摘 要

我们目前生活的时代是处于一个信息大爆炸的时代，而用电方面，随着家家户户的智能电表的普及，诸如各相功率因素在内的多项用电的功率数据之类的大量数据急需通过智能的算法进行归类，整理，总结，以提取出对发展有用的信息。

在火热的智能算法领域，聚类算法是目前公认的较好的解决方法，聚类算法一般是将数据点集合通过算法分成多个类别，因为没有固定的解答，因此采用的方法不同，其复杂程度以及结果均可能不同，而其中K-Means算法采用了欧几里得距离来判断数据点的归属，因此是比较常见且易操作的算法，它通过不断迭代将初始质心的位置最终改变至相对稳定的位置从而确定聚类的最终结果，从而得出聚类的模型参数。

K-Means算法的精准与否基本取决于三个参数：初始质心的选择，初始质心的个数以及数据的预处理程度，本论文希望从电力行业自身的角度出发，基于具体的数据，找寻解决三个参数的办法，使对用户的用电行为分析模型更加的精准，从而能让其更好的服务于电力公司的研究以及发展。

本论文中逐一介绍了K-Means算法的国内外研究进展，具体实现，算法精准相关参数的选择以及与用电数据的具体应用配合,以及最终代码的实现过程以及数据测试与呈现。

**关键词：**聚类算法，K-Means，Spark，大数据，用电行为分析

**ABSTRACT**

At present, we are living in an era of information explosion, and with the popularization of smart meters in every household, a large number of data, such as power data of various power factors, need to be sorted, made and summarized by intelligent algorithms to extract useful information for development.

In the hot field of intelligent algorithms, Clustering algorithm is recognized as a better solution at present, Clustering algorithms usually divide data points into several categories, because there are no fixed answers, so the methods used are different, The complexity and results may vary, K-Means algorithm adopts Euclidean distance to judge the attribution of data points, so it is a common and easy-to-operate algorithm. It determines the final result of clustering by changing the initial centroid position to a relatively stable position through continuous iteration, thus obtaining the model parameters of clustering.

The accuracy of the K-Means algorithm basically depends on three parameters: The choice of initial centroid, the number of initial centroid and the degree of data pre-processing, this paper hopes to find a solution to the three parameters from the perspective of the power industry itself, based on specific data, so that the analysis model of consumer behavior more accurate, so that it can better serve the research and development of power companies.

 This paper introduces the research progress of K-Means algorithm at home and abroad, the specific implementation, the choice of algorithm precision parameters and the specific application of power data, as well as the implementation process of the final code and data testing and presentation.

**KEY WORDS:**collaborative Clustering algorithm，K-Means，Spark，Big Data，Electricity Behavior Analysis

目 录

[摘 要 I](#_Toc17609)

[**ABSTRACT** II](#_Toc16795)

[第1章 绪论 1](#_Toc16464)

[1.1 研究背景及意义 1](#_Toc16536)

[1.2国内外研究现状 1](#_Toc15400)

[1.2.1国内研究现状 1](#_Toc12397)

[1.2.2国外研究现状 2](#_Toc18993)

[1.3 论文研究内容及目标 2](#_Toc22235)

[1.3.1 论文主要研究内容 2](#_Toc10734)

[1.3.2 论文研究目标 2](#_Toc28850)

[1.4 论文组织结构 2](#_Toc6878)

[第2章 关键技术分析 4](#_Toc9631)

[2.1 相似度计算 4](#_Toc7041)

[2.2协同过滤推荐算法 5](#_Toc8604)

[2.2.1基于用户的协同过滤推荐算法 5](#_Toc6520)

[2.2.2基于项目的协同过滤推荐算法 5](#_Toc3858)

[2.3推荐算法评价指标 6](#_Toc25512)

[2.4 Apache Mahout框架介绍 7](#_Toc7912)

[2.5 本章小结 9](#_Toc30354)

[第3章 电影智能推荐系统模型设计 10](#_Toc30216)

[3.1 相似度计算分析 10](#_Toc15413)

[3.2 特征向量提取 10](#_Toc21549)

[3.3 推荐模型设计 11](#_Toc25656)

[3.3.1 User-based的协同过滤算法推荐模型 11](#_Toc9713)

[3.3.2 Item-based的协同过滤算法推荐模型 13](#_Toc24109)

[3.4 推荐模型验证 14](#_Toc1136)

[3.5 本章小结 19](#_Toc3887)

[第4章 电影智能推荐系统模型实现 20](#_Toc30472)

[4.1系统整体设计 20](#_Toc25897)

[4.2 数据库设计及数据预处理 21](#_Toc12275)

[4.2.1数据库处理 21](#_Toc15743)

[4.2.2数据预处理 22](#_Toc718)

[4.3推荐模型实现 23](#_Toc508)

[4.3.1User-based CF推荐实现 24](#_Toc3151)

[4.3.2 Item-based CF推荐实现 24](#_Toc13871)

[4.4推荐结果 24](#_Toc31927)

[4.5本章小结 25](#_Toc16534)

[第5章 总结与展望 27](#_Toc17090)

[5.1论文工作总结 27](#_Toc22096)

[5.2未来展望 27](#_Toc803)

[参 考 文 献 28](#_Toc26826)

[致 谢 30](#_Toc30525)

**第1章 绪论**

1.1 研究背景及意义

在电力大数据的背景之下不断深挖和研究，怎么样从巨量的用户用单电数据中，准确精准地识别出并挖掘有用的信息，最终应用在电力大环境下是目前亟待解决的。基于对大数据环境下的用电数据特征关联信息，选取强有力的大数据分析方法，对电力用户进行十分合理的分类，并基于分类的结果，迅速、精准地挑选出用电行为、用电潜力等有价值的信息，发现用户用电规律等工作，实施在智能电网的大背景下，是目前智能电力相关的一个重大问题。对电力用户进行合理有效的分类，有利于电力公司指定针对性的营销策略以及合理的价格区间，有利于合理引导用户调整其用电模式，减少能源产出的过渡浪费以及提高机组的运行效率，从而提高能效，节约电力企业的运行成本[1]。因此，利用大数据相关算法以及技术对电力用户特性进行专业分析，可以帮助电力公司精准认知用户的差异化、个性化服务需求，对客户对电力企业的满意度的提升有着重要的意义。

