文章编号：2095-2163(2020)11-0092-05 中图分类号：TN929.5 文献标志码：A

基于大数据的用户负荷特性分析系统的研究与应用

张丽华[1]，张伟民[2]，刘 春[1]，贾美娟[1]，杨 瑞[1]，邹雨轩[1]
（1 大庆师范学院 计算机科学与信息技术学院，黑龙江 大庆 163712；
2 大庆油田有限责任公司试油试采分公司，黑龙江 大庆 163712）

摘 要：针对重点行业（如采油、冶炼加工等）的具体负荷特性，构建一个分析模型至关重要。本文通过对比分析、经验总结及仿真实验等研究方法，对各类用电负荷的特性及电力负荷特性曲线、指标等因素的掌握，借助服务器和 Hadoop 等相关技术搭建数据处理平台，择优选取 k-means 聚类算法搭建模型，实现负荷预测分析结果的判定，完成了标签库的设计及体系架构。该系统可以更精准的掌握市场潜力，预判采油开发、冶炼加工等行业，特别是对未来用电的总体趋势，充分发挥相应指标及权重适应性，为电力市场分析、电力系统负荷预测提供了相应的参考依据与保障。

关键词：数据挖掘；负荷特性；k-means 聚类 k 均值；极大值标准化

User load characteristic analysis system based on big data Design and implementation

ZHANG Lihua[1], ZHANG Weimin[2], LIU Chun[1], JIA Meijuan [1], YANG Rui[1], ZOU Yuxuan[1]
(1 School of Computer Science and Information Technology, Daqing Normal University, Daqing Heilongjiang 163712, China;
2 Daqing Oilfield Co., Ltd. Oil Testing and Production Test Branch, Daqing Heilongjiang 163712, China)

【Abstract】 In view of the specific load characteristics of key industries (such as oil extraction, smelting and processing, etc.), it is important to construct an analysis model. Through comparative analysis, experience summary and simulation experiment and other research methods, the characteristics of various power loads and power load characteristic curves, indicators and other factors are globally grasped, and the data processing platform is built with the help of servers and Hadoop and other related technologies, and then the preferred k- The means clustering algorithm builds a model, realizes the judgment of the load forecast analysis results, and completes the design and system architecture of the final tag library. After construction, it can more accurately grasp the market potential, predict the oil production development, smelting and processing industries, especially the general trend of electricity consumption in the future, and give full play to the adaptability of the corresponding indicators and weights, providing power market analysis and power system Load forecasting provides the corresponding reference basis and guarantee.

【Key words】 Data mining; Load characteristics; k-means clustering; k-means; Maximum standardization

0 引 言

随着电力行业负荷管理调整政策出台，以及信息化水平的不断提高，重点行业（如冶金、石油开采等）因电量数据残缺与分类范畴不明等问题，得到了极大的改善。从而为开展电力需求等各项指标相关联问题的研究工作提供了重要的数据基础，使该类研究工作的开展成为可能。负荷的大小与特征，对于电网规划及电网运行管理，都是极为重要的因素。所以，对负荷的变化和特点，事先估计是电力系统规划与运行研究的重要内容。

本文采用数据挖掘技术，对某地区主要行业各类用电负荷的特性及电力负荷特性曲线、指标等因素进行分析，有利于全面掌握该地区的用电结构及用电特征。其研究结果可应用于电力营销的用户分类、电费预警等相关工作中。通过进一步研究，负荷特性指标数据还可挖掘得到统计指标间潜在的内在规律，给电力系统负荷预测、电网规划和安全稳定运行提供相应的参考依据。

1 数据挖掘技术

数据挖掘，是指基于数据库原理、云技术、机器学习、人工智能、现代统计经济学等跨学科的综合技术。伴随着所涉及学科的增多，其实用价值及应用可在很多领域加以展现。所相关联涉及到的算法也多种多样，诸如神经网络、决策树、基于统计学习理

基金项目：大庆市指导性科技计划项目（zd-2019-69）；黑龙江省教育科学“十三五”规划年度重点课题（GJB1319002）；2019 大庆师范学院科学研究基金资助项目（19ZR10）；大庆市指导性科技计划项目（zd-2019-60）。

作者简介：张丽华（1980-），女，硕士，讲师，主要研究方向：计算机网络及云计算；张伟民（1975-），男，学士，高级工程师，主要研究方向：油气田；刘 春（1970-），女，硕士，教授，主要研究方向：数据库、云计算及软件工程。

