16级电气工程及其自动化专业毕业设计中期报告

基于因子隐马尔科夫模型的非侵入式负荷分解

学生：刘晋松 16级电气工程及其自动化专业3班，1605020316

指导教师：秦川

1. 研究背景与意义[1,2]

人类社会的发展伴随着能源消耗的增长，经济的发展又与能源消费息息相关。当前社会进步的同时，能源短缺、环境问题也随之越来越突出，成为制约经济的一个重要因素，如何节约地、高效地利用能源，寻找清洁的可再生能源成为社会发展的一个重要议题受到人们的广泛重视。而电能作为一种清洁、便利、高效能源的载体，在终端能源消费中占据着不可替代的重要地位，所以如何经济、有效地利用电能对节约能源有着重要意义。

电力需求侧管理可以使电能得到更经济的利用，它与高级量测体系同属于智能电网的关键技术。电网的智能化需要供电机构通过监测精确得知用户的用电规律，从而对需求和供应有一个更好的平衡。负荷监测作为高级量测体系重要的组成部分之一，是实现智能电网的第一步。

非侵入式负荷分解是负荷监测的一种途径，通过在用户电力入口处安装监测设备，对电气量进行采样、分析，得出每个电器在一段时间内的运行状态情况和耗能情况。用户便可以对自己的用电行为做出调整优化，合理安排用电设备的使用时间和使用方法，提高用户自主节能的意识，降低电能消耗。同时相较于侵入式负荷监测方法，非侵入式负荷分解具有成本低、用户更易接受、更易于实现的优点。而对于电网公司来说对用户的负荷分解有助于其制定科学的动态电价和需求响应激励政策。

1. 国内外文献综述与分析

2.1 负荷特征和数据集的选取[1,2]

负荷特征总的来说分为暂态特征和稳态特征。电气特征量的选择和提取是实现负荷分解的基础和前提，利用不同的负荷特征得到的负荷分解效果和精度也不尽相同，同时负荷特征量采集、处理的难易程度也是实现符合分解必须考虑的一个重要因素。

相较于稳态数据，暂态数据的获得往往需要高频率的采样设备，并且得到的数据量也是巨大的，这就对设备的采样率、精度和数据的处理计算能力提出了较高的要求，投资较高，故适用性较低。而稳态数据更易获得，而且在负荷分解的过程中抗干扰的能力更强，也有很多可以用来测试的数据集，故应用更加广泛。

本次采用的数据集是domestic UK2015数据集。该数据集包含了6个不同房间内的每个电器各自有功功率采样数据和每个房间内所有电器的总有功功率采样数据并且采样周期为6s。

2.2 聚类算法以及电器状态数的识别

首先聚类算法实施的对象是一段时间内单个电器的有功功率采样数据，聚类的目的是从这些数据当中学习得到最佳的电器工作状态数及其相应的功率水平，因为当电器工作在不同状态下时有功功率往往也会不同。而在之后的HMM参数学习时，是以聚类结果——电器状态数为依据的。

K-Means聚类算法是一种基于距离划分类簇的简单有效的聚类方法。其原理和步骤可分为：

1).确定类簇数并随机初始化类簇中心点。

2).计算数据集当中每一个点到各个类簇中心点之间的距离，并将其分配到距离最近的类簇中心点所在的类簇中去。

3).求出各个类簇中所有数据的平均值，并将该值更新为新的类簇中心点。

4).计算新的类簇中心点与上一次中心点之间的距离是否收敛于一个特别小的数，或者达到了最大迭代次数。如果是，则新的类簇中心点就是这个数据集的聚类结果；如果否，则返回到第2)步直到满足4)的条件。

可以看到K-Means聚类算法的原理非常简单并且易于实现，而且相较于其他聚类算法效率也很高。但这一算法的缺点在于：必须事先给定聚类类簇数，并且聚类中心点的初始化也是随机的，而这种随机性对于聚类效果有较大的影响。为了减弱这种随机性，采用了改进后的K-Means聚类算法：K-Means++，这种算法在初始化类簇中心点时，不是随机选取所有中心点，而是在随机选取一个后，计算数据集中的每一个点到与它最近的已选择的聚类中心点之间的距离，根据距离越大，被选择为初始中心点的概率越大的原则，依次选出所有初始类簇中心点。其余步骤均与K-Means算法相同。