随着大数据时代的来临，用于实现电网的安全、可靠、经济、高效、环境友好和使用安全目标的智能电网在快速发展和普及，电力公司对用户的电力负荷分析就显得尤为重要。但过去电力分析只是基于经验法，所采取的措施是拉闸限电等，不仅无法处理如今海量的大数据，也阻碍了智能电网的快速发展，因此需要更科学的分析法来解决当下的难题。

用户用电行为特征分析主要包括以下步骤。

( 1) 从用电采集数据源中进行选择性抽取历史用户用电功率数据和用户档案信息数据。

( 2) 对步骤( 1) 中的两个数据集进行数据关联、数据探索分析与预处理，包括数据缺失值与异常值的探索分析，数据的清洗和变换等。

( 3) 利用步骤( 2) 中形成的已完成数据预处理的建模数据，利用 K-Mean 聚类算法进行用户用电特征分群，对各个用户群进行特征分析，识别出各类型用户特征。

( 4) 针对模型结果得到的不同特征类型的用户，采取不同的营销手段，提供定制化服务。

聚类分析法是理想的多变量统计技术，能通过迭代分层原理对大数据进行筛选聚类，能快速建立起用户的电力负荷模型，帮助电力公司能通过模型训练更好的预测分析用户的用电行为，掌握用户的用电习惯，进一步对评估电力市场工序，指导上网电价制定具有重要意义。

**二、国内外研究现状**

自从上世纪80年代改革开放始，我国进入了高速发展阶段，电力需求极度旺盛，能源极度紧张，电力供电一度出现供不应求的状况，过去中国采取一刀切的拉闸限电等策错，极大影响了用户的日常用电。随着社会的发展，结合了数学模型的聚类算法成了大多数电力公司的选择，例如广东电网中山供电局的李京平等采用PCC聚类算法对电力数据进行分类，成功做出了用户用电负荷模型，对预测用户用电负荷以及电力调度提供了科学可靠的依据[2]。

目前聚类算法的基本分类包括：划分法，层次法，密度算法，模型算法等，能处理任意形状的高维数据，并提供了对于噪点数据的针对处理方式。1957年，Lloyd基于划分的思想首先提出了K-Means聚类算法，但其本身只能针对圆形数据进行聚类，而1990年，Kaufman和Rousseeuw分别提出了基于K-Means的算法PAM和CLARA。1996年Ester等人抛弃了距离的概念，基于密度提出了聚类算法DBSCAN，解决了K-means算法对圆形数据的依赖[3]。日趋完善的数学模型推动了电力大数据的应用。

为了提高聚类算法的速度，不少专家学者开始研究聚类算法的并行计算。文献提出了一种新的基于自适应 K-means 的分布式聚类算法，实现了聚类算法的并行计算，有效减少了数据的通信成本。文献和文献建立了多维度用电特征评价指标，采用优选策略提取负荷曲线的最佳特征集，实现了用户用电行为的聚类优选。文献和文献对并行化 K-means 算法进行了改进，解决了 K-means 并行化后仍存在的聚类初始中心难以确定、聚类结果不稳定等问题。用户用电数据随时间不断增长，呈海量趋势，大大增加了数据分析的时间和空间复杂度。为了避免电力数据维度灾难，学者们在降维方面也做了一些努力。文献利用核矩阵降低维度，保留核矩阵中值较大的值。但该方法需要提前确定核函数参数且运行时间较长。文献对于海量高维的负荷曲线采用了主成分分析的方法降维，得到了有效的聚类结果，但该方法中重要参数的选取的合理性没有得到验证，还有一定的改进空间。

1.3 论文研究内容及目标

1.3.1 论文主要研究内容

* 研究聚类算法的原理、应用和发展现状；
* 研究基于用户行为数据的聚类算法的原理；
* 基于负荷用电数据公开数据集，利用聚类算法，分析用户用电行为习惯。

1.3.2 论文研究目标

* 了解和学习聚类算法的基本理论和方法；
* 基于k-means算法实现对用户的用电行为分析；
* 对模型进行深入的分析与训练，不断完善聚类结果；
* 结合学习和开发的内容完成毕业设计论文。

1.4 论文组织结构

本文分为五个章节，论文的组织结构如下：

第一章是绪论，主要介绍了本次研究的社会历史背景以及国内外研究现状的探讨以及在相同问题下的探究进度。主要介绍了目前参考文献中的部分解决示例，并基于此提出自己的通用解决方案，以符合论文的要求，为后续提供一个大的导向。

第二章为本次毕业设计所需要的一些基本技术的介绍，我在这一章做了关于聚类算法，K-Means算法的基本原理以及K-Means算法模型评价指标的介绍。所应用的语言和平台也是离不开的，因此还加入了针对python以及为何使用python进行科学技术运算的脚本语言的相关原因。

第三章为关于电力设计的相关实施细节，包括针对数据公开集的数据预处理以及具体python语言 sklearn包的运用的掌握。本章以需求分析开头，是希望自己的论文能扎根于实际之中，希望能摆脱纯理论的知识，使最终结果能使用在现实情况之中。

第四章为基于用户电力数据的程序设计过程，包括设计步骤，设计过程，设计遇到的问题等等。本章对一般情况下的通用聚类算法在电力的实际应用过程时应采取的优化以及怎样具体应用的思考，并验证聚类算法的最终聚类结果是否与实际情况相似。

第五章为总结与展望，总结了本论文的相关内容，包括相关技术，原理，以及展望。展望部分主要分析目前方法的不足以及是否有更佳的聚类方法的思考。

**第2章 关键技术分析**

2.1 聚类算法简介

在大数据的背景下，海量的数据并不具有明显的标签，因此分类起来相当麻烦，聚类算法应运而生。聚类算法就是按照一个标准，将一堆看不出规律的数据进行分割，分出不同的类或簇，使得一个类或簇中的数据非常相似，同时，在不同的类或簇中的数据差异性大。聚类算法包括：划分法、密度算法、图论聚类法、网格算法、模型算法等。例如K-means算法，属于聚类算法中划分方法的一种。算法中每个中心点称为质心，而在划分过程中，首先要确定质心的个数。然后通过不断地迭代计算，直到簇内的数据相对稳定，因此簇内的数据之间有较大的相似度，而不同簇的数据之间的相似度较低。

