

基于最小生成树的层次 K-means 聚类算法

贾瑞玉, 李 振

(安徽大学 计算机科学与技术学院, 安徽 合肥 230601)

摘 要: 针对 K-means 算法初始化时需要指定聚类数目, 和随机选择初始聚类中心对聚类结果产生不稳定的问题, 结合图论中最小生成树和层次算法的分裂、凝聚思想, 提出一种基于最小生成树的层次 K-means 算法. 该算法初始时根据数据样本生成一颗最小生成树, 然后利用层次分裂思想把数据分成多个较小的簇, 通过 K-means 算法迭代操作得到每次操作的评价函数值来判断是否进行簇的合并, 进一步确定聚类簇数目. 实验结果证明, 该算法能够较准确地判断聚类数目, 并且聚类结果的稳定性比基本 K-means 算法要好.

关键词: K-means 算法; 聚类簇数; 初始聚类中心; 层次结构; 最小生成树; Prim 算法

中图分类号: TP301.6

文献标识码: A

文章编号: 1000-7180(2016)03-0086-03

The Level of K-means Clustering Algorithm Based on the Minimum Spanning Tree

JIA Rui-yu, LI Zhen

(School of Computer Science and Technology, Anhui University, Hefei 230601, China)

Abstract: Aiming at the number of clusters and the problem of the clustering results instability result of random selection of the initial cluster centers. Combine minimum spanning tree (MST) with the thinking of split and cohesion of hierarchical algorithm, proposed a MST-based hierarchical K-means algorithm. A MST is generated by sample data at the time of the algorithm initialization, and then use the thought of disunion, divide the data into smaller cluster. Gain the evaluation function value of each operation by iterative operation of the K-means algorithm, determine whether to cluster merging with value of evaluation function, to further determine the clustering number of cluster. Experimental results show that the algorithm can more accurately determine the number of clusters, and the stability of the clustering results of the algorithm is better than the basic K-means.

Key words: K-means algorithm; clustering number of clusters; the initial cluster centers; hierarchy; minimum spanning tree (MST); Prim algorithm

1 引言

数据聚类是数据挖掘中一项重要的研究内容, 在图像分割、模式分类、机器学习等领域有着广泛的应用^[1]. 数据聚类是根据数据样本的内在属性把相似性较大的数据聚在一起, 形成一个相似数据的簇. 聚类的结果往往是同一类别中的数据具有较大的相似性, 而不同类别之间差异性较大. 现阶段常用的聚类方法主要分为基于划分的聚类方法, 基于层次的聚类方

法、基于密度的聚类方法和基于网格的聚类方法等^[2].

层次聚类是一种应用广泛的非监督聚类方法. 分为两种类型: 自底向上的可凝聚层次聚类和自上而下的可分裂层次聚类. K-means 是一种经典的基于划分的聚类算法, 该算法以聚类中心为基础不断迭代, 优化聚类结果^[3]. 但使用该算法必须在聚类之前就指定要聚类的个数 K 和初始的聚类中心, 聚类结果往往受到随机指定聚类中心的影响. 解决初始聚类中心随机

选择对聚类结果的影响和确定聚类数目 K 值一直是该算法改进的方向. 胡伟通过把簇不断分裂, 根据终止条件最终得到较好的聚类数目^[4]. 徐沁为解决 K-means 陷入局部最小值问题, 提出一种结合自适应 means-shift 与最小生成树的 K-means 聚类算法^[5]. 在文献[6]中从图的角度, 用 Prim 算法生成数据的最小生成树, 利用 Prim 轨迹确定聚类的类别和中心点. 李翔宇在数据集上构建网格最小生成树, 由此确定潜在的簇结构, 优化聚类结果^[7].

本文融合层次聚类的分裂和凝聚的思想, 提出一种基于最小生成树的层次 K-means 聚类算法.

2 K-means 聚类算法

K-means 算法是一种基于划分的聚类算法, 初始化时给定聚类簇的数目和聚类中心, 不断迭代更新聚类中心, 以达到最优解. 解的评价常使用目标函数 J :

$$J = \sum_{j=1}^k \sum_{i=1}^m \text{dist}(p_i, c_j) \quad (1)$$

式中, m 代表属性数目, p_i 代表属于聚类中心 c_j 的数据对象, dist 是基于欧式距离, 定义如下:

$$\text{dist}_{ij} = \left(\sum_{k=1}^p |x_{ik} - x_{jk}|^2 \right)^{\frac{1}{2}} \quad (2)$$

式中, p 表示数据的属性个数.