由于这种算法并未直接给出最佳的电器状态数，故引入了DBI、CHI两种指标来对不同类簇数的聚类结果进行评价。DBI衡量了每一类与其它类相似度中最大相似度的均值，最小值为0，并且越小代表聚类效果越好。CHI指标综合考虑了类内紧密性和类间分散性，类别内部数据的协方差越小越好，类别之间的协方差越大越好，故CHI越大聚类效果越好。最佳聚类数的选择综合考虑了这两种指标并以有功功率采样数据的图像为辅助依据确定出了最佳聚类数即电器最佳状态数。

2.3 FHMM建模与负荷分解

2.3.1隐马尔科夫模型[3]

首先，隐马尔科夫模型描述的是一个随机变量的随机过程，是一个关于时序的概率模型。它由不可观测的隐状态序列和可以被观测到的观测序列组成，序列的每一个位置可以看作一个时刻。其中隐状态序列遵循马尔科夫过程，其状态空间必须为离散的；观测序列的每一个观测由与其对应的隐含状态生成，状态空间可以为离散或连续。马尔科夫模型由初始概率分布、状态转移概率分布以及观测概率分布确定。以上用数学语言描述为：

Q是所有可能的隐含状态的集合，V是所有可能的观测的集合：

其中N是可能的状态数，M是可能的观测数。

I是长度为T的状态序列，O是对应的观测序列：

A是状态转移概率矩阵：

其中，

是在时刻t处于状态的条件下在时刻t+1转移到状态的概率。

A矩阵是阶数为N的方阵，且每一行的值总和为1。

B是观测概率矩阵：

其中，

是在时刻t处于状态的条件下生成观测的概率。

B矩阵阶数为，每一行的值总和也为1。

是初始状态概率分布：

其中，

是时刻t=1处于状态的概率。

矩阵为阶，总和为1。

隐马尔科夫模型由初始状态概率分布，状态转移概率矩阵A，观测概率矩阵B决定。和A共同决定隐含状态序列，B决定观测序列，因此隐马尔科夫模型用三元符号表示，即

马尔科夫过程遵循了两个基本假设：

1)齐次马尔科夫性假设，任一时刻t是某个隐含状态的概率只与该时刻的前一个时刻隐含状态有关，而与其他时刻的隐含状态或者观测状态均无关，即：

2)观测独立性假设，任一时刻t时的观测状态只与该时刻的隐含状态有关而与其他时刻的隐含状态或者观测状态均无关，即：

基于以上两个基本假设，再根据隐马尔科夫模型参数可生成一条具体的隐马尔科夫链。关于隐马尔科夫模型有三个基本问题：

1)概率计算问题，已知模型参数，求某一确定的观测序列发生的概率。解决这一问题需要用到前向算法或者后向算法。

2)学习问题，已知某一确定的观测序列，估计模型参数，使得在此模型参数下，产生这一观测序列的概率最大。解决这一问题需要用到Baum-Welch算法即EM算法。根据电器有功功率序列求出相应HMM模型参数就属于学习问题。

3)预测问题，已知模型参数和一个确定的观测序列，预测最有可能的隐含状态序列。解决这一问题需要用到Viterbi算法。根据FHMM模型参数和总有功功率观测序列计算相应的电器工作状态序列便属于预测问题。

2.3.2基于HMM的单个电力负荷模型

利用单条隐马尔科夫链对一个电器进行负荷建模，每个电器的工作状态随时间的变化抽象为马尔可夫过程下的隐含状态序列，而电器的有功功率随时间变化是隐马尔科夫链的可观测序列，二者一一对应，并且电器的工作状态是不可观测的，实际可观测到的是这个电器有功功率随时间变化的离散序列。

其次，因为即使电器在同一工作状态下其有功功率也不可能是一个定值，而应该是在一定范围内随机波动，所以观测随机变量的状态空间不可能由离散值全部列举出来，故观测概率矩阵每一行应该是基于一定概率分布的连续函数。这里选择我们最常见的高斯概率密度分布函数作为矩阵B的元素，矩阵共N行，每行是一个不同的正态分布函数，B的参数学习也就是求出每个正态分布的均值，和方差。