介绍聚类算法中较常见的几个聚类算法：

1) K均值聚类算法步骤：

(1) 首先我们选择一些类/组，并随机初始化它们各自的中心点。中心点是与每个数据点向量长度相同的位置。这需要我们提前预知类的数量(即中心点的数量)。

(2) 计算每个数据点到中心点的距离，数据点距离哪个中心点最近就划分到哪一类中。

(3) 计算每一类中中心点作为新的中心点。

(4) 重复以上步骤，直到每一类中心在每次迭代后变化不大为止。也可以多次随机初始化中心点，然后选择运行结果最好的一个。

2) 均值漂移聚类是基于滑动窗口的算法，来找到数据点的密集区域。这是一个基于质心的算法，通过将中心点的候选点更新为滑动窗口内点的均值来完成，来定位每个组/类的中心点。然后对这些候选窗口进行相似窗口进行去除，最终形成中心点集及相应的分组。

具体步骤为：

具体步骤：

(1). 确定滑动窗口半径r，以随机选取的中心点C半径为r的圆形滑动窗口开始滑动。均值漂移类似一种爬山算法，在每一次迭代中向密度更高的区域移动，直到收敛。

(2). 每一次滑动到新的区域，计算滑动窗口内的均值来作为中心点，滑动窗口内的点的数量为窗口内的密度。在每一次移动中，窗口会想密度更高的区域移动。

(3). 移动窗口，计算窗口内的中心点以及窗口内的密度，知道没有方向在窗口内可以容纳更多的点，即一直移动到圆内密度不再增加为止。

(4). 步骤一到三会产生很多个滑动窗口，当多个滑动窗口重叠时，保留包含最多点的窗口，然后根据数据点所在的滑动窗口进行聚类。

3) 基于密度的聚类方法(DBSCAN) 与均值漂移聚类类似，也是基于密度的聚类算法。

具体步骤：

(1). 首先确定半径r和minPoints. 从一个没有被访问过的任意数据点开始，以这个点为中心，r为半径的圆内包含的点的数量是否大于或等于minPoints，如果大于或等于minPoints则改点被标记为central point,反之则会被标记为noise point。

(2). 重复1的步骤，如果一个noise point存在于某个central point为半径的圆内，则这个点被标记为边缘点，反之仍为noise point。重复步骤1，知道所有的点都被访问过。

4) 用高斯混合模型（GMM）的最大期望（EM）聚类。

使用高斯混合模型（GMM）做聚类首先假设数据点是呈高斯分布的，相对应K-Means假设数据点是圆形的，高斯分布（椭圆形）给出了更多的可能性。我们有两个参数来描述簇的形状：均值和标准差。所以这些簇可以采取任何形状的椭圆形，因为在x，y方向上都有标准差。因此，每个高斯分布被分配给单个簇。

具体步骤：

(1). 选择簇的数量（与K-Means类似）并随机初始化每个簇的高斯分布参数（均值和方差）。也可以先观察数据给出一个相对精确的均值和方差。

(2). 给定每个簇的高斯分布，计算每个数据点属于每个簇的概率。一个点越靠近高斯分布的中心就越可能属于该簇。

(3). 基于这些概率我们计算高斯分布参数使得数据点的概率最大化，可以使用数据点概率的加权来计算这些新的参数，权重就是数据点属于该簇的概率。

(4). 重复迭代2和3直到在迭代中的变化不大。

5) 凝聚层次聚类算法分为两类：自上而下和自下而上。凝聚层级聚类(HAC)是自下而上的一种聚类算法。HAC首先将每个数据点视为一个单一的簇，然后计算所有簇之间的距离来合并簇，直到所有的簇聚合成为一个簇为止。

具体步骤：

(1). 首先我们将每个数据点视为一个单一的簇，然后选择一个测量两个簇之间距离的度量标准。例如我们使用average linkage作为标准，它将两个簇之间的距离定义为第一个簇中的数据点与第二个簇中的数据点之间的平均距离。

(2). 在每次迭代中，我们将两个具有最小average linkage的簇合并成为一个簇。

(3). 重复步骤2知道所有的数据点合并成一个簇，然后选择我们需要多少个簇。

本次论文从实际与应用角度出发，将各算法的复杂程度与实际应用相结合，最终选择了易上手且针对电力大数据易操作的K均值聚类算法作为本次论文对电力用户用电分析模型建立的最终算法。

2.2 K均值聚类算法

K-Means是较为简单的聚类算法，其思想是首先对给定的样本集，分成K个簇，让簇内的点紧密排列，不同簇之间距离鲜明。

其计算公式:

其中μi是簇Ci的均值向量，有时也称为质心，表达式为：

假设输入是样本集D={x1,x2,...xm},聚类的簇树k,最大迭代次数N

假设输出是簇划分C={C1,C2,...Ck}

算法流程：

1. 从数据集D中随机选择k个样本作为初始的k个质心向量： {μ1,μ2,...,μk}
2. 对于n=1,2,...,N
3. 将簇划分C初始化为Ct=∅ t=1,2...k
4. 对于i=1,2...m,计算样本xi和各个质心向量μj(j=1,2,...k)的距离：dij=||xi−μj||22，将xi标记最小的为dij所对应的类别λi。此时更新Cλi=Cλi∪{xi}
5. 对于j=1,2,...,k,对Cj中所有的样本点重新计算新的质心
6. 如果所有的k个质心向量都没有发生变化，则转到步骤3）
7. 输出簇划分C={C1,C2,...Ck}

K-means 算法存在不少的缺点，比如质心的个数需要提前指定，而具体的数据集初始的质心个数往往能难判断。初始质心的具体数值的选取往往很大程度的决定了迭代后收敛的最终结果，因此不同的单次聚类方法其结果会有很大的不同。针对于噪点数据，即不正常的数据或离群的数据，会很大程度的影响聚类的效果。

2.3 K均值算法评估指标

K-Means算法的性能因为以上缺点需要对聚类的效果进行性能评估。评估的方法主要包括：SSE误差平方和、SC系数(Silhouette Cofficient)轮廓系数法、CH系数(Calinski Harabasz Index)轮廓系数法。本次论文主要采用SSE误差平方和的方法进行不同k值得初始化评估。

