基于 SSE 的全局最优 K—means 算法

文/董炎焱

摘要

传统的 K-means 聚类算法对初值敏感，随机的初始聚类中心会造成簇的不稳定。本文采取全局搜索的方法避免了局部最优解，实验证明，采用 SSE 作为分类的标准，可以提高簇的稳定性。

【关键词】K-means 聚类 SSE 全局最优解 初始聚类中心

1 引言

聚类分析能够实现数据的归类，是数据挖掘的重要方法。K-means 在聚类算法中的收敛速度较快，可以对数据进行预处理，产生数据的基本分布规律，但是传统的 K-means 算法中人为确定聚类数 k，初始聚类中心的 k 个点随机选取，均影响到了聚类结果。本文针对 k 个聚类点的选取提出改进，增加 K-means 算法的稳定性。

2 K-means聚类算法

2.1 传统的K-means算法

设有数据点集 $\{X_u\}$，X_u 是 u 维空间的一个点，u 表示全部属性个数，人工设定聚类数为 k：

（1）在 $\{X_u\}$ 中任取 k 个初始聚类中心点，记为 $\{W_u\}$，k<u，$W_u \in X_u$；

（2）计算 X_u 和 W_u 的欧氏距离，归于最近的簇；

（3）更新聚类中心点，以各簇的均值代替原聚类中心点；

（4）重复（2）（3），直到连续两次聚类中心的距离小于或等于某阈值。

2.2 K-means的收敛测度SSE

聚类效果体现于聚类函数 SSE 的值，若 SSE 的值越小，认为聚类效果越好。设 $SSE=\sum\sum \| Xu-Wu \|^2$，对 W_u 求偏导数，并取为 0，得到 $Wu=\frac{1}{m}\sum Xu$，m 是以 W_u 为聚类中心的点个数，$\frac{1}{m}\sum Xu$ 就是 SSE 函数在 W_u 类的最优解，每一次迭代，SSE 将减小，最终趋于收敛。

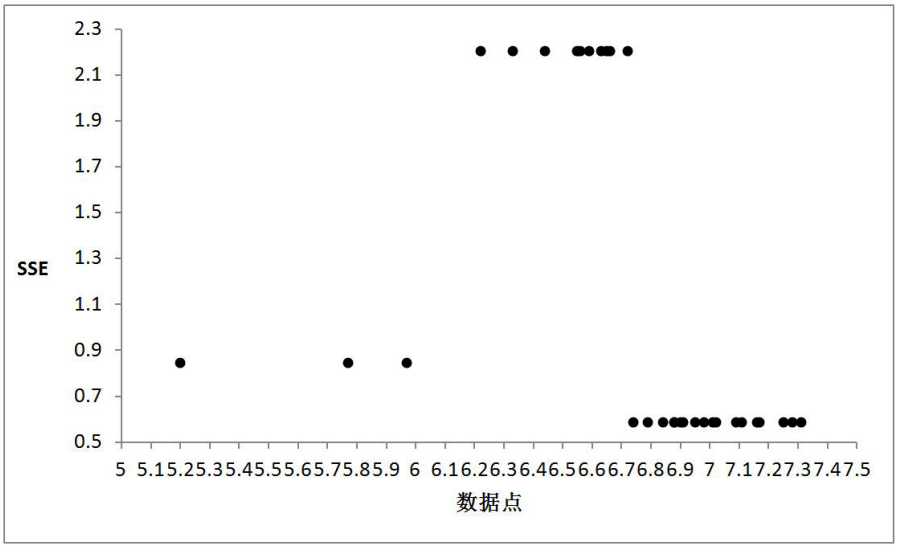

图 1：各数据点作为初始聚类中心的 SSE

2.3 传统K-means的局限性

聚类数人为确定，在大多数情况下，以人的先验知识不足以分清类别，要么 k 值偏小，忽略差别，要么 k 值偏大，过分强调类别，因此 k 值的选取需要多次的尝试，得到较为合理的聚类数。

SSE 是非凸函数，由于初始聚类中心的选取是随机的，会形成局部最小值，不能保证是全局最小值，可多次更新初始聚类中心，重复算法，取其中最小的 SSE。

3 全局最优解的K-means聚类算法

3.1 算法原理

设数据集点 $\{X_u\}$，k 为聚类个数，$\{W_u\}$ 为聚类中心点集：

（1）k=1，求解 $\frac{1}{m}\sum Xu$，其中 m 为数据点个数，得到第一个初始聚类中心 w_u^1；

（2）k=2，将第一个聚类中心 w_u^1 分别与 x_u^1，x_u^2，x_u^3，……x_u^m 进行 K-means 聚类，分别求出每次聚类的 SSE_i，找到 $\min\{SSE_i\}$，记录与之对应的第二个聚类中心 w_u^2；

（3）k=3，将 w_u^1、w_u^2 分别与 x_u^1，x_u^2，x_u^3，……x_u^m 进行 K-means 聚类，记录与 $\min\{SSE_i\}$ 对应的第三个聚类中心 w_u^3；

（4）依次类推，其中 k<m。

3.2 对比实验

实验数据来源为中华人民共和国统计局发布的“第六次人口普查”中“1-8 各地区分性别、受教育程度的 6 岁及以上人口”的统计数据，选用该数据的原因是不考虑异常数据对实验的影响，取对数后进行聚类分析。

传统的 K-means 算法对初始聚类中心点随机选取，得到的聚类结果不稳定，设 k=3，三次实验分别取不同的初始聚类中心，结果如表 1。

实验四为全局最优解的 K-means 聚类算法，第一个初始聚类中心是数据集的均值 w_u^1=6.73，数据集的每个点分别作为第二个初始聚类中心进行 K-means 聚类，得到 SSE，如表 2。