K-means 算法流程具体描述如下:

- (1) 给定要聚类簇数 K ;
- (2) 在数据样本点中随机选取 K 个数据作为初始聚类中心;
- (3) 根据距离度量(本文采用基于欧式距离), 将数据对象分配给距离最小的聚类中心;
- (4) 重新计算聚类中心;
- (5) 采用式(1)计算目标函数 J 的值;
- (6) 重复步骤(3) ~ (5) 直到达到迭代次数或者簇 J 值不再变化.

在使用 K-means 聚类算法时必须在聚类开始时给出 k 个簇和初始聚类中心点, 其中 k 值较难选取. 并且初始的聚类中心确定一个初始的划分, 其后的步骤是对初始的划分进行优化. 初始聚类中心的选择对聚类结果的影响较大, 一旦初始值选择不好, 可能无法得到有效的聚类结果. 同时 K-means 算法也有其一定的优势. 其算法思想简单, 实现容易; 另外算法运算时间复杂度低, 收敛速度较快.

3 Prim 算法

普里姆算法(Prim 算法), 在加权连通图里搜索

最小生成树的一种算法, 该算法能够得到权值最小的连通图. 详细步骤如下:

- (1) 输入: 一个加权连通图, 其中顶点集合为 V , 边集合为 E ;
- (2) 初始化: $V_{\text{new}} = \{x\}$, 其中 x 为集合 V 中的任意一节点(起始点), $E_{\text{new}} = \{\}$, 为空;
- (3) 重复下列操作, 直到 $V_{\text{new}} = V$:
 - ① 在集合 E 中选取权值最小的边 $\langle u, v \rangle$, 其中 u 为集合 V_{new} 中的元素, 而 v 不在 V_{new} 集合当中, 并且 $v \in V$ (如果存在有多条满足前述条件即具有相同权值的边, 则可任意选取其中之一);
 - ② 将 v 加入集合 V_{new} 中, 将 $\langle u, v \rangle$ 边加入集合 E_{new} 中;
- (4) 输出: 使用集合 V_{new} 和 E_{new} 来描述所得到的最小生成树.

4 基于最小生成树的层次 K-means 聚类算法

基本的 K-means 聚类算法在进行聚类初始化时候需要给定确切的聚类数目 K , 和初始聚类中心. 许多情况下数据要聚成几簇是不能提前确定的, 随机给出初始聚类中心会使聚类结果产生较大的波动, 聚类结果不稳定. 针对以上问题, 本文提出了一种基于最小生成树的层次 K-means 聚类算法. 该方法首先对给定的数据样本集进行处理产生一个最小生成树, 然后给出模糊的分裂数 $m-1$, 根据最大距离分裂把树分成 m 棵子树, 这里把每棵树都看做独立的一个簇, 从中选择含有元素较多的簇作为初始的 K 个簇, 并计算其初始聚类中心. 使用 K-means 算法进行迭代操作, 通过评价函数

$$I = \sqrt{\frac{J}{n}} \quad (3)$$

判断是否进行簇之间的合并操作, 最终得到较好的聚类数目和稳定的聚类结果. 式中, J 为式(1) 的值, n 为聚到该簇的数目.

算法步骤如下:

- (1) 对于给定 n 个数据对象, 采用经典的 Prim 算法生成一棵最小生成树.
- (2) 设置模糊的分裂数目 $m-1$, 根据最大距离分裂原则把步骤(1) 的最小生成树分成 m 部分, 把每一部分看做一个簇.
- (3) 从 m 个簇中选择含有较多元素的簇 K 个, 看做初始的簇别, 并计算其对应的簇中心 $c_j (j = 1, 2, \dots, k)$.
- (4) 进行 K-means 算法迭代处理.

(5) 选择聚类中心中最小距离的两个簇进行合并,重新计算聚类中心.

(6) 对评价函数(3) 进行评价分析,公式 $\Delta = \frac{I^{(t)} - I^{(t-1)}}{C^{(t)} - C^{(t-1)}}$,如果 $\Delta > \omega$ 则返回步骤(4) 继续执行,否则执行结束. 其中 C 表示聚成簇数目, $\omega = -0.19$ 是经验值.