收稿日期：2020-09-30

论的支持向量机、分类回归树和关联分析等。数据挖掘的内容主要包括分类、关联分析、聚类和异常检测等。数据挖掘技术[1]是综合了传统数据分析法与新型复杂算法的新方法，该方法能在大型数据存储库之中，自动的发现有用信息。数据挖掘作为数据库知识发现（KDD）不可缺少的一部分，在整个数据库知识发现过程中发挥着重要作用[2]。

“大数据”时代的数据挖掘，是从大量数据的定义中发现有意义的模式或知识。数据挖掘离不开数据库技术的发展和成熟，具体地说，对数据库中的原始数据进行一系列的计算和分析，得到有用的信息。该过程通常包括数据预处理、数据挖掘、后处理 3 个方面。其中，数据预处理是指将原始数据转换成一种适合分析的形式，包括多数据源的数据融合、数据清理、尺寸标注等。后处理是指对模型预测的结果进行进一步的处理和推导。

本文主要分析电力负荷（主要为负荷曲线）的相关特性，对不同用电特性的用户加以区分。目前利用配用电数据进行异常用户检测、需求侧管理与能效管理、用电客户精细分类等都需要对负荷曲线进行聚类分析。因此，有必要通过对负荷曲线的聚类对其特性进行分类研究。

2　聚类分析

聚类分析是对描述对象的信息和相似性进行分析和分组。常用的聚类方法包括 K 均值、凝聚层次聚类和 DBSCAN 聚类等方法。其作为数据挖掘中数据筛选的重要方法，在电力行业中也开始广泛的应用。

聚类分析是对描述对象的信息进行分析，按照数据对象的相似性进行分组。常用的聚类方法包括 K 均值、凝聚层次聚类和 DBSCAN 聚类等方法。其作为数据挖掘中数据筛选的重要方法，在电力行业中也开始广泛的应用。周晖、王毅等人利用灰色聚类方法以及 K-Means 聚类算法等，不仅对各行业用电量[3]，也对客户欠费特征数据等进行分类[4]，为电网公司提供负荷管理和电价制定的依据。

为了更高效的提供电力服务，对电力负荷的预测则需要更加精准。传统 BI 已不能满足现今的需求，而大数据分析能更好进行预测分析，这也是电力大数据应用与传统数据仓库和 BI 技术的关键区别之一[5]。随着电力系统信息化程度不断提高，如何合理地利用这些数据并从中提取有价值的信息，是目前电力系统面临的一个重要问题。电力大数据是能源变革中电力工业技术革新的必然途径，而不是简单的技术范畴；其不仅仅是技术进步，更是涉及整个电力系统在大数据时代下发展理念、管理体制和技术路线等方面的重大变革，是下一代智能化电力系统在大数据时代下价值形态的跃升。

“配用电侧”是电力系统数据源的主要指标，特别是随着智能电网技术的快速发展，各种先进的检测装置和计量设备在配电网中得到了广泛应用[6]。电网公司的多种不同业务系统集成，对数据集成系统的多源海量数据进行有效的数据挖掘是智能电网发展的必然趋势。

目前，在新的智能化技术形势下，负荷数据量急剧加大，数据缺失也更加频繁，本次研究将针对这些问题，采用数据准备、数据预处理与数据挖掘技术解决相应问题，为电网企业不断发展提供支持。

3　电力负荷数据准备及预处理

3.1　技术路线

本研究的技术路线如图 1 所示，首先研究电力负荷预测在当下的现状，主要分析电力负荷（主要为负荷曲线）的相关特性，并对不同用电特性的用户加以区分；然后研究数据处理平台的搭建结构以及设计新的网络层次结构，并对所需要分析处理的数据做相应计算；最后运用 k-means 聚类算法，确定“最佳聚类数”搭建模型，并将其应用于某重点行业（如采油、冶炼加工等）负荷特性预测分析。

3.2　数据处理平台搭建的硬件结构和软件环境

该硬件架构如图 2 所示，共计使用 3 台服务器，每台服务器为 8 * 8G 内存，2 * 4T 硬盘，2 * E5 处理器，并通过 8 口 300M TP-link 进行网路连接，以保证数据传输过程的效率与稳定性。软件环境主要采用的 Hadoop 为基于 CentOs 6.7 的 2.6 版，通过 Yarn[7]分布式计算。

