基于大数据的配网数据挖掘研究

刘宏森[1]　李　娟[1]　程　斌[1]　李　睿[2]　张建国[2]

（1.国网铜川供电公司，陕西铜川 727031；2.保定市睿为电气科技有限公司，河北保定 071051）

摘　要　针对海量的配网数据挖掘需求，提出一种基于 Hadoop 的海量运维数据存储与分析方法。在该方法中，采用 Hadoop 对运维数据进行存储；然后针对常用的数据 K 均值分类方法，针对该方法存在的问题，提出采用密度聚类对初始点选择进行改进，然后运用手肘法对 K 值进行确定，以提高聚类的精度。最后搭建 Hadoop 集群环境对配网数据进行挖掘，结果表明，上述方法有效。

关键词　Hadoop 框架；K 均值聚类；数据分类；手肘法

Research on distribution network data mining based on big data

Liu Hongsen[1]; Li Juan[1]; Cheng bin[1]; Li Rui[2]; Zhang Jianguo[2]

(1. State Grid Tongchuan power supply company, Tongchuan Shaanxi 727031;

2. Baoding Ruiwei Electrical Technology Co., Ltd.,　Baoding Hebei 071051)

Absrtact　Aiming at the demand of massive distribution network data mining, a storage and analysis method of massive operation and maintenance data based on Hadoop is proposed. In this method, Hadoop is used to store the operation and maintenance data; then, aiming at the common K-means classification method, the density clustering method is proposed to improve the initial point selection, and elbow method is used to determine the K value to improve the accuracy of clustering. Finally, the Hadoop cluster environment is built to mine the distribution network data, and the results show that the above method is effective.

Key words　Hadoop framework; K-means clustering; data classification; elbow method

信息化技术逐步渗透到了各个行业中，各种类型的信息化系统在配电网系统中得到了较多的应用，改善了配电网运行和管理的质量。配电网运行的数据存储在信息化管理系统中，可以通过查询获取到需要的数据，同时采用数据挖掘等方法可以深入挖掘隐藏在数据中的信息，以此能够为配电网的维护提供准确的依据。随着配电网信息化建设的持续深入，无论是数据的规模还是复杂度都在增大，对于数据的管理和应用提出了更高的要求，如果只是依赖于传统的挖掘方式已经难以获取到有价值的信息。大数据技术在多源异构数据处理中具有广阔的应用潜力，因此可以在配电网数据处理中引入大数据技术，基于此方式挖掘数据中的深层次信息，并将这些信息应用到配电网管理决策中，从而提升配电网运行的质量。与此同时，如何结合海量的数据进行配网数据挖掘，是当前思考和研究的重点。对此，本研究结合当前的大数据相关技术，提出一种基于大数据平台的配网数据分类方法。

1 K—Means 基本原理

K-均值算法(K-means)的原理较为简单，实施难度较小，广泛应用在分类以及决策等领域中。K-means 算法主要使用欧式距离对样本间的距离进行计算，可以将给定的数据划分为 k 个类别,该算法显著提升了对大规模数据集的处理效率。此外,通过这种方式还能够解决大数据集聚类处理时的实时性问题，降低了对资源的占用，因此 K-means 算法广泛应用。尽管

收稿日期：2020−9−1

作者简介：刘宏森（1966−），男，汉族，陕西渭南，本科，高级工程师，研究方向为电网建设与运行管理。

K-means 算法被应用到大数据处理以及分类等领域中，但仍然存在一定不足。首先是容易陷入局部最优解的问题，由于算法执行过程中需要先确定初始聚类中心，然后通过迭代的方式来判断是否满足收敛条件，所以初始聚类中心选择的合理性将会直接影响到最终的结果；其次是确定 k 值的难度较大，如果用户对于数据比较熟悉，则容易得到更为合适的 k 值，但是在实际应用中对于数据往往缺乏了解，因此难以保证 k 值选取的合理性。

2 K—Means 算法改进

基于以上对 K-Means 算法问题的法纳西，提出以下几点改进：

2.1 初始点选取

本研究中，根据数据密度选取合适的初始点。具体是先对数据间的平均密度进行计算，将密度最大的点当作首个初始聚类中心；然后对第二次聚类中心进行选择，即保证密度高于$\sqrt{n}$并且与初始聚类中心的距离最大，接着选择距离二者最大的最小距离，并将其作为第三个聚类中心；基于这种方式，可以得到 k 个聚类中心。详细的执行过程如下所示：

1）首先对全部数据间的欧式距离进行计算，具体公式如下所示。

$$MeanDis(D)\frac{\sum_{i=1}^{n}\sum_{j=n+1}^{n}d(x_i-x_j)}{\frac{n(n-1)}{2}} \tag{1}$$

