K 均值聚类算法的研究与优化

陶 莹,杨 锋,刘 洋,戴 兵
(广西大学 计算机与电子信息学院,广西 南宁 530004)

摘 要:聚类分析是数据挖掘的重要组成部分,K 均值聚类算法是聚类分析方法中一种基本的划分式方法,也是无监督的机器学习方法。其具有效率高、容易理解和实现等优点,同时,可以对多种数据类型进行聚类,广泛应用于诸多领域。但是,K 均值聚类算法也有一些局限性。算法中合理的 k 值难以确定,而且算法选择初始聚类中心的随机性会导致聚类结果不稳定,同时,算法对噪声和离群点数据也有很强的敏感性。为了解决初始聚类中心随机性的问题,通过全局化思想对 K 均值聚类算法进行了改进,改进的聚类效果评价使用常用的误差平方和准则。实验结果表明,相较于一般的 K 均值聚类算法,全局 K 均值聚类算法得到了更好的聚类效果,同时提升了算法的稳定性。

关键词:数据挖掘;K 均值聚类;中心点;误差平方和
中图分类号:TP393 **文献标识码**:A **文章编号**:1673-629X(2018)06-0090-03
doi:10.3969/j.issn.1673-629X.2018.06.020

Research and Optimization of K-means Clustering Algorithm

TAO Ying, YANG Feng, LIU Yang, DAI Bing
(School of Computer and Electronic Information, Guangxi University, Nanning 530004, China)

Abstract: Clustering analysis is an important part of data mining. The K-means clustering algorithm is a basic partition method of clustering analysis, and it is also an unsupervised machine learning method with the advantages of high efficiency, easy understanding and implementing. At the same time, the clustering data type can be various, so it is widely used in many fields. However, the K-means clustering algorithm exists some limitations. For example, the reasonable value of k is difficult to determine, and choosing the initial clustering center is random, which can lead to the result unstable, also with strong sensitivity to noise and outliers. In order to solve the problem of the randomness for initial clustering center, we improve the K-means clustering algorithm through the idea of global change. The evaluation criterion of the clustering effect is the error sum of squares. Experiment shows that compared with normal K-means clustering algorithm, the global K-means clustering algorithm can get better clustering effect, while increasing its stability.

Key words: data mining; K-means; center point; error sum of squares

0 引 言

数据挖掘在实际应用中的主要任务之一是聚类分析,其是数据挖掘中一个很热门的研究领域,同时与其他学科的研究方向有很大的交叉性[1]。聚类分析可以发现数据隐含的结构,对数据进行自动归类,从而得到数据的大致分布,在诸多领域为决策提供支持信息[2]。

K 均值聚类算法是聚类分析方法中一种基本的划分式方法[3]。由于该算法简便易懂,且计算速度较快,通常被作为大样本聚类分析的首选算法[4]。

但是一般的 K 均值聚类算法初始聚类中心的选择是随机的,这样会导致聚类结果的不稳定。而且算法中 k 的值需要人为提前设定,k 值设定的合理与否会直接影响聚类的效果[5]。此外,噪声和离群点也会对聚类效果产生影响,使得聚类中心偏离主要数据区域[6]。因此,文中针对 K 均值聚类算法的随机性较强的特点进行改进。

1 K 均值聚类算法

1.1 算法基本思想

首先在整个数据集中任意选择 k 个数据作为初始聚类中心,然后计算其他数据对象与 k 个聚类中心的距离,将数据对象划分到距离最近的聚类中心所在的聚类域。所有数据划分好后,重新计算 k 个聚类中每个聚类的全部数据对象的平均值,该平均值所在的数

收稿日期:2017-06-04 **修回日期**:2017-10-12 **网络出版时间**:2018-02-24
基金项目:广西壮族自治区中青年教师基础能力提升项目(KY2016YB026);广西自然科学基金(2014GXNSFBA118274)
作者简介:陶 莹(1993-),女,硕士研究生,研究方向为计算机视觉;杨 锋,博士,副教授,硕导,研究方向为网络信息安全和人工智能。