3.3　数据来源

数据主要来源于采集系统和调度系统。主要表关系如图 3 所示。其中对内部数据的采集主要包括：

（1）客户基础信息数据（ C_CONS）。用户数据来源此表，其中包含用户编号、用户标识、用户中文名称、地区信息、合同容量信息、行业分类等。此表主键为用户标识，并通过用户编号与外表进行关联。

（2）日负荷基础数据（E_MP_POWER_CURVE）。本文所使用的日 96 点负荷数据来源于此表。表中包含计量点编号、采集数据类型、96 点实际负荷、电压电流比率等信息。此表主键为计量点编号，并通过主键与外表关联。

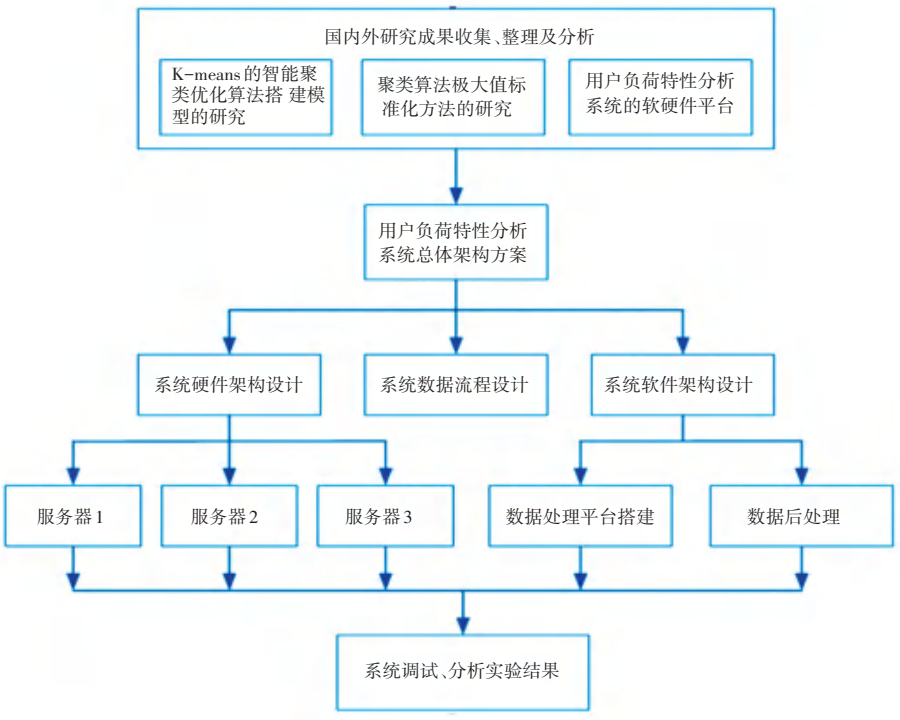

图 1 技术路线图

Fig. 1 Technology roadmap

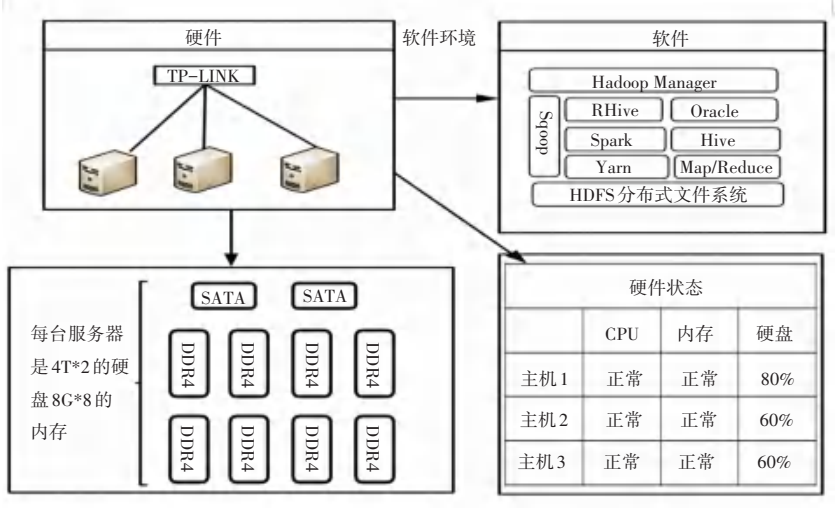

图 2 硬件架构图

Fig. 2 Hardware architecture diagram

(3)日电量数据(E_MP_DAY_READ)。日电量表底示数来源此表。表中包含计量点编号、采集数据类型、日电量表底示数、电压电流比率等信息。此表主键为计量点编号,并通过主键与外表关联。