SSE误差平方和简单来说即是计算当前迭代得到的中心位置到各自中心点簇的欧式距离总和，因此值越小代表当前的聚类效果越好。

其计算公式为：

其中P代表点位置，Mi为中心点的位置。

SC系数(Silhouette Cofficient)轮廓系数法结合了聚类的凝聚度(Cohesion)和分离度(Separation),用于评估聚类算法的效果。

其计算公式为：

a表示C1簇中的某一个样本点Xi到自身簇中其他样本点的距离总和的平均值。bC2表示样本点Xi 到C2簇中所有样本点的距离总和的平均值。bC3表示样本点Xi 到C3簇中所有样本点的距离总和的平均值。

本次论文主要采用SSE误差平方和来对聚类算法进行一个简单的结果评估，来辅助判断k值得最终应该选取的结果。

2.4 Python 语言特性介绍

Python 是一个高层次的结合了解释性、编译性、互动性和面向对象的脚本语言。Python 是由 Guido van Rossum 在八十年代末和九十年代初，在荷兰国家数学和计算机科学研究所设计出来的。

Python 本身也是由诸多其他语言发展而来的,这包括 ABC、Modula-3、C、C++、Algol-68、SmallTalk、Unix shell 和其他的脚本语言等等。

由于Python语言的简洁性、易读性以及可扩展性，在国外用Python做科学计算的研究机构日益增多，一些知名大学已经采用Python来教授程序设计课程。例如卡耐基梅隆大学的编程基础、麻省理工学院的计算机科学及编程导论就使用Python语言讲授。众多开源的科学计算软件包都提供了Python的调用接口，例如著名的计算机视觉库OpenCV、三维可视化库VTK、医学图像处理库ITK。而Python专用的科学计算扩展库就更多了，例如如下3个十分经典的科学计算扩展库：NumPy、SciPy和matplotlib，它们分别为Python提供了快速数组处理、数值运算以及绘图功能。因此Python语言及其众多的扩展库所构成的开发环境十分适合工程技术、科研人员处理实验数据、制作图表，甚至开发科学计算应用程序。

像 Perl 语言一样，Python 源代码同样遵循 GPL(GNU General Public License)协议。Python是一种解释型，交互式，面向对象的语言。且Python拥有pip包管理工具，包含大量的关于聚类算法的开源代码。本次论文即是通过python的sklearn包的cluster类进行聚类算法的处理。

2.5 Python 中Sklearn拓展包特性介绍

官网上列举了Sklearn的四大特性：

1. 简单有效的数据挖掘和数据分析工具；
2. 可供所有人访问，并且可以在各种环境中重复使用；
3. 基于NumPy，SciPy和matplotlib构建；
4. 开源，拥有商业可用的BSD许可证。

Scikit-learn(sklearn)是机器学习中常用的第三方模块，对常用的机器学习方法进行了封装，包括回归(Regression)、降维(Dimensionality Reduction)、分类(Classfication)、聚类(Clustering)等方法。

sklearn是Scipy的扩展，建立在NumPy和matplotlib库的基础上。利用这几大模块的优势，可以大大提高机器学习的效率。sklearn拥有着完善的文档，上手容易，具有着丰富的API，在学术界颇受欢迎。sklearn已经封装了大量的机器学习算法，包括LIBSVM和LIBINEAR。同时sklearn内置了大量数据集，节省了获取和整理数据集的时间。

2.6 开发相关环境

2.6.1 vscode开发环境

Visual Studio Code是一个轻量级但功能强大的源代码编辑器，可在桌面上运行，适用于Windows，macOS和Linux。它内置了对JavaScript，TypeScript和Node.js的支持，并具有丰富的其他语言（如C ++，C＃，Java，Python，PHP，Go）和运行时（如.NET和Unity）的扩展生态系统。

Vscode的宣传语是：一个运行于 Mac OS X、Windows和 Linux 之上的，针对于编写现代 Web 和云应用的跨平台源代码编辑器。

其特性基本为：

1. Free但不开源（目前已经开源）
2. Build（构建）和 debug（调试） 现代web和云应用(尤其是JavaScript、TypeScript、C#、ASP.NET v5 和 Nodejs)
3. 跨平台支持Linux, Mac OSX, and Windows
4. 支持语法自动补全，智能提示
5. 内置html开发神器emmet

6) 速度、调试效率非常快

7) 支持多主题（配色方案）0.9.1之后更爽

8) 对于其他编辑器的高效操作和快捷键都有继承

9) 插件系统

2.6.2 csv文件介绍

逗号分隔值（Comma-Separated Values，CSV，有时也称为字符分隔值，因为分隔字符也可以不是逗号），其文件以纯文本形式存储表格数据（数字和文本）。纯文本意味着该文件是一个字符序列，不含必须像二进制数字那样被解读的数据。CSV文件由任意数目的记录组成，记录间以某种换行符分隔；每条记录由字段组成，字段间的分隔符是其它字符或字符串，最常见的是逗号或制表符。通常，所有记录都有完全相同的字段序列。通常都是纯文本文件。

可以总结，CSV为特殊的纯文本文件，因为其文件中单个数据分隔符的问题，而python语言可以完美的读取CSV文件并且可以将其转化为自己所创的DATAFRAME类并可以进行一系列针对性操作，这也是选取python进行数据处理和算法实现的关键理由之一。

**第3章 基于聚类算法的负荷用电模型分析**

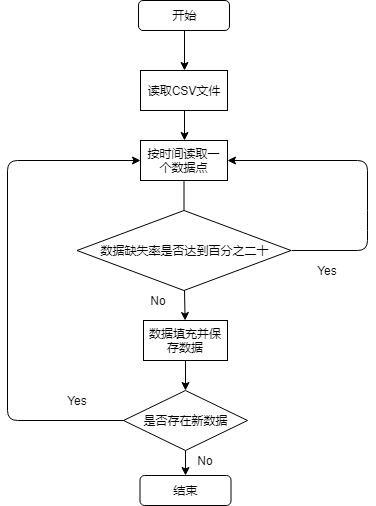
3.1 基于聚类算法的用户用电行为需求分析

目前用电公司存在两大难题，一是用电高峰的错误预估导致用电高峰期只能进行拉闸限电等原始措施，二是用电低谷期电机空转导致的电能浪费。目前，中国正处于风口浪尖之上，各国纷纷指责中国的环境污染问题，因此在如此严峻的形势下，电力公司对用户用电的预测不仅仅只帮助其减少公司运维成本，增加公司的利益，指导其合理制定电价，更是对环境保护的一种极其重要的措施。