据点成为新的聚类中心。最后进行多次迭代,直到连续两次的聚类中心相同,说明此时数据对象类别划分完毕,即得到 k 个聚类[7]。

1.2　算法评价准则—误差平方和

评价聚类效果,可以定义一个数据对象和其所在聚类域的目标函数。通过目标函数的取值情况评价聚类效果[8]。常用的聚类算法评价准则是中心误差的平方和,即:

$$J = \sum_{j=1}^{k} \sum_{u \in c_j} d(\boldsymbol{X}_u, \boldsymbol{m}_j) \tag{1}$$

其中,$\boldsymbol{X}_u$ 是数据 u 的全部属性值所构成的矢量;k 是聚类个数;$m_1, m_2, \cdots, m_k$ 是 k 个聚类中心对应的矢量;c_j 是聚类中心为 $\boldsymbol{m}_j$ 的聚类域。

聚类中心矢量 $\boldsymbol{m}_j$ 表示为:

$$\boldsymbol{m}_j = \frac{1}{N_j} \sum_{u \in c_j} \boldsymbol{X}_u \tag{2}$$

其中,N_j 为聚类域 c_j 中数据的个数。

式 1 中,目标函数 J 代表 k 个聚类里的全部数据与其聚类中心 $\boldsymbol{m}_j$ 之间的误差平方和,值越小表明聚类中数据的集中性越好,即得到的聚类效果越好[9]。

1.3　算法局限性分析

(1)K 均值聚类算法中的 k 值(待聚类簇的个数)必须由用户输入。

k 值必须是用户最先确定,即分为多少个聚类。但是在一些实际情况下,合适的聚类数目 k 用户也是未知的,在这种情况下,就需要运用其他办法来获得到聚类的数目[10]。

(2) k 个聚类中心的选择是随机的。

一般 K 均值聚类算法初始中心是随机选择的,然后进行聚类和迭代,并最终收敛达到局部最优结果[11]。因此,聚类结果对于初始中心有着严重的依赖,随机选择初始中心会造成聚类结果有很大的随机性。

(3)K 均值聚类算法对于噪声和离群点数据非常敏感。

K 均值聚类算法中每个聚类的中心都由每个聚类里所有数据求均值得到。当有与其他数据不一致的数据或者差异比较大的数据时,计算出的聚类的中心易受干扰,偏离主要数据区域,影响聚类效果。因此,K 均值聚类算法对数据中的噪声和离群点非常敏感[12-13]。

2　全局 K 均值聚类算法

2.1　算法基本思想

全局 K 均值聚类算法从 $k = 1$ 的聚类开始,即先求出所有数据的聚类中心 m。当 $k = 2$ 时,将聚类中心 m 作为其中一个初始聚类中心,然后依次将数据集的每个点作为另一个聚类的初始中心,即运行 n 次 K 均值聚类算法,计算误差平方和 J 后,取值最小的点确定为第二个聚类中心,其中聚类后得到了新的聚类中心 m_1, m_2。如果 $k = 3, 4, \cdots, n$,以此类推[14]。

全局 K 均值聚类算法最终得到了所有 k ($k < K$) 的聚类解决方案,不存在初始化的问题并且相对稳定[15]。

2.2　实验数据分析

当 $k = 2$ 时,对 40 个数据重复使用 K 均值聚类算法得到 4 种不同的结果,如表 1 和图 1 ~ 4 所示。根据误差平方和评价准则,最好的实验结果是实验一,其误差平方和是 112.923 864 62,但 K 均值聚类算法本身的不稳定性使得每次产生不一样的聚类结果,不一定是最优的,甚至是很差的聚类结果。