5 实验结果及分析

本文采用三个经典的 UCI 数据集,如表 1 所示,并给出与 K-means 算法比较结果. 通过 100 次的实验记录 J 的最优、最差和平均值. 实验运行环境采用 inter core2,2.20 GHz CPU,Java 编程.

表 1 实验数据集

数据集	数据个数	属性值
Iris	150	4
Wine	178	13
Glass	214	9

图 1 中 Iris 数据的变化如下:在聚成 3 簇之前的评价价值变化比较平稳,而在过了 3 簇之后变化则是剧烈的,因此选择聚成 3 簇是比较合适的. 对于数据 Wine,把数据聚成 3 类之前,评价价值逐渐增加,因为聚类数目的减小,评价价值也会随之增加,当聚成 3 簇之后,评价价值增加迅速,是变化较大的转折点,可以把 Wine 聚成 3 类. 数据 Glass 在簇数 2 到 5,6 到 10 之间,评价价值变化都很平稳,5 和 6 是转折点,考虑到 J 值越小聚类也就最紧密,所以选择簇为 6. 实际中数据集 Iris、Wine、Glass 的簇数目分别为 3,3,6. 该算法能较好地解决传统 K-means 算法聚类初始化时必须确定聚类簇数目.

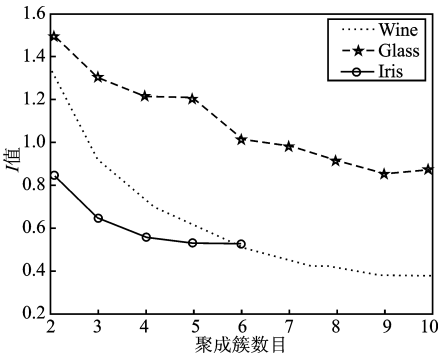


图 1 聚成不同数目簇变化曲线

(注:由于进行合并操作,所以在图 1 中随着聚类数目的减小来进行观察. 为了能在同一图中反映所有数据,我们对 J 值进行了了除 10 处理)

从表 2 可以看出在对数据集 Iris、Wine、Glass 进行聚类时,基于最小生成树的层次 K-means 算法比传统的 K-means 算法在均值一栏更小,反映出聚类结果的变化幅度较小. 在图 2 ~ 图 4 中可以更加直观地看到基于最小生成树的层次 K-means 算法在数据集 Iris、Wine、Glass 上的变化曲线波动较小,而基本 K-means 算法曲线幅动较大. 因此,改进的层次 K-means 算法更好地解决了随机选择聚类中心对聚类结果波动性较大的影响.

表 2 100 次试验 J 值

算法数据集	Iris	Wine	Glass
K-means 算法	最大值 149.52	18 436.95	259.7
	最小值 97.32	16 530.58	213.22
	平均值 111.576	17 343.54	236.86
改进层次 K-means	最大值 97.34	16 555.68	220.34
	最小值 97.19	16 480.97	213.69
	平均值 97.227	16 541.99	217.97

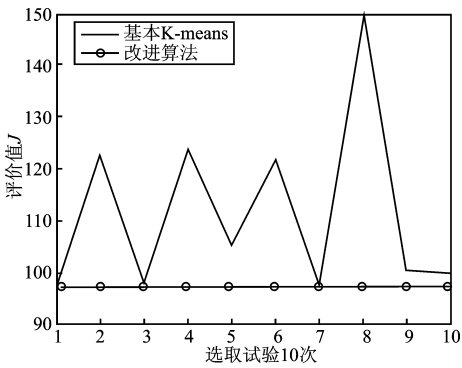


图 2 Iris 数据集

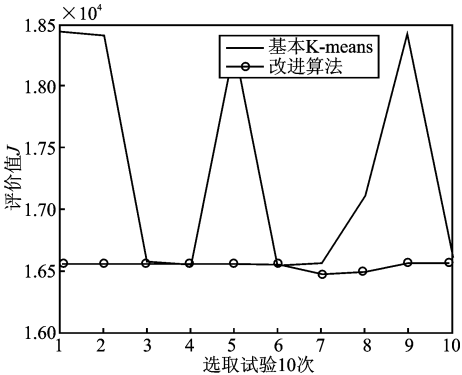


图 3 Wine 数据集

6 结束语

本文融合层次算法思想,对基本 K-means 算法 (下转第 93 页)