如果能通过对用户用电进行模型建立，准确预测用户的用电高峰期以及用户的具体模型，则可以在用电低谷存储电量或停止电机运转，在用电高峰进行提前预估，还可以针对模型出台相关收费措施，合理引导用户的用电行为，从而避免高峰期的频繁触发，降低高峰期的峰值，减轻用电公司的负担。

而普通的分类方法无法处理数量级于百万甚至千万以及亿级的数量级，因此需要利用聚类算法对用户数据进行合理的聚类，得出预测分析模型，从而有利于用电公司的发展。聚类算法本身就是在大数据的环境下诞生的，是针对高维、极大量、分布规律极其复杂的数据进行运算，分类的一种处理算法，因此十分符合电力公司对用户用电数据模型的建立。

随着计算机相关技术的发展，一款语言因其上手简单、易读易维护且拥有丰富的拓展库逐渐被各个大公司用来开发大规模的项目，世界上比较著名的几个人工智能框架，例如TensorFlow、Caffe、MXNet等均由python所写或者提供了基于python的开放api，所以考虑到python拥有大量的开源智能算法集，且对csv数据文件的处理非常成熟，因此可利用python脚本，调用sklearn包进行数据的预处理以及聚类运算，并将结果进行评估。

数据处理的流程图如下：

3-1 数据处理流程图

有了数据模型，既可以通过通用的算法对数据集进行聚类，再针对行业特点进行算法的优化，则最终即可满足用户用电模型的建立要求。

考虑到电力行业数据的特点，在基本可预测的问题，例如数据缺失，数据噪点，数据作假等进行针对性处理，考虑到实际操作难度，本次论文只针对数据缺失进行填充或删除，不进行更进一步的优化处理。

在算法优化的基础上，采用随机选取等距离的数据点做初始化处理，并选取多个K值做算法结果评估，因而比普通的聚类算法所得到的数据模型更加精确，符合行业的标准。

因目标客户是电力企业或相关公司，鉴于其本身行业特性，图标能直观的显示出函数的特性的特点，因此数据的结果应以函数的形式显示在图表中以代表用户的典型特征，方便电力相关公司发现其规律并基于此规律指定相关的策略方针。

3.2 数据预处理

1）为什么要预处理数据。因为大数据具有四个特点，分别为：Volume（大量）、Variety（多样）、Velocity（高速）、Value（价值），其中数据的多样性决定了数据拥有大量的无用属性以及属性缺失，因此需要针对数据进行具体的预处理，使其能满足统一算法的需求，本次论文采用的是西班牙一百多家用户2014-2016年的公开数据集，数据格式为：时间点、用电量，时间间隔为一分钟或十五分钟，且数据有缺失，因此要进行统一化处理才能把数据代入算法进行统一化处理。

2）数据点集的数据选取。因为聚类算法是将数据进行聚类，因此数据的格式至关重要，不同的数据选取会导致不同的聚类结果，要通过目的去选取数据模型，本次论文要研究基于负荷用电数据公开数据集，利用聚类算法，分析用户用电行为习惯，因此选取单个用户的每15分钟的用电数为坐标，将一天具象化为96维的特征向量，进行聚类处理。

3）缺失数据的填充。对于缺失率达到百分之二十的数据，进行简单的剔除处理。对于缺失率不到百分之二十的数据，若有坐标缺失，将相邻的数据坐标取均值并插入，若相邻左边依然缺失，则继续向两边查找，直至满足条件。针对于异常数据，因为数据集不含有负值，因此不做正负值处理。

3.3 模型大致设计

3.3.1 设计步骤

1）首先编写脚本读取所有的csv数据集文件，利用相关规则进行数据统一化处理以及缺失数据的填充，并将结果存放在固定的文件中，因为涉及到大量文件的读取，因此采用流式方式对文件进行读取，降低内存的使用率，并且将结果存在单个文件中，方便后续的数据读取。

2）本次论文采取的K均值算法对普通的K均值算法进行少许优化，首先在初始质心的值选取时，将初始质心的值选取尽可能的分散，依然采取随机方式选取，但每选取一个质心，不同位置的数据点即拥有不同概率的选取率，其概率决定于数据点距离已选质心的距离。其次选取多个K值进行聚类分析，并进行聚类结果评估，分别求取其SSE误差平方和并进行比较，因为K值越大，其SSE误差平方和越小，因此选取离后续节点差距较小的K值进行最终的聚类初始质心的个数选择。

3）将上述结果所得到的K个质心数据进行处理，并将其通过前端程序呈现在网页上，并对最终的结果进行分析，并与最初的预测结果和实际情况进行对比，从而判断最终的实现结果是否符合规律。

3.3.2 算法实现

根据前文原理以及数据处理的内容，首先将处理后的内容（即CSV文件）作为输入文件，其单条数据即为数据点，其中包含96个坐标点，是代表一天的96维的特征向量。依据python的numpy处理包，可以将处理后的CSV文件转化为Dataframe类数据，并提供一系列api去对数据进行处理，以便后续操作。

依据sklearn包的cluster类方法，并赋予聚类的结果个数参数n\_clusters，其值分别从1至8，其处理过程大致是：

1. 依据n\_clusters参数选定相关个初始质心，初始质心的选取为数据的前K个点，这里并未针对初始质心的选取进行优化。
2. 根据欧几里得计算公式，计算出不同的96维特征向量距离质心的欧氏距离从而判断其所属的质心。
3. 将所聚之类的点的各个维度值相加取平均值，作为其簇的新的质心点，并重复步骤2、3，若质心未改变，则聚类计算结束，导出最终质心坐标。

