一代并开始进化。在每代进化中，我们都用K均值方法对每个个体进行进一步的优化。这相当于在每一代都要对所有个体计算以其为初始值的K均值问题的局部最优结果，并以这些局部最优结果替换掉原来的个体并继续进化，直到达到最大代数或者结果符合要求为止。这种方法力图通过遗传算法来保证获取全局最优解，而用K均值方法提高算法的收敛速度。

*)本文得到国家自然科学基金(60174021)资助。

算法流程图如下：

(1)初始化：确定遗传参数；产生初始群体。

(2)对当前群体的每个个体，用K均值方法将其优化为以该个体为初始值的K均值问题的局部最优结果。

(3)对这些局部最优个体进行选择。

(4)动态改变交叉概率进行交叉。

(5)动态改变变异概率进行变异。

(6)若到达最大代数或者满足适应度要求，继续执行。否则，转②。

(7)输出结果。

染色体编码策略 我们将各个类别的中心坐标编码为染色体。例如对于一个类别为3的聚类问题，假设数据集为5维的，初始的三个聚类中心点为(10,20,30,40,50)，(20,40,60,80,100)，(30,50,70,90,110)，则我们的染色体编码就为(10,20,30,40,50,20,40,60,80,100,30,50,70,90,110)。这种基于聚类中心的编码方式减少了染色体的长度，大大提高了遗传算法的收敛速度，对于求解大量数据的复杂聚类问题效果较好。

适应度函数 我们将适应度函数定义成误差平方根值($\sqrt{\sum_{i=1}^{k}\sum_{p\in C_i}(P-Mi)^2}$)。其中，K为聚类的数目，P为问题空间的点，Mi是簇Ci的平均值，其中P，Mi都是多维的。这里，我们只考虑了数值属性的聚类，对于非数值属性，可以用适当方法转化为数值属性之后再使用该方法。

初始群体 初始群体有多种生成办法。为了获得全局最优解，这里我们的初始群体完全随机生成。考虑到要对大量数据进行聚类，我们不能将初始群体设得很大。但是为了使遗传算法数据元素能够多样性，我们在时间和硬件条件允许的条件下，应该尽可能地提高群体的规模。

选择，交叉与变异 为了让遗传算法能够获得全局最优解，我们采用了随机选择，最佳个体保留的方式。交叉则采用一点交叉方式。对本问题，我们经过仿真试验发现，变异是能够跳出局部最优的关键，变异算子对最终能否获得全局最优解影响重大。为此，我们对变异率实行动态变化。开始时，变异率较小，随着整代的平均适应度趋于最优个体的适应度时，变异率就自适应地增大。

孤立点的影响 由于本方法的进化目标是找到最优的聚类中心，因此与K均值方法的本质是一样的。少量的孤立点势必导致聚类中心的偏移，影响全局的聚类效果。但由于遗传算法的全局寻优特性，本算法对孤立点的敏感性要小于单纯的K均值法。

5 仿真实验

我们用VC++实现了新算法。实验数据采用在聚类中心附近产生的高斯分布数据。我们用文中所谈及的K均值方法，基于单纯遗传算法的聚类方法和新方法对于300，1000，5000，10000个有5维特征的点集进行了3类(K=3)聚类分析和10类(K=10)聚类分析。下面把一组用新方法和K均值方法进行比较的结果通过表格显示如下：

表1 3聚类分析(表格内容为误差平方根，进化：20代)

	K均值法	新方法	性能提升
300点	102.478	27.974	72.2%
1000点	189.617	50.075	73.6%
5000点	576.496	113.381	80.3%
10000点	812.978	159.91	80.3%

表2 10聚类分析(表格内容为误差平方根，进化：20代)

	K均值法	新方法	性能提升
300点	124.669	79.1883	36.5%
1000点	228.511	144.426	36.8%
5000点	529.764	235.96	55.5%
10000点	799.979	523.878	34.5%

结论 K-Means方法速度最快，可处理的数据集合最大。但是，所得到的结果通常只是局部最优，或者说结果还可以被进一步优化。在一些对时间要求较高，对精度要求不高的情况下，K均值方法还是最好的选择。单纯遗传算法对于类别数较小，数据量也较小的问题聚类效果最好，可以找到全局最优解。但是，一旦当数据量加大，类别加多的时候，该算法效果很差，对于复杂聚类问题基本不适用。对于数据挖掘中最常用到的类别数较大、数据量也很大的复杂聚类问题，新方法效果最好。相对于单纯K均值方法常常会有30%-50%的性能提高。新方法所用时间介于K均值方法与单纯遗传算法聚类之间。所能处理的数据规模也因为遗传操作的限制而不能过大(当然，这个问题可以通过随机抽取数据集中的特征数据再聚类获得解决)。这种方法的缺点是：由于在遗传的每代都进行了大量的K均值分析，因此运算速度相对较慢。

参考文献

1 Han Jiawei, Kamber M. Data Mining: Concepts and Techniques, The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor Morgan Kaufmann Publishers

2 徐勇，刘奕文，陈贺新，戴逸松. 一种基于自适应遗传算法的聚类分析方法. 系统工程与电子技术，1997，(9)

3 王磊，戚飞虎. 大矢量空间聚类的遗传K-均值算法. 上海交通大学学报，1999(9)

4 韩斌. 基于单亲遗传算法和模糊C-均值算法的混合聚类算法. 华东船舶工业学院学报，2000(3)

5 Maulik, Ujjwal Bandyopadhyay, Sanghamitra. Genetic algorithm-based clustering technique Pattern Recognition, Elsevier Science Ltd, 2000, 33(9)

6 Cowgill M C, Harvey R J, Watson L Z. Genetic algorithm approach to cluster analysis Computers & Mathematics with Applications, 1999, 37(7)