在上述公式中，n、D 代表的是数据量、数据集。

2）其次计算对象间的密度，提取高于$\sqrt{n}$的点得到对应的密度集。

$$Den(x_i)=\sum_{j=1}^{n}u(MeanDis(D)-d(x_i-x_j)) \tag{2}$$

3）选择首个初始点 k_1，主要从先前得到的密度集中选择密度值最大的点，在此基础上提取与 k_1 距离最大的点 k_2。

4）继续选取与 k_1、k_2 最大的最小距离的点 k_3。

5）然后选取与已有聚类中心最大的最小距离的点。

6）在初始点数量达到 k 个时，开始采用 k -均值算法计算。

在本次研究中根据$\sqrt{n}$来划分密度点。

2.2 k 值个数确定

在研究过程中基于误差平方和（SSE）来确定 k 值个数，使用较多的方法是手肘法，其中 SSE 的具体公式如下所示。

$$SSE=\sum_{i=1}^{k}\sum_{p\in c_i}\left|p-m_i\right|^2 \tag{3}$$

在上述公式中，c_i 代表聚类 i，m_i、P 分别对应着 c_i 的质心和样本点。

SSE 的思想如下所示：在聚类个数增多时，样本划分的精细度会逐步提高，由此进一步增强了各个聚类的聚合度，所以对应的 SSE 减小。根据 k 与真实聚类数量之间的大小关系可以对 SSE 的变化趋势进行确定，如果前者高于后者，则在 k 增大时，SSE 的变化趋势相对平缓，不会出现剧烈的变化；反之，在 k 增大时导致 SSE 急剧减小。因此只需要依据 SSE 与 k 值关系图的拐点即可确定对应的聚类个数。此外，根据轮廓系数可以确定数据集的类别信息。

已知样本集 D 划分为 k 个簇，分别表示为 $c_1,c_2,...,c_k$。其中，p 属于 c_i 内的样本点，由此可以得到：

$$a(p)=\frac{\sum_{p\in c_i,p\neq p}dist(p,p')}{\left|c_i\right|-1} \tag{4}$$

$$b(p)=\min_{c_i 1\le j\le k,j\neq i}\left\{\frac{\sum_{p\in c_i}dist(p-p')}{\left|c_j\right|}\right\} \tag{5}$$

该点的轮廓系数表示为：

$$s(p)=\frac{b(p)-a(p)}{\max\{a(p),b(p)\}} \tag{6}$$

在上述公式中，$b(p)$代表各个簇之间的远离程度，$a(p)$代表相同簇中个体之间的相似度，轮廓系数取值范围是[-1,1]，如果基本等于 1，则对应着更佳的聚类

效果，如果远离 1，则聚类效果不佳。在本次研究中选择 k 值的具体方法如下所示：

1）需要先对各个 k 值的 SSE 进行计算，其中 k 值取值范围为 $2 \leqslant k \leqslant \sqrt{n}$。

2）考虑到肘形图分析时可能出现的误判问题，可以计算出各个 k 值对应的夹角，具体公式如下所示；

$$\theta = \pi - \alpha + \beta \tag{7}$$

最佳的 k 值选择点应该是 90 度的夹角。

3）如果在第 2）步中有多个 k 值的夹角基本一致，则需要进一步对平均轮廓系数进行计算，并选择较大值对应的 k 值。

3 算法验证

3.1 运维数据预处理

为更好的做好海量运维数据的预处理，本文搭建基于大数据技术 Hadoop 的配网运维数据处理方法整体结构图，如图 1 所示。

图 1 基于 Hadoop 的运维数据处理

在上述的架构中，以 Hadoop 框架对数据的存储、分析等进行搭建，通过 HDFS 文件管理系统完成对数据的存储。同时通过数据合并、数据清洗、数据处理三个环节完成对运维数据的预处理。而运维数据的获取方面则主要源自于 95598 营销系统、配网自动化系统等。

3.2 集群环境搭建

在本次研究中需要搭建 Hadoop（2.8.1 版本）集群，集群中总计包括四个节点，分别是一个主节点（Master）和三个从节点（Slave）。此外，还需要配置 CentOS 系统、JDK1.8.0_144、Tableau 工具，挖掘算法选用的是 PrefixSpan 与 K-means 算法。集群具体结构如下图所示。