从上述步骤我们会获得K个96维的特征向量，其中每个向量所代表的是同一类型的用户用电典型模型，我们可以将结果画在曲线图中从而直观的判断15、16两年一百多户西班牙家庭的每天的用电情况所属的不同聚类结果，并进而分析其与当地现实情况的对比与更进一步的分析，从而判断算法最终是否成功，因为聚类的结果模型要应用于现实的电力公司或者电力企业，因此将理论的结果与实际进行结合分析是十分必要的。

另外，我们需要将结果进行评估，因此在不断迭代计算质心位置时，要注意SSE误差平方和公式的计算，所以要在最后计算每个点到其所属的簇的质心的距离的平方，并进行累加从而得到最终的结果，用于对最终的聚类结果进行评估。

**第4章 聚类结果的分析评估**

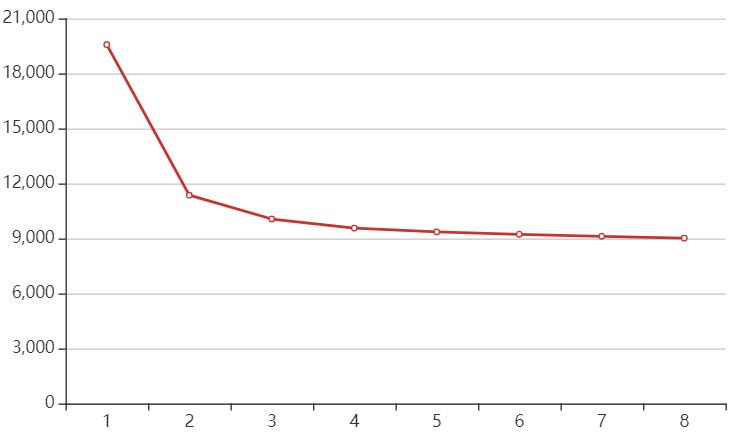
4.1 算法角度

分别利用SSE误差平方和公式对不同k初值的方法进行计算，其相关数据如下。

|  |  |
| --- | --- |
| 初始质心个数 | SSE误差平方和 |
| 1 | 1957271419.168 |
| 2 | 1136369924.831 |
| 3 | 1006366055.242 |
| 4 | 956838108.120 |
| 5 | 935939365.350 |
| 6 | 922958839.425 |
| 7 | 912031715.624 |
| 8 | 902326630.293 |

表4-1 初始质心与SSE误差平方和的对应关系

SSE误差平方和之所以能够检验聚类效果是因为其值越小，则代表每个数据点距离其所属的簇质心的距离越少，从而代表其簇内的点越聚集，因此简介验证了其聚类效果的好坏，因本次论文的数据格式均为96维度的特征向量，因此不能通过二维或者三维的坐标系直观的表现出来，所以通过SSE误差平方和的方式来代表最终结果是最为合适的。表格的形式难以看到规律，因此提供图像版本探讨并选取最终的初始质心个数，结果如下图：

图4-1初始质心与SSE误差平方和的对应关系

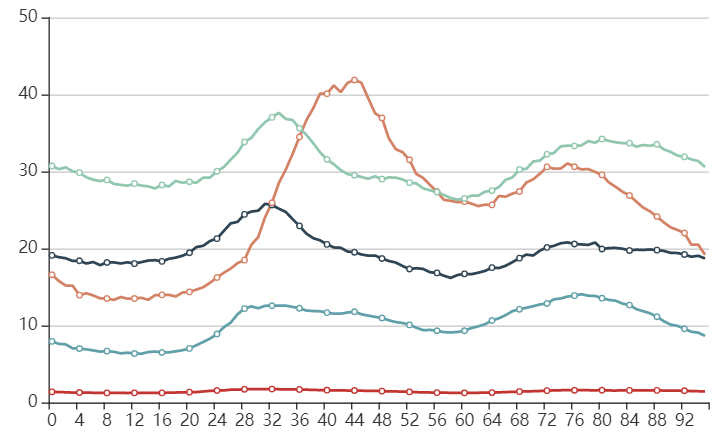
鉴于K=3时初始质心的个数较少，且K值过大时的SSE误差平方和差距较小，因此选取K=5时的聚类结果作为最终的聚类结果的初始质心个数，聚类的质心结果如下图：

图4-2 K=5时聚类结果的质心所对应的的曲线图

用电情况非常直观的显现在图内，在此先不对数据进行分析，先从初始质心的选取对最终情况的结果进行对比，以此来验证初始质心个数对聚类结果的影响，因此需要不同初始质心个数曲线图用来对比。作为对比，在此插入K=3以及K=8时的结果进行对比，结果图分别如下：

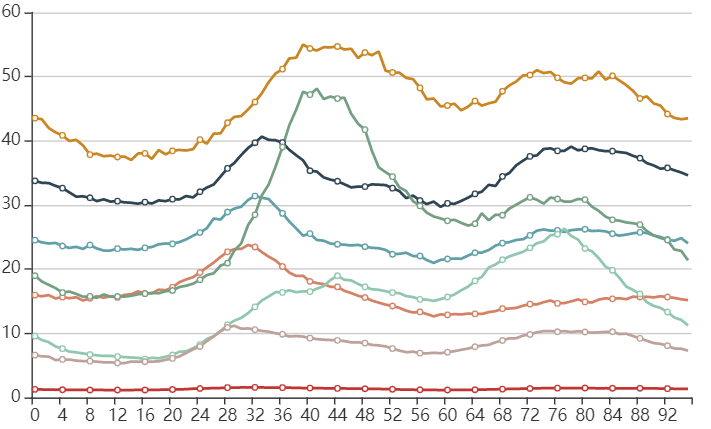
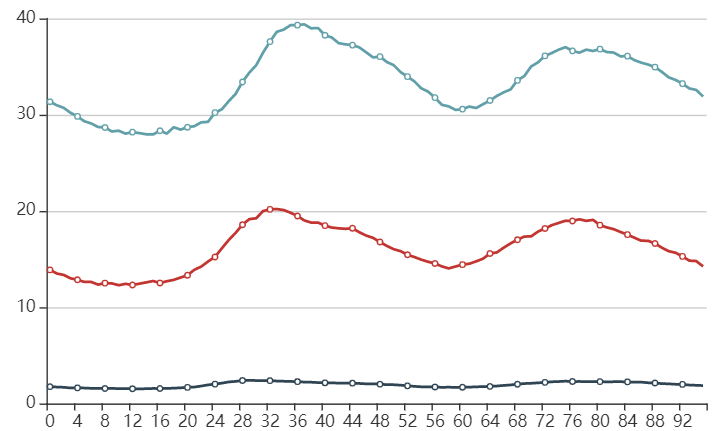
图4-3 K=3时聚类结果的质心所对应的的曲线图

