基于模拟退火的 K 调和均值聚类算法①

刘国丽，甄晓敏

(河北工业大学 计算机科学与软件学院，天津 300401)

摘 要：K 均值算法是最通用的划分聚类算法，然而它有高度依赖初始值和收敛于局部最小的缺点，K 调和均值算法采用数据点与所有聚类中心的距离的调和平均替代了数据点与聚类中心的最小距离，解决了 K 均值算法对初值敏感的问题。这样虽然解决初始值敏感问题，局部最小收敛问题仍然存在。为了获得全局最优解，提出一种新的算法：基于模拟退火算法的 K 调和均值聚类。该算法将一种优秀的随机搜索算法——模拟退火算法引入 K 调和均值聚类，来解决局部最小收敛的问题，并将改进后的算法用于 IRIS 数据集的聚类分析，聚类结果与 K 均值算法结果对比，证明了改进算法的优越性。

关键词：聚类；K 均值；调和均值；模拟退火；局部最小

K-Harmonic Means Clustering with Simulated Annealing

LIU Guo-Li, ZHEN Xiao-Min

(Department of Computer Science and Software, Hebei University of Technology, Tianjin 300401, China)

Abstract: K-means algorithm is a frequently-used methods of partition clustering. However, it greatly depends on the initial values and converges to local minimum. In K-harmonic means clustering, harmonic means fuction which apply distance from the data point to all clustering centers is used to solves the problem that clustering result is sensitive to the initial valve instead of the minimum distance. Although the problem above is solved, the problem converged to local minimum is still existed. In order to obtain a glonal optimal solution, in this paper, a new algorithm called K-harmonic means clustering algorithm with simulated annealing was proposed. This alhorithm is introduced into simulated annealing to solve the the problems of local minimum. Then the algorithm was used to analyse IRIS dataset and get a conclution that the new algorithm get a glonal optimal solution and reached a desired effect.

Key words: clustering; K-means; K-Harmonic means; simulated annealing; local minimum

1 引言

K 均值（KM）算法是一种简单高效的聚类算法，实际应用价值很高，因此很多学者对 K 均值算法的研究非常感兴趣，他们试图从不同的角度来改善 K 均值算法，使它能够有更好的效果。现在已经有了许多对 K 均值改进的方法，有基于密度蚂蚁思想的 K 均值算法,基于分层聚类的 K 均值算法，基于粗糙集的 K 均值聚类算法，基于遗传算法的 K 均值聚类；另外还有引入学习特征权值、最小生成树原理、三角不等式原理以及核学习方法来改进 K 均值的算法。本文研究并改进了一种由 K 调和均值（KHM）算法和模拟退火（SA）算法相结合而成的新算法，这种新算法称为基于模拟退火的 K 调和均值（SAKHM）聚类算法。

2 K调和均值聚类算法

2.1 K 均值算法

KM 算法是一种常用的基于划分的聚类算法,是最早提出的较为经典的聚类算法。算法以 k 为参数,把 n 个数据对象分为 k 个类,使类内具有较高的相似度,而类间有较低的相似度。这里相似度的计算采用欧几里德距离,类的中心则用类中对象的平均值来表示。KM 算法思想简单、易实现,而且收敛速度较快,具有可伸缩

① 收稿时间:2010-10-26;收到修改稿时间:2010-12-03

性和高效率性;但是也有两个缺点：算法结果依赖于初始值并且可能收敛到局部最优,而 K 调和均值聚类算法就解决了对初值敏感的问题。

2.2 K 均值算法

KHM 聚类方法由张[1]在 2000 年提出。它相对于 KM 算法,进行了以下几点改动：

1) KHM 算法中用调和平均函数取代了 KM 算法中的最小距离函数，调和平均公式为：

$$\frac{|C|}{\sum_{c\in C}\frac{1}{d^p(x,c)}} \tag{1}$$

其中，$x\in X$ 代表数据集中的对象，$c\in C$ 代表聚类中心，$d^p(x,c)$ 是距离测度。

2) KHM 算法目标函数为

$$\sum_{x\in X}\frac{|C|}{\sum_{c\in C}\frac{1}{d^p(x,c)}} \tag{2}$$

即为每个数据点到所有聚类中心的距离。

3) 中心迭代公式

$$c_k=\frac{\sum_{x\in X}\frac{1}{(\sum_{x\in C}\frac{d^p(x,c_k)}{d^p(x,c)})^p}x}{\sum_{x\in X}\frac{1}{(\sum_{x\in C}\frac{d^p(x,c_k)}{d^p(x,c)})^p}} \tag{3}$$