表 1　K 均值聚类算法 4 次实验结果

指标	1	2	3	4
J	112.923 864 62	118.700 924 84	117.223 879 84	116.528 861 12

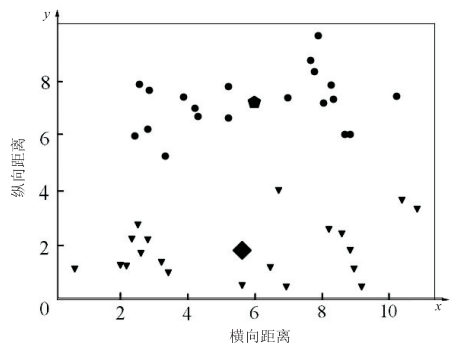

图 1　K 均值聚类结果一

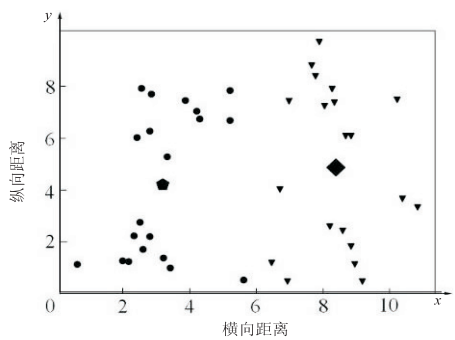

图 2　K 均值聚类结果二

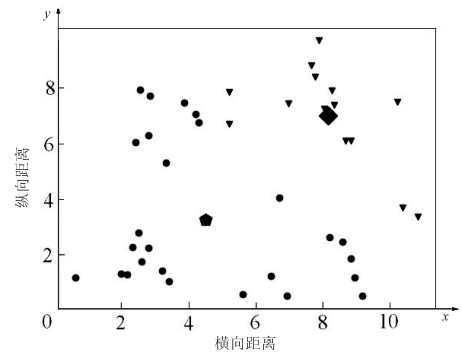

图 3　K 均值聚类结 5 三

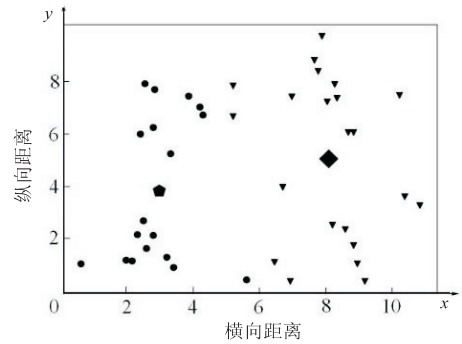

图 4 K 均值聚类结果四

下面使用改进的 K 均值聚类算法即全局 K 均值聚类算法对数据进行实验。当 $k=1$ 时,得到聚类中心 $m=[-0.207\ 404\ 58, 0.055\ 375\ 1]$,再用 m 分别和每一数据点为中心进行 K 均值聚类算法。实验数据见表 2。

表 2 全局 K 均值聚类算法实验数据

数据点	J
1-5	113.124 482 112.923 864 118.700 924 112.923 113 112.923 864
6-10	118.700 924 112.923 864 112.923 113 117.223 879 112.923 864
11-15	123.975 412 113.083 026 117.223 879 112.923 864 118.700 924
16-20	113.083 026 117.223 879 112.923 864 123.975 412 113.083 026
21-25	117.223 879 112.923 864 118.700 924 113.083 026 113.124 482
26-30	112.923 864 118.700 924 113.083 026 113.124 482 118.700 924
31-35	112.923 864 113.083 026 117.223 879 118.700 924 118.700 924
36-40	113.083 026 113.124 482 112.923 864 123.975 412 113.083 026