将表 2 的数值以图形表示，如图 1。

3.3 实验分析

传统的 K-means 聚类算法对初始聚类中

●课题名称：基于职教云平台的混合式学习实践研究；课题类型：规划课题；课题编号：GH-17154。

表 1：传统 K-means 算法的 3 次实验结果

	实验一			实验二			实验三		
类别	1	2	3	1	2	3	1	2	3
初始聚类中心	7.17	7.28	6.74	5.97	5.20	6.72	7.31	6.63	6.59
最终聚类中心	6.59	7.06	5.65	6.14	5.2	6.91	7.05	6.57	5.65
SSE	0.32	0.36	0.31	1.52	0	0.21	0.36	0.32	0.31
聚类情况	1	北京、天津、内蒙古、吉林、上海、福建、海南、重庆、贵州、云南、甘肃、新疆		1	天津、海南、青海、宁夏		1	河北、山西、辽宁、黑龙江、江苏、浙江、安徽、江西、山东、河南、湖北、湖南、广东、广西、四川、陕西	
	2	河北、山西、辽宁、黑龙江、江苏、浙江、安徽、江西、山东、河南、湖北、湖南、广东、广西、四川、陕西、		2	西藏		2	北京、天津、内蒙古、吉林、上海、福建、海南、重庆、贵州、云南、甘肃、新疆	
	3	西藏、青海、宁夏		3	北京、河北、山西、内蒙古、辽宁、吉林、黑龙江、上海、江苏、浙江、安徽、福建、江西、山东、河南、湖北、湖南、广东、广西、重庆、四川、贵州、云南、陕西、甘肃、新疆		3	西藏、青海、宁夏	

表 2：各数据点作为初始聚类中心的 SSE

数据点	SSE	数据点	SSE	数据点	SSE	数据点	SSE
6.73	2.204	7.16	0.585	7.02	0.585	6.72	2.204
6.33	2.204	6.95	0.585	7.09	0.585	5.20	0.847
7.17	0.585	6.91	0.585	7.31	0.585	6.84	0.585
6.88	0.585	7.01	0.585	6.9	0.585	6.55	2.204
6.63	2.204	6.79	0.585	6.22	2.204	5.77	0.847
6.98	0.585	6.88	0.585	6.3	2.204	5.97	0.847
6.74	0.585	7.25	0.585	7.11	0.585	6.56	2.204
6.59	2.204	7.28	0.585	6.65	2.204	6.66	2.204

心选取敏感，造成簇的不稳定性。分析数据集 $\{X_u\}$，$\bar{X}$=6.73，σ=0.462，$\bar{X}$±3σ 的范围是 (5.344，8.116)，当数据点在这个范围内时，随机选择的初始聚类中心对聚类的稳定性影响小，否则会产生奇异的簇，如实验二。

实验四中如果按照全局最优解的理论算法，需要找到 min{SSE}，才能确定第二个聚类中心，但是通过数据点与 SSE 的图 1 就可以发现数据点已明显的分为三类，三个簇的 SSE 分别是 0.585，0.847，2.204，数据点小于 w_u^1，SSE 就大。将 SSE 所对应的聚类中心作为初始聚类中心进行传统的 K-means 聚类，迭代 5 次后，最终的聚类中心是 6.57，7.05，5.65，每个簇的 SSE 是 0.32，0.36，0.31。实验四与实验一、三的最终聚类中心和 SSE 接近。

3.4 全局最优解的K-means聚类算法的改进

设数据集点 $\{X_u\}$，k 为聚类个数，$\{W_u\}$ 为聚类中心点集：

（1）k=1，求解 $\frac{1}{m}\sum Xu$，其中 m 为数据点个数，得到第一个初始聚类中心 w_u^1；

（2）将第一个聚类中心 w_u^1 分别与 w_u^1，w_u^2，w_u^3，……x_u^m 进行 K-means 聚类，分别求出每次聚类的 SSE_i，按照 SSEi 进行分类，相同 SSE_i 的归为一簇；

（3）将不同簇的最终聚类中心作为初始聚类中心，进行 K-means 聚类，得到聚类结果。

4 结论

传统的 K-means 聚类高效而简单，应用范围广，但是随机的初始聚类中心和局部最优的存在影响了聚类的稳定性。本文从结合前人的研究成果对全局最优解的 K-means 聚类提出改进，缩短多次全局搜索的时间，增加聚类的稳定性。

参考文献

[1] 谢娟英，蒋帅，王春霞，张琰，谢维信．一种改进的全局K-均值聚类算法 [J]．陕西师范大学学报（自然科学版），2010,38(02):18-22

[2] 王晓东，张姣，薛红．基于蝙蝠算法的K均值聚类算法 [J]．吉林大学学报（信息科学版），2016,34(06):805-810.

[3] 周世兵，徐振源，唐旭．新的K-均值算法最佳聚类数确定方法 [J]．计算机工程与应用，2010,46(16):27-31.

[4] Park H S, Jun C H.A simple and fast algorithm for K-medoids clustering [J]. Expert SystemswithApplications, 2009, 36(02): 3336-3341.

[5] 王红睿，赵黎明，裴剑．均衡化的改进 K 均值聚类法 [J]．吉林大学学报（信息科学版），2006,24(03):172-176.

作者简介

董炎焱（1972-），女，山西省太谷县人。大学本科学历。晋中师范高等专科学校数理科学系讲师，主要从事数据挖掘研究。

作者单位

晋中师范高等专科学校　山西省晋中市 030600