图 4-4 K=8时聚类结果的质心所对应的的曲线图

从图上对比，可以发现K=3时并不能很好的代表数据的聚类，即使从数学角度也无法将其结果作为最终的聚类结果，而K=8时聚类结果虽更加明显，但依据SSE误差平常和，其差距和K=5时的结果相差较小，因此K=5可以代表用电用户的典型特征。

贴图的目的只是从算法角度说明同一K均值聚类算法，其初始质心的个数不同，对结果的影响非常大，一般在选取初始质心的个数时，应选取不同的初始质心代入运算，将结果进行评估处理，例如最简单的SSE误差平方和算法，即将各个簇内的数据点与簇心的误差平方和进行相加，得到结果评估，如前文所述。

从算法角度分析，K=5时五类数据大致可以分成两大类，其中一类为贴近X轴的函数，另外四类都是呈现先增后减再增再减的趋势。

4.2 实际角度

若从实际角度进行分析，则根据图4-2的结果表明，几乎所有的用户都会出现早晚双高峰的情况，其中三类的区别是并不大的，有两类是特殊情况，一类是贴近X轴的曲线，全天基本没有什么用电量，说明这种类型的用户基本不怎么在这个地方住，这其实是与显示情况相符合的，因为有很多的用户有多个住处，而一般情况下一个家庭住在一个住处，因此其他住处的用电量基本非常少，可以忽略，只有少部分例如热水器等电器还在工作，因此全天基本用电量少的可怜。

而三类相似的用电曲线也符合一般规律，因为早上班晚回归是几乎所有家庭上班族的生活常态，唯一的区别是不同的人在家时有不同的生活方式，例如单身青年租房不存在大规模的用电，因此其曲线是相对较低的，而例如一些生活家具齐全，全家五口一起生活的家庭其用电量及用电规模是比上述单身青年用电多很多的，不过这几种结果都符合早晚双高峰的规律，因此可以判断当前的聚类结果是符合现实情况的，从实际角度出发，每条曲线都可以找到一类用户进行验证并得到最终确认，是符合科学规律的。

4.3 结果分析

依据图4-2聚类的结果，可以基本得出样本地区早晚会出现高峰用电情况，且早高峰可能更会出现“用电阻塞”现象，因此需要电力公司针对早上用电进行合理引导，指定相关政策，例如增加早7至9点的用电费用、鼓励用户将一些预备用电例如热水器用电的时间设置到凌晨即预热以及考虑其他时间段用电的奖励政策从而将早晚双高峰用电合理引导至其他时间段，并能通过用电模型合理的预测以及对高峰用电的预备措施来降低“用电阻塞”情况的发生。

首先针对全体用户，我们应该首先秉承不浪费资源的原则，倡导用户停止不必要的用电行为，再通过不同的用户特征曲线进行相应的政策处理，从而达到电力行业对环境以及盈利的双赢。

针对贴近X轴的用户，我们应采取奖励的措施，号召其他组用户向其学习，例如针对不同的时间点，用户用电量相比同时间段降低即对其采取降低费用处理，从而合理引导用户自主降低峰值用电。针对早晚双高峰特征曲线的三类用户，我们需要采取固定时间段的用电奖励，例如凌晨或者中午时间段用电的费用减免，引导其一天用电曲线的扁平化，从而降低用电高峰指数，缓解企业的发电压力，并避免电机空转所带来的经济以及资源的浪费。针对早高高晚低高的用户，首先要对其早用电进行惩罚性措施，例如强制用电限制、提高早时间段收费等行为，要强制降低其早用电，另外开源节流，将其不可避免的用电行为合理引导至晚高峰或更合理的时间段，因此也需要降低其他时间段的用电收费辅以配合，才有可能使其用电更加扁平化。

从上述三个方面可以总结，企业在拿到用户用电模型后，针对用电高峰进行合理限制，针对用电低峰进行合理引流，能够达到经济和资源的双盈利。当然，其大前提是用户的用电模型必须非常精准，而聚类方法中的K均值算法经过验证，其结果与实际相对比较符合，因此可以用来给与电力公司来参考，从而指定合理的相关政策。

**第5章 总结与展望**

5.1论文工作总结

本课题的主要内容是基于聚类的用户用电行为分析，本次课题开篇讲述了电力大数据问题在国内外的研究现状以及背景介绍，分析了电力大数据处理的困难以及正确处理的意义。之后分析了所用到的相关算法、部分思想以及相关平台，其中包含了聚类算法的介绍、K均值算法的介绍以及初步优化和算法评估并阐述了电力数据的处理方式以及其被处理前后的意义，最后是针对所涉及的python编程语言以及vscode编辑器的介绍与使用原因。之后则是具体的算法处理以及结果分析和结果评估。论文总体分为以下几个方面：

（1）第一章是本篇论文的背景与现状的介绍，只有真正了解了相关算法在实际应用的适用以及进展才能站在巨人的肩膀上，从而更快更精准的建立模型，否则自己埋头研发，不问世事，最终只能是一纸空谈，不仅进度会非常慢，而且其实用性也会受到质疑。只有充分了解了相关问题的具体应用，才可以有理可依，从而优化具体算法在相关领域的应用，达到最终目的。

（2）在论文的第二章主要介绍了基于聚类的用户用电行为分析过程所需要的关键技术。首先阐述聚类以及本次课题所采取的K均值算法的计算过程、结果评估。接着开始阐述所使用的程序语言python的相关特性以及所涉及到的相关文件包，其中阐述了其特点以及使用原因。最后，对所使用的平台做了相关介绍，vscode是一款强大的编辑器，适用于很多语言且拥有丰富的拓展，因此在这里使用了vscode这款编辑器。