此处的距离测度采用明氏公式,在 KHM 算法里，p>2 时效果比较好；当 p=2 时，距离测度则为欧氏距离，是 KM 算法的距离测度公式。

4) 在 KM 中的每一个迭代里,目标函数给所有的数据点相等的权重,KHM 通过修正方程

$$w_{KHM}(x)=\frac{\sum_{c\in C}d^{-p-2}(x,c)}{(\sum_{c\in C}d^{-p}(x,c))^2} \tag{4}$$

给每个数据点分配基于调和均值的动态权重。

KHM 算法很好的减少了初始值对聚类结果的影响，使 KM 算法对初始值敏感问题得到很好的解决，但是聚类结果局部收敛的问题仍然存在，为了解决局部收敛问题，学者们将启发式规则引入算法，模拟退火就是比较优秀的启发式算法之一。它通过移动个别聚类中心产生扰动，并且按照一定概率接受新解，这种方法使 KHM 算法跳出局部最优，得到全局最优解，可取得较理想的聚类效果。

3 模拟退火算法

模拟退火（SA）算法是 1982 年由 Kirpatrick[2]提出的一种启发式随机优化算法。它的出发点是基于物理中固体物质的退火过程与一般组合优化问题的相似性。算法描述如下：

1）在初始温度下,随机产生初始解;

2）保持温度不变的情况下,随机产生遍历整个解空间的若干个候选解，;

3）按选定的概率函数在候选解中选出合适的解;

4）检验是否满足收敛条件,如满足条件则终止运算,否则,重复步骤 2)和 3),进入下一次迭代.

4 模拟退火K调和均值算法

4.1 算法理论

用模拟退火思想优化 K 调和均值算法，算法以优化过程的求解与物理退火过程的相似性为基础，通过对下降温度的控制使搜索过程向最优化方向进行，并依据 Metropolis 准则，以一定概率接受使目标函数变差的状态，通过具有概率突跳特性的随机搜索和温降重复抽样，最终得到问题的全局最优解，避免陷入局部极值[3]。该算法是一种具有并行性和渐进收敛性的全局优化算法。因此用模拟退火算法对 KHM 算法进行优化 ,可以改进 KHM 算法的局限性 ,提高原有算法的性能。算法具体设计：

1）参数初始化：最高温度 Tmax，最低温度 Tmin，内循环次数 MaxInnerLoop，退火系数 DR；

2）对样品进行 KHM 聚类 ,将聚类后样本用最小距离原则分配类别，此聚类划分结果赋给初始解 w，并计算相应的目标函数 F（1）。

3）将循环计数变量 InnerLoop 置 0；

4）数据样本归类，计算新聚类中心 w（k）。计算新状态下目标函数 F(k+1)，若 F(k+1)<F(k)，

则接受新状态；若 F(k+1)>=F(k)，求概率 $p=e^{\frac{F(k+1)-F(k)}{aT(k)}}$，T(k)为当前温度，a 为常数，产生一随机概率 r∈[0,1],若 P>=r，则接受新状态；否则，不接受新状态，以当前状态继续迭代；

5）若 InnerLoop<MaxInnerLoop，InnerLoop 加 1，转 4）；否则转 6）；

6）若 T(k)<Tmin，算法终止，否则，强制降温 T(k+1)=DR*T(k)，转 3）。

4.2 参数的选择及算法改进

基于模拟退火的 K 调和均值算法的研究重点在于

选择算法的结合方式和设置算法中的关键参数，下面四点讲述了算法中参数的设置，除此外，在 3）中改进了初始控制参数的获取方式，在 4）中讲述了算法的结合方式。这两点对算法进行了改进。

1）目标函数的选择

本文选择误差平方和函数作为本算法的目标函数。

2）温度更新方式

本文采用了由 Kirkpatrick 等人提出的设置退火系数[4]的方法。假设 DR 为退火系数，是一个接近 1 的常数，则温度更新公式为 T(k+1)= DR*T(k)，k 为降温次数。通过 DR 的值控制温度下降的快慢。本文将 DR 设为 0.98。