全局 K 均值聚类的结果如图 5 所示。

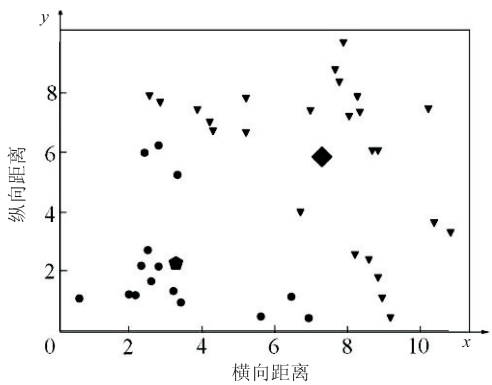

图 5 全局 K 均值算法聚类结果

与 K 均值聚类算法相比,全局 K 均值聚类算法的误差平方和 $J=112.923\ 113$,改善了 K 均值聚类算法的随机性所导致不理想结果的可能性。全局 K 均值聚类算法不受初始聚类中心位置的影响,并且通过一种确定有效的方法能够最小化误差平方和。

3 结束语

聚类算法发展到今天,已经衍生出了多种算法。其中,经典的 K 均值算法作为划分聚类算法中最基础的算法,由于其高效性和简单性被广泛应用于各领域[16]。然而,K 均值算法也有其固有的局限性,而很多针对 K 均值算法的改进都极大地降低了算法本身的性能,这显然是得不偿失的[17]。对此,文中对 K 均值聚类算法进行了改进,降低了算法的不稳定性,提高了聚类的有效性。

参考文献:

[1] 周 涛,陆惠玲. 数据挖掘中聚类算法研究进展[J]. 计算机工程与应用,2012,48(12):100-111.

[2] 盘俊良,石跃祥,李娉婷. 一种新的粒子群优化聚类方法[J]. 计算机工程与应用,2012,48(8):179-181.

[3] ABDEYAZDAN M. Data clustering based on hybrid K-harmonic means and modifier imperialist competitive algorithm[J]. Journal of Supercomputing,2014,68(2):574-598.

[4] HUNG C H,CHIOU H M,YANG Weining,et al. Candidate groups search for K-harmonic means data clustering[J]. Applied Mathematical Modelling,2013,37(24):10123-10128.

[5] 丁祥武,郭 涛,王 梅,等. 一种大规模分类数据聚类算法及其并行实现[J]. 计算机研究与发展,2016,53(5):1063-1071.

[6] 万 静,张 义,何云斌,等. 基于 KD-树和 K-means 动态聚类方法研究[J]. 计算机应用研究,2015,32(12):3590-3595.

[7] 罗军锋,锁志海. 一种基于密度的 k-means 聚类算法[J]. 微电子学与计算机,2014,31(10):28-31.

[8] 王 涛,卿 鹏,魏 迪,等. 基于聚类分析的进程拓扑映射优化[J]. 计算机学报,2015,38(5):1044-1055.

[9] SAHOO A K,ZUO M J,TIWARI M K,et al. A data clustering algorithm for stratified data partitioning in artificial neural network[J]. Expert Systems with Applications,2012,39(8):7004-7014.

[10] 贾洪杰,丁世飞,史忠植. 求解大规模谱聚类的近似加权核 k-means 算法[J]. 软件学报,2015,26(11):2836-2846.

[11] 朱建宇. K 均值算法研究及其应用[D]. 大连:大连理工大学,2013.

[12] GÜNGÖR E,ÖZMEN A. Distance and density based clustering algorithm using Gaussian kernel[J]. Expert Systems with Applications,2017,69:10-20.

[13] 赵 丽. 全局 K-均值聚类算法研究与改进[D]. 西安:西安电子科技大学,2013.

[14] 雷小锋,谢昆青,林 帆,等. 一种基于 K-Means 局部最优性的高效聚类算法[J]. 软件学报,2008,19(7):1683-1692.

[15] 张建萍. 基于计算智能技术的聚类分析研究与应用[D]. 济南:山东师范大学,2014.

[16] AHMAD A,HASHMI S. K-harmonic means type clustering algorithm for mixed datasets[J]. Applied Soft Computing,2016,48:39-49.

[17] TU Chunhao,JIAO Shuo,KOH W Y,et al. Comparison of clustering algorithms on generalized propensity score in observational studies:a simulation study[J]. Journal of Statistical Computation and Simulation,2013,83(12):2206-2218.