(4)计量点基础信息(E_DATA_MP,C_MP,C_METER_MP_RELA)。计量点信息表包含计量点的ID、倍率、表属性等信息。

各表间关系如图 3 所示。

3.4 数据存储与转换

本数据取自采集系统,采用以 Oracle 10g 作为数据库主要版本,导出的数据泵文件,通过数据泵进行导入导出,其数据格式为 Dump。鉴于 Hadoop 作为数据处理主平台,为了便于 HDFS[8] 文件系统进行分布式处理,需要进行相应数据中转,将数据导入 Oracle 后生成 Oracle 数据,通过 Sqoop 将数据导入 Hadoop,最终将其导入 Hive[9],具体实现过程如下:

图 3　表关系图

Fig. 3　Table diagram

（1）将 Oracle 数据由 exp/imp 导出至外部存储（存储格式为 Dump）。基础数据以月做数据基础表进行导出；档案数据以整体导出，五地市数据存储于一张表内。

（2）根据数据字典创建数据表，并在服务器内分配表空间与物理存储空间。由于单个表文件无法存储超规模数据量，且固定表的处理效率远高于自增表，因此共创建 100 个数据文件。为避免由于表空间不足导致存储失败，每个数据文件均分配 30G 的固定存储空间。建表后创建索引和分区，以便加快后续查询速度。

（3）通过数据泵将数据文件导入 Oracle，并通过 Sqoop 将数据导入 HDFS。由于数据量大，在数据导入 Hive 前先将 Oracle 中的表增加时间字段以便于分区，利用 Linux 的 Shell 命令将时间字段加入表中后，将表导入 Hive 中，进而在 Hive 中对表进行分区操作，便于数据查询及分析。

3.5　负荷数据准备

3.5.1　数据选取

根据日测量点功率曲线表给出的字段来选取数据，筛选得到正向有功的数据；

根据行业代码，选取（如采油、冶炼加工等）行业数据进行重点分析，经筛选得到完整记录数据。

3.5.2　数据填补和清洗

“大数据”的数据量在累积中，面临数据类型繁杂且量级参差不齐，模型无法将其直接使用等问题。因此，需要对数据异构海量信息进行归一化处理、存储与相应转换。此外，还需对负荷数据值中缺损部分给予及时填充和清洗修正。针对不同数据综合采用以下几种处理方式：

（1）首末端缺失数据处理。若用户当日的末端数据（如某点数据）缺失，则以该用户次日首端数据进行填补；反之，当用户首端数据缺失，则以该用户前日末端数据进行填补。

（2）单个数据空缺处理。在已知当日单个缺失数据前后负荷点数据时，可借助列插值法，即该点数据前后均值来填补数据。

（3）连续多个数据缺失处理。利用行插值法，需要考虑到缺失数据个数的奇偶性。若寻找的数据在中间有连续多个数据缺失时，需寻找缺失数据 1/2 中心点数据。再由此方法，依次找寻 1/4 和 3/4 点处的数据。重复该操作，即可补全所缺失数据。

（4）删除空缺值多的记录。经由数据填补之后，仍有记录存在大量空缺，则进行删除处理。表中有用字段对应数据缺失值过多时，填补等方法将不能起到相应作用，且无法保证填补数据的有效性，将其删除以减小计算误差。

（5）删除表中的无效记录。首先删除表中含有负值记录，其次删除表中数据全为 0 的记录，最后删除表中出现极大值的记录。

3.5.3　数据归一化

根据聚类方法的要求，采用数据归一化方法对日负荷数据进行处理，将数据限制在 0~1 范围内。数据归一处理主要解决数据的可比性，常用的有“最大最小值化”、“0 均值标准化”和“极大值标准化”等。经过归一化处理，方便进行研究分析。本研究采用数据归一化方式的极大值标准化。极大值标准化可由式（1）表示：

$$x_{ij}^{'}=\frac{x_{ij}}{\max\{x_{ij}\}}(i=1,2,\cdots,m;j=1,2,\cdots,n). \tag{1}$$