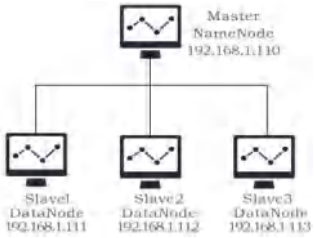

图 2 平台结构示意图

3.3 运维数据信息

在研究中导出的运维数据超过 300 万条，设计对应的数据表来管理不同的运维数据，具体数据信息如下表所示。

表 1 运维数据表

数据表类型	数量（条）
故障停电	高于 8000
非计划停电	7000 上下
计划停电	17000 上下
配变重过载	大约 6 万
配变轻空载	大约 7 万
运检投诉	大约 1000 条
用户过电压	高于 160 万
用户低电压	120 万
遥信数据	高于 8000
遥控数据	高于 6000

3.4 聚类结果评价

采用表 1 所述方法计算运维数据的 SSE，并绘制折线图，如下图所示，折线标注为该点的处直线的夹角，从图中可以看出 k = 5 时为夹角最小，最接近于折线的拐点，因此确定聚类个数为 5 。同时从图 3 也能看出 k = 5 时，平均轮廓系数也较大。

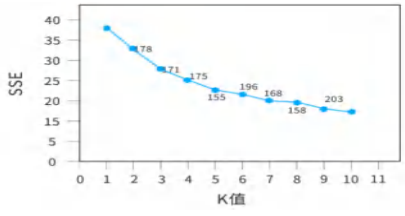

图 3 SSE−K 折线图

基于先前所述的方法计算得到了对应的指标数据，详细信息参见下表。

表 2 聚类中心信息表

	聚类序号	1	2	3	4	5
	数量	50	1	33	15	16
	数量占比	43.48%	0.80%	28.69%	13.04%	13.91%
聚类中心	客户投诉	0.15	1.00	0.15	0.24	0.27
	故障停电	0.25	1.00	0.24	0.47	0.60
	计划停电	0.34	0.16	0.19	0.49	0.33
	非计划停电	0.22	0.32	0.23	0.44	0.41
	配电轻空载	0.22	0.11	0.19	0.41	0.15
	配变重过载	0.28	0.77	0.29	0.16	0.50
	摇信错误率	0.79	0.08	0.26	0.36	0.70
	过电压占比	0.14	1.00	0.12	0.16	0.38
	低电压占比	0.08	1.00	0.08	0.08	0.31

在本次研究中针对两种选择初始点的方法进行了对比，分别是 K-means 算法和基于密度的方法，通过聚类分析过程对各个 k 值聚类结果轮廓系数的平均值进行了计算，最终得到的结果如图 4 中所示。根据图中的信息可知，基于密度的选择方法能够达到更优的聚类效果。

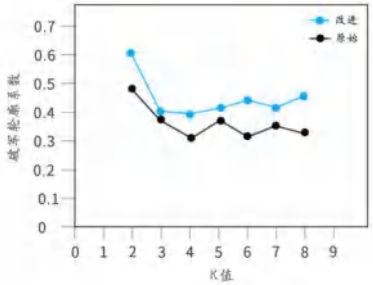

图 4 本文算法与 K 均值算法的平均轮廓系数对比

4 结束语

本文针对 K-means 在选择初始点中的不足设计了一种基于密度的改进方法，以减少初始点传统选择的主观性。然后采用误差平方和SSE对k值进行确定，以确定聚类种类。最后，通过搭建 Hadoop 仿真平台，对上述的改进进行验证，得到本改进要优于传统算法。

参考文献

[1] 张国宾，王晓蓉，邓春宇.基于关联分析与机器学习的配网台区重过载预测方法[J].大数据，2018，4(01)：105−116.

[2] 李端超，王松，黄太贵，程栩，许小龙，窦万春.基于大数据平台的电网线损与窃电预警分析关键技术[J].电力系统保护与控制，2018，46(05)：143−151.

[3] 乔俊峰，王一清，杨佩，谢韬.基于大数据技术的配电网全景监测系统的研究与实现[J].供用电，2018，35(03)：41−46+51.

[4] 张嵩，刘洋，许芳，侯喆瑞.配电网中大数据的挖掘应用[J]. 电力大数据，2018，21(02)：8−12.

[5] 补敏，丁泽俊，钟锦群，刘桓瑞.基于配用电大数据的供电电压监测与分析研究[J].电力大数据，2019，22(03)：1−7.

[6] 谢涛，石雪梅，钟成元.基于态势感知驱动的配电网投资决策体系架构设计探讨[J].电力与能源，2018，39(03)：329−332+343.

[7] 杨东宁，冯磊，马龙，罗骥辉.基于多源异构数据集成开发的主配网规划数据应用[J].电子技术与软件工程，2018(24)：148−149.

[8] 朱琨.美林：打造 BI+AI 工业大数据平台[J].软件和集成电路，2018(12)：28−32.

[9] 邢杰，唐泽洋，刘焱，周鲲鹏，邱丹，曹侃.基于电网运行数据的配网线变关系校验影响因素分析[J].电力大数据，2019，22(02)：13−19.