5 结束语

本文通过测试函数的测试和比较,验证了混沌人工鱼群算法(CAFSA)的准确性和有效性.同时运用该方法可以实现对 RBF 核函数的 SVM 参数组合(C, γ)的优化,并适用于语音识别系统.实验结果表明混沌人工鱼群算法比基本人工鱼群算法更易跳出局部极小,收敛速度更快,系统识别率更高.但这种方法仍然存在一些不足之处,在某些条件下(如 0 dB40 词、5 dB50 词和 10 dB40 词)识别率并没有显著的提高,且总体最佳参数的寻优时间较长,所以该问题有待进一步地研究和解决.

参考文献:

- [1] 邓乃扬,田英杰.数据挖掘中的新方法:支持向量机[M].北京:科学出版社,2004.
- [2] 袁小芳,王耀南.基于混沌优化算法的支持向量机参数选取方法[J].控制与决策,2006,21(1):111-113.
- [3] 李晓磊,邵之江,钱积新.一种基于动物自治体的寻优

模式:鱼群算法[J].系统工程理论与实践,2002,22(11):32-38.

- [4] 高雷阜,赵世杰,高晶.人工鱼群算法在 SVM 参数优化选择中的应用[J].计算机工程与应用,2013,49(23):86-90.
- [5] Jiang Mingyan, Cheng Yongming. Simulated annealing artificial fish swarm algorithm [C] // Proceedings of the 8th World Congress on Intelligent Control and Automation. Jinan, China:IEEE, 2010: 1590-1593.
- [6] 李晓磊,薛云灿,路飞,等.基于人工鱼群算法的参数估计方法[J].山东大学学报,2004,34(3):84-87.
- [7] 杨利红.基于 AFSA 的 SVM 参数优化及其应用[D].太原:太原理工大学,2012.

作者简介:

朱文静 女,(1990-),硕士研究生.研究方向为语音信号处理.

白 静(通讯作者) 女,(1965-),教授,硕士生导师.研究方向为语音信号处理. E-mail: bj613@126.com.

(上接第 88 页)

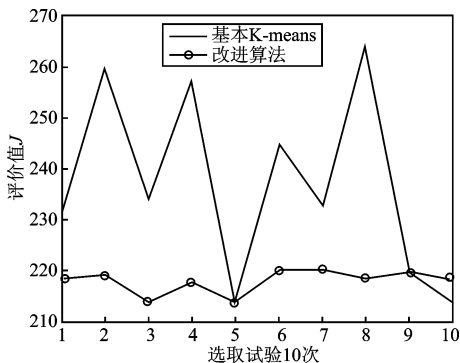


图4 Glass 数据集

在初始聚类中心随机选择导致聚类结果波动较大和聚类数目确定的缺点进行改进.但在确定聚类数目值的时候,采用了与经验值比较的方法,影响了准确率.下一步工作是提出一种合理判定聚类结束条件的方法,减少人为因素影响,以便能更加准确地获得聚类数目.

参考文献:

- [1] 雷小锋,谢昆青,林帆,等.一种基于 K-Means 局部最优的

高效聚类算法[J].软件学报,2008,19(7):1683-1692.

- [2] 贺玲,吴玲达,蔡益朝.数据挖掘中的聚类算法综述[J].计算机应用研究,2007,23(1):10-13.
- [3] 孙吉贵,刘杰,赵连宇.聚类算法研究[J].软件学报,2008,19(1):48-61.
- [4] 胡伟.改进的层次 K 均值聚类算法[J].计算机工程与应用,2013,49(2):157-159.
- [5] 徐沁,罗斌.结合 means-shift 与 MST 的 K-means 聚类算法[J].计算机工程,2013,39(12):214-210.
- [6] Galluccio L, Michel O, Comon P, et al, Graph based kmeans clustering[J]. Signal Processing, 2012, 92(9): 1970-1984.
- [7] 李翔宇,王开军,郭躬德.基于网格最小生成树的聚类算法选择[J].模式识别与人工智能,2013,26(1):34-41.

作者简介:

贾瑞玉 女,(1965-),副教授,硕士生导师.研究方向为智能计算与数据挖掘. E-mail: 764702886@qq.com.

李 振 男,(1991-),硕士生.研究方向为数据挖掘.