因为FHMM参数是由所有隐马尔科夫链的参数合成而来，所以必须首先根据每个电器有功功率采样数据即每条HMM链的观测序列学习得到HMM的参数。学习参数需要事先知道单个电器的工作状态数，因为这样三个矩阵的维度才能确定下来，这也就是为什么要进行聚类确定电器状态数的原因。

2.3.3从HMM模型到FHMM模型

由于所有电器隐状态随机变量之间相互独立，所以每个电器由一条马尔科夫链单独表示，当每条链的观测序列即每个电器有功序列无法观测，可以观测的是由所有观测序列线性相加得到的总有功功率序列时的模型便是FHMM模型。

由以上所述，FHMM模型是由与电器个数相等条数的隐含状态序列和一条总的有功功率可观测序列构成。

这样包含K个用电器的FHMM负荷模型同样可以用参数表示出来：

1)参数为初始状态概率分布，其中的元素可以表示为,代表了所有电器工作状态排列组合，如果第j个电器有种工作状态的话，那么序列共有种组合,即有个元素。由于各条马尔科夫链之间相互独立，所以。

2)参数A为状态转移概率方阵，其中的元素可以表示为, 同样是代表了所有电器工作状态的排列组合，方阵A共阶。同样由于各条马尔科夫链之间相互独立，所以。

3)参数B为观测概率矩阵，其中的元素可以表示为，概率P实际上应该是一个高斯概率密度分布函数，如果第j个电器t时刻处于状态，则它所对应的观测高斯概率密度函数的均值为，方差为，由于各条马尔科夫链隐状态随机变量之间相互独立，所以根据正态分布性质，概率密度函数P的均值为，方差为。同样B矩阵也有行。

当参数的每个元素都确定下来后，FHMM模型也就唯一确定了。进一步，如果当中的元素都是按一定顺序排列的话，比如：以代表第j个电器工作状态中的第一个状态，则参数的第一个元素中，；方阵A第一行第一列位置的元素中，，；矩阵B第一行元素中。以此类推，每一个互相对应位置上的元素都与同一种状态的排列组合有关，且每个参数中元素排列的顺序遵循排列组合的先后顺序，那么FHMM实际上就可以看成是隐含状态集合为的HMM模型，其中是状态的组合，是状态的组合，以此类推。集合的排列顺序也和的元素排列顺序一致。

实际计算中，顺序对每个HMM的使用矩阵的克罗内科积便可以满足以上所有要求。

2.3.4基于FHMM的负荷分解

将每个HMM参数按照如2.3.3所述合成FHMM参数后便可以利用此模型进行负荷分解了。

根据电力入口处总的有功功率采样数据即FHMM的唯一的一条有功观测序列，可以通过求解基本的预测问题，利用Viterbi算法计算得到最有可能的电器工作状态组合序列。

又因为FHMM的隐含状态集合中的每一个元素是一种确定的各电器状态的排列组合，所以当与FHMM等价的HMM的状态序列确定后，便可根据对应关系得到每个电器的工作状态序列，即得到了每条HMM状态序列，这样就实现了负荷状态识别。

要得到每个电器具体的有功功率观测序列，需要求解一个条件极值问题。因为每个电器的工作状态序列已经预测得到，所以要求解的极值问题是在某一时刻t，，约束条件是，所有电器同一时刻的有功功率之和等于总的有功功率，即。求解出所有时刻下的条件极值问题，便得到了每个电器的有功功率观测序列，这样就解决了负荷分解问题。

2.3.5分解精度的评价 [1]

模型得到的结果是一种近似的结果，不可能完全和实际值相同，总会存在或大或小的误差，故需要有一些参数来衡量这些误差的大小，对负荷分解模型的精确度做出评价。

评价负荷分解的精度时采用了、总分解误差和能耗占比的绝对误差三种评价指标。

1)