（3）首先根据需求进行分析，分析包括数据合理化处理，算法所使用的具体实例，然后介绍了数据的具体预处理过程，包括缺失数据以及错误数据的处理。接下来就具体实现步骤与算法细节进行介绍，具体步骤主要介绍了源代码的几部分，其中包括预处理程序以及主程序的设计，算法细节主要介绍了一般性算法在本次课题设计的相关细节实施。

（4）主要通过结果首先对不同聚类结果进行算法评估，挑选出误差最小的聚类结果进行实际对比，包括算法角度以及实际角度的两方面对比，从而探讨聚类结果的实际意义并通过两方面的分析给出针对性的建议策略从而使得聚类结果与实际更加贴合，进一步加强聚类结果与实际的粘合程度，因为本次课题实质是用户的用电行为分析，若不与实际相结合，那么聚类本身不存在任何意义。

5.2未来展望

本次的数据来源是西班牙一百多户家庭2015至2016年两年的用电数据公开集合，所使用的算法是最基本也是最简单的K均值算法，所以存在以下几个问题：

1. 初始质心的选择并没有进行优化，理论上应该选取尽可能远的点进行迭代，每个质心与其他质心的距离应保持足够远。这样迭代的最终结果才会比较均匀，迭代次数相对较少，迭代的结果才能更加精确。
2. 并未对最终结果内每个簇中的个数进行统计，无法排除样本数量较小的簇，或对样本过大、分散程度较大的簇进行判断和二次聚类。
3. 向量机中核函数的思想，将所有样本映射到另外一个特征空间中再进行聚类，就有可能改善聚类效果。本次课题只是简单的对数据进行简单的处理以及转换，并未将其转换到另外一个聚类空间。

参 考 文 献

[1]李翠平.大数据与推荐系统[J].中国人名大学数据工程与知识工程教育部重点实验室,2015,1(3):23-35

[2]韩春阳.基于标签的个性化电影推荐算法研究与实现[D].北京邮电大学,2017

[3]俞美华.融合电影兴趣度与项目相关度的电影推荐算法研究[J].电脑知识与技术,2017,13(8):22-26

[4]兰艳,曹芳芳.面向电影推荐的时间加权协同过滤算法的研究[J].计算机科学,2017,44(4):295-301

[5]刘佳玮.基于电影系统的协同过滤算法的改进[J].电子技术与软件工程,2018,(01):181-182

[6]闫燕.基于协同过滤算法的电影推荐应用研究[D].河北大学,2014

[7]周魏，文俊浩.推荐系统中基于目标项目分析的托攻击检测研究[D].重庆大学出版社,2017 14-15

[8]赵雪.基于用户兴趣的个性化协同过滤推荐算法研究[D].河北:燕山大学,2014

[9]弗朗西斯科.里奇等编.李艳民.胡聪等译.推荐系统技术、评估及高效算法.机械工业出版社,2015 186-188

[10]李爱梅.融合环境信息的协同过滤推荐算法[D].燕山大学硕士论文,2012

[11]蔡观洋.个性化推荐中协同过滤算法的改进研究[D].吉林：吉林大学硕士论文,2013

[12]高献卫,师智斌. 基于Mahout的新用户推荐算法的设计与实现[J]. 计算机工程与科学,2015,37(08):1444-1449

[13]甘鹤.基于攻击识别的鲁棒推荐算法研究[D].燕山大学,2016

[14]袁义龙. 基于Mahout的电影推荐引擎的设计与实现[D].北京：北京邮电大学,2014

[15]李媛.基于位置的社交网络推荐算法的研究与应用[D].中国科学院研究生院（沈阳计算技术研究所):硕士论文,2015

[16]陆嘉恒.Hadoop实践.机械工业出版社,2013 304-307

[17]Sean Owen,Robin Anil著，王斌，韩冀中译.Mahout实战.人民邮电出版社,2014 10-57

[18]冷亚军,陆青,梁昌勇.协同过滤推荐技术综述.CSDN：模式识别与人工智能,2014,08 720-734

[19]汪然.基于图像内容的隐写分析技术研究[D].解放军信息工程大学博士论文,2014

[20]高晶晶.基于需求响应的电动汽车充电智能推荐的应用研究[D].北京：华北电力大学,2016 22-23

[21]李小川.大数据数据处理分析技术的研究[EB/OL].互联网论文库,2015

[22]李雪.云健康[EB/OL].互联网论文库,2016

[23]张勇.时间序列模式匹配技术研究[D].华中科技大学博士论文,2012

[24]张权.基于协同过滤家装方案推荐算法的研究与应用[D]..上海:上海交通大学,2015

[25]孔维梁. 协同过滤推荐系统关键问题研究[D].武汉：华中师范大学,2013

[26]冯婧禹.泛在学习环境下学习资源推荐系统的研究与设计[D].北京:北京交通大学,2015

[27]常江. 基于Apache Mahout的推荐算法的研究与实现[D].成都：电子科技大学,2013.

[28]曾志浩,张琼林,姚贝,孙琪. 基于Mahout分布式协同过滤推荐算法分析与实现[J]. 计算技术与自动化,2015,34(03):67-72

[29]王伟.基于资源特征的协同过滤推荐算法研究[D].长沙：中南大学,2012

[30]高拓.基于Mahout的分布式视频推荐系统的研究与实现[D].大连理工大学硕士论文,2014

[31]贾忠涛.基于协同过滤算法的电影个性化推荐系统设计与实现[J].软件导论,2015,14(01):86-88

致 谢

时间不知不觉得过去了，我们也不知不觉得大学毕业了。我们从小就期盼的大学四年生涯就这么慌慌张张的走到了尽头。借此，感谢大学四年来所有帮助过我的老师、同学、学长学姐们。

首先感谢我大学四年的班主任，齐林海老师，同时齐老师也是我这次的毕业论文辅导老师。谢谢齐老师在大学四年来对我们的教诲，对我们提供的帮助。在我们遇到问题时老师总是能够及时的给予帮助，引导着我们一步步往前走。齐老师在这次的毕业论文中也给了我专业的指导和重要性的建议。

其次要感谢我的舍友们，大学相处了四年，多谢你们给了我很多的包容和友爱。我会记得我们最后一个学期伴随着毕业论文一起进步的过程。

最后感谢华北电力大学，我的母校，谢谢你这四年来对我的大力栽培。同时向所有论文审阅组的老师感谢。