式中，$x_{ij}^{'}$ 为标准化后的数据，x_{ij} 为聚类元素所对应的原始数据。经式（1）处理后，最大值为 1，其余各值均小于 1。

采用年最大负荷作为极大值，会导致对异常数据的极度敏感，可以通过前期对数据异常值的清洗消除影响。因此，采用其作为归一化极大值指标。

4　基于 k-means 聚类的重点行业负荷特性分析

4.1　模型搭建

4.1.1　k-means 聚类算法

采用 k-means 方法对用户的日负荷曲线进行聚类，能获取反映行业特性典型的日负荷曲线。k-means 可以处理数据量较大的情况，是基于划分的计算方法，时间复杂度较低，在给定 k 值后，可以较快的完成收敛；其次，该算法较为简单，不需要复杂的逻辑和方法；此外，对足够大的数据量进行聚类，可以保持结果准确一致，与初始随机选择的聚类中

心无关。

4.1.2 基于评价指标的最佳聚类数确定

聚类有效性研究是通过建立有效性指标,评价聚类质量并确定最佳聚类数的过程。本文选择运算复杂度较低的 SSE 和 CC 这二类完全相反的指标,作为选取最佳聚类数的考核指标。

(1)数据计算和抽取生成周负荷曲线。计算每日最大负荷,选取连续 7 天日最大负荷生成周负荷曲线。

(2)取不同聚类数 k(初始值取 2,逐次增加 1),采用 k-means 聚类方法对周负荷曲线进行聚类。

(3)聚类结果采用 R 语言计算 SSE 和 CC 指标,根据指标趋势图获取最佳聚类数 k。

4.2 重点行业负荷特性分析

本研究利用 k-means 聚类方法,对特殊行业负荷特性进行分析,及对行业用户某点日负荷数据进行预处理。综合处理步骤如下:

(1)筛选某点日负荷数据,并进行数据填补与清洗。

(2)按日期对该行业或其子行业的所有用户日负荷加和,计算该行业或其子行业每日综合日负荷曲线,进行初步挖掘分析。

(3)关联每个用户档案,以便于获取其年最大负荷,并采用极大值标准化方法,对数据作归一化处理。

(4)通过 SSE 和 CC 指标,计算该行业最佳聚类 k,随机选取初始聚类中心,对预处理后数据进行 k-means 聚类。

(5)随机更换初始聚类中心,进行 k-means 聚类。

(6)对不同初始聚类中心的聚类结果进行评价,选取最优聚类结果。

(7)根据最优聚类结果,对该行业用户日负荷特性进行分类,分析其生产特性并给出典型用户。

5 构建基于用户负荷特性的标签库

将负荷特性分析结果用于标签体系库建设的应用;将聚类结果进行整理并以标签库形式展现。如,冶炼加工行业标签库界面如图 4 所示。

图 4 标签库界面

Fig. 4 Tag library interface

6 结束语

本文提出的基于大数据的用户负荷特性分析系统的总体架构,以相对较为稳定的日 96 点电力负荷数据建立分类模型,获得针对某个重点行业,如采油、冶炼加工等典型场景用电负荷特性。通过实验仿真 SSE 和 CC 等指标变化图,实现对该行业用电特征分析;再抽取该行业用户用电特征,建立用户用电行为标签库,有利于网络收敛与稳定,提高负荷数据聚类化。

参考文献

[1] 孟小峰, 慈祥. 大数据管理:概念、技术与挑战[J]. 计算机研究与发展, 2013, 16(1): 146-169.

[2] 吴新垣. 从数据挖掘到知识发现[J]. 舰船电子工程, 2012, 6(2): 31-34.

[3] 周晖, 王毅, 王玮,等. 市场条件下电力客户欠费预警模型[J]. 中国电机工程学报, 2012, 15(4): 107-112.

[4] 吴新垣. 从数据挖掘到知识发现[J]. 舰船电子工程, 2014, 9(2): 41-43.

[5] 中国电机工程学会电力信息化专业委员会. 中国电力大数据发展白皮书[J]. 中国电力出版社, 2013, 19(2): 30-35.

[6] 张斌, 庄池杰, 胡军, 等. 结合降维技术的电力负荷曲线集成聚类算法[J]. 中国电机工程学报, 2015, 31(15): 3741-3749.

[7] 董西成.深入解析 YARN 架构设计与实现原理[M]. 北京:机械工业出版社,2013, 153-184.

[8] 林闯,苏文博,孟坤, 等. 云计算安全:架构、机制与模型评价[J]. 计算机学报, 2013, 21(3): 1765-1784.