其中pre称为精确度rec称为召回率，计算公式如下：

式中,表示负荷j在t时刻通过分解算法得到的预测功率值,表示负荷j在t时刻的真实功率值,T是指预测样本总数。

为了分析分解算法的整体性能，需要计算预测集中所有负荷的Pre和Rec，取每个负荷精确度和召回率的平均值，如下：

式中,K为预测集中的负荷总数。

的值越接近于1表示分解精确度越高。

2) 总分解误差

的值越接近于0表示分解精确度越高。

3) 能耗占比的绝对误差

可以用每种负荷所消耗能量占比的绝对误差来评估每个负荷所消耗电能的情况，计算公式如下：

式中，表示负荷j分解所得消耗能量占比，表示真实情况下负荷j消耗能量占比，为以上两种占比的绝对误差。的值越小，说明能量消耗的预测越精度，每种负荷消耗能量的占比越接近真实值。

1. 研究内容

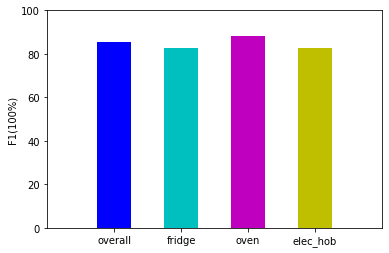
本课题选取了domestic UK2015数据集当中部分电器的有功功率采样数据作为研究对象，分别对其进行聚类和HMM参数的学习。之后，将所有HMM参数合成FHMM参数，所有单个电器的有功功率观测序列合成电力入口处总的有功功率观测序列，然后利用FHMM模型进行状态识别和负荷分解，最后利用评价指标对负荷分解的精度做出评价。

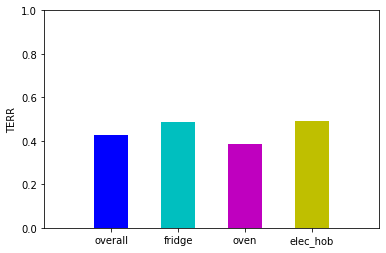
1. 现阶段成果

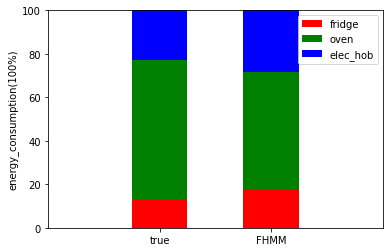
1．能够熟练运用python进行编程，掌握了python的sklearn库，hmmlearn库在负荷分解中的应用，并且利用python编程完成了FHMM模型的搭建，实现了基于FHMM的负荷分解。

2．掌握了HMM的基本结构和原理，研究了HMM三大基本问题的数学过程，探究了FHMM实现方法，实现了从HMM到FHMM的转变。

3.实际的应用过程中，在domestic UK2015数据集当中挑选了电冰箱、电烤箱、电炉三个电器，并对它们的总有功功率采样样本进行了非侵入式负荷分解，结果显示达到了0.85，达到了0.43。下图为各个负荷的评价指标数据柱状图：







1. 问题与建议

目前，FHMM模型对于一些运行工作状态较为复杂的负荷分解精度不够理想，并且当负荷数较多时，分解精度会有不同程度的降低。

另外，该数据集所给出的一个房间内总的有功功率采样数据有较为严重的偏差，即同一时刻下各个电器有功功率之和与总的有功功率不相等，且这一差值的数量级与数值本身相同甚至高于数值本身，根本无法忽略，所以数据集中的总有功功率数据无法使用，必须先自己合成。但是，这样又会出现一个问题，就是同一段时间内每个电器的采样点数不同，也就是说时间点无法一一对应，这样又对数据的预处理造成了困难。

1. 后期工作计划

1)考虑使用Explicit-Duration HMM对模型进行改进。

2)考虑使用字典学习和稀疏表示的方法进行NILM。

1. 主要参考文献
   1. 孙晓伟．基于FHMM的非侵入式负荷分解研究[D]．郑州:郑州大学,2018．
   2. 刘玲．基于稳态特征的非侵入式负荷监测算法研究[D]．广州：华南理工大学,2019.
   3. 李航,统计学习方法 [M].北京：清华大学出版社,2012.