3）初始温度的选择

初始温度 Tmax 对算法全局搜索的能力有很大的影响，为了使算法进程在合理的时间内尽量搜索尽可能大的解空间，根据平衡的理论，控制参数 T 的初始值应该选得足够大，设初始接受率为 v_0，则应满足 $v_0=\frac{\text{接受新解数}}{\text{提出新解数}}\approx 1$，由 Metropolis 接收准则可推知：$\exp(\frac{-\Delta F}{T\max})\approx 1$，要想使上式成立，Tmax 的值应该足够大，但是如果 Tmax 的值过大的话就会增加迭代的次数，增加计算时间。Tmax 最好选择可以保证算法取得全局最优解的最小数。Kirkpatrick 等人提出一种选择初始温度的方法，叫做经验法。它首先选定一个大值作为 Tmax 并进行若干次变换，如果接受率 v 小于预定的初始接受率 v_0（一般取 0.8），则 Tmax 值加倍，直至得到使 $v<v_0$ 的 Tmax 值；本文 Tmax 的选择使用将上述的经验法与 KHM 的最终目标函数相结合的方法，并将经验法进行了扩充。方法如下：首先将 KHM 的目标函数作为 Tmax 值，再按照上述的方法进行若干次变换，如果接受率 $v<v_0$（v_0=0.8），则当前 Tmax 值加倍，直至得到使 $v>v_0$ 的 Tmax 值；如果接受率 $v>v_0$（v_0=0.8），则当前 Tmax 值减半，直至得到使 $v<v_0$ 的 Tmax 值，此时 v 取 $v>v_0$ 成立的最小 Tmax 值。

4）初始解及新解的产生

为了使算法在开始时就达到准平衡。先对数据集进行 KHM 聚类，将聚类结果作为初始解。由于在接下来的模拟退火迭代过程中需要不断进行迭代，而 KHM 算法本身的计算量就很大，如果继续将 KHM 聚类的中心更新法和目标值计算用于退火迭代，计算量会相当大。为减少算法的时间复杂度，可以将 KM 算法的中心更新方法和目标函数为用于迭代过程这样，可以显著降低时间复杂度，也可以得到很好的结果。

基于模拟退火的 K 均值算法中，模拟迭代过程中新解的产生是对当前解进行随机扰动得到的，即随机变化一个或几个聚类样品的当前所属类别,产生一种新的聚类划分,从而使算法有可能跳出局部极小值。但是初始解—KHM 聚类的结果中，各个数据并没有明确唯一的所属类别，而从扰动过程来看，必须明确所属类别，为此，可以在最初 KHM 聚类后，用最小距离原则为数据分配类别，作为聚类后的数据所属的类别。

综上所述，我们将数据进行 KHM 聚类后，以聚类结果为簇中心，用最小距离原则为数据分配类别，计算相应目标函数，以此作为初始解。并以随机扰动方法产生新解。

5　实验研究

为测试 SAKHM 聚类算法的聚类性能，在常用的 IRIS（鸢尾花）数据集上进行了实验. IRIS 数据集由 150 个 4 维数据组成。数据集分为三个类，每个类代表一种类型的 iris 植物，各包含 50 个实例，每个实例带有 4 个数值属性,分别代表：花萼长（cm）,花萼宽（cm）,花瓣长（cm）,花瓣宽（cm），没有缺失的属性值。三个类中，其中一个类可与其它两个可直线分割;另外两个相互间不可用直线分割。

聚类分析前要对数据集进行标准化，本文的标准化公式为 $X'=\frac{X-mean}{std.Deviation}$。标准化后的变量描述如表 1：

表 1　IRIS 数据集标准化后变量描述

变量名称	数据个数	最小值	最大值	平均值	标准差
花萼长	150	-1.86	2.48	0.00	1.00
花萼宽	150	-2.47	2.48	0.00	1.00
花瓣长	150	-1.56	1.78	0.00	1.00
花瓣宽	150	-1.44	1.71	0.00	1.00

用三种算法：KM,SAKM,SAKHM 在数据集上各运行 10 次。其中 KM 算法要输入参数 k=3；SAKM 和 SAKHM 还要输入 Tmax=36, Tmin=0.0001, MaxInnerLoop=5,此外，SAKHM 还有参数 p=3.5。

图 1、图 2 分别是 SAKM 算法和 SAKHM 算法第 5 次运行中目标函数值与参数 t 的对应图:

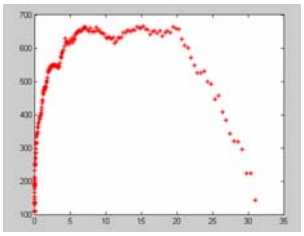

图 1 SAKM 算法中参数 t 与目标函数 J 的对应图示

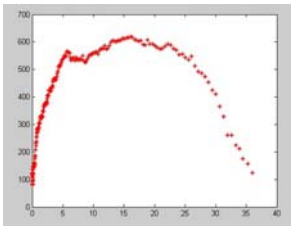

图 2 SAKHM 算法中参数 t 与目标函数 J 的对应图示

我们从几个方面来分析上图：

1） 两图起始点和终止点描述

图 1 的右下角和左下角起始点分别是 SAKM 算法的起始解和最终优化解对应的目标函数值。同理，图 2 的右下角和左下角起始点分别是 SAKHM 算法的起始解和最终优化解所对应的目标函数值；经对比，两算法最终的目标函数值都小于起始解对应的目标函数值，可知算法对目标函数是有优化作用的。

2）对比两图的目标函数值

SAKHM 算法的初始解目标函数小于 SAKM 的初始解目标函数，说明了调和均值函数对得到较小目标函数所起的作用。两图对比最终解对应的目标函数看出，SAKM 算法和 SAKHM 算法的目标函数值分别为 108.6364 和 79.7981，SAKHM 算法的目标函数要小很多，说明它的优化作用更好。

3）目标函数的变化过程分析

由于模拟退火算法的搜索过程是随机的，且当 t 值较大时可接受部分恶化解，所以当前解到达到最优解时必须经过暂时恶化的“山脊”。但最终随着温度 t 的降低，恶化解被接受的概逐渐减小直至趋近于零，只有优化的解才会被接受，目标函数在此时迅速下降，在接近 0 处取得了最优值。

通过对图 1、图 2 多个角度的对比，一方面说明了模拟退火技术的优化能力，另一方面证明了 KM 算法引入调和均值函数的作用。最重要的是证明了 K 均值算法、调和平均函数、模拟退火技术结合得很成功，算法效果非常好。

另外，将一个三种算法各 10 次运行的一些重要数据进行了记录汇总，包括目标函数的最大值、最小值和平均值，与实际中心误差和平均 CPU 时间，其中后两者为 10 次运行的平均值。

表 2 三种算法运行结果汇总

方法	最小值	平均值	最大值	中心误差	CPU 时间
KM	140.94	175.56	207.06	18.98	0.05
SAKM	106.73	124.74	179.65	9.13	12.83
SAKHM	76.32	79.82	84.98	2.63	16.94

由表中数据可知,虽然 KM 算法运行时间最少，但对初始值高度敏感，因为它的性能函数范围最大。相反，SAKM 算法的聚类中心与实际中心误差减小,但时间明显增加，并且结果不是最好的。虽然 SAKHMC 算法需要大量计算，但是得到的结果是最好的。

好的算法的主要标准有两个：

1）聚类中心到实际聚类中心的误差最小；

2)性能函数值较小。综上分析,通过与 KM、SAKM 两种算法运行结果的比较分析，我们可以知道，虽然 SAKHM 算法在时间上相对复杂一些，但是无论是从目标函数方面还是从与实际中心的接近程度上，它都优于 KM 和 SAKM 算法，得到了很好的聚类效果。

6 结论

本次研究提出了一个基于模拟退火的 K 调和均值聚类算法。用该算法在 IRIS 数据集上进行了测试，实验表明,虽然本算法时间上相对复杂，但无论是从性能函数方面还是从与实际中心的接近程度上，它都（SAKHM)优于 KM 和 SAKM 算法。

该方向未来的研究有：基于其他启发式搜索方法的 KHM 数据聚类,例如基于蚁群的优化，基于遗传算法的优化,还有基于其他的优化。也可做基于现实生活领域的研究，以展示我们的算法能够有效应用于那些领域。

参考文献

1 赵恒,杨万海.一种基于调和均值的模糊聚类算法.电路与系统学报,2004,9(5):114–117.

2 吴晓燕等.基于遗传模拟退火算法的高维离群点挖掘.微计算机信息,2010,7(3):139–140.

3 Zülal Güngör, Alper Ünler. K-harmonic means data clustering with simulated annealing heuristic. Applied Mathematics and Copputatuin,2007,32(6):199.

4 谢磊,张旭毅,郑仕勇.模拟退火 K 均值算法在文本分类中的应用.软件导刊,2010,9(6):41–42.