基于余弦相似度分类负荷预测

南京易司拓电力科技股份有限公司 罗耀强 陈延彬 陈昌友

摘要：采用余弦相似度算法对历史负荷数据进行曲线特征抽取并进行聚类，对每个分类用机器学习算法进行构建模型，依据曲线相关度来决定使用哪种模型对当前负荷进行预测。

关键词：负荷预测；余弦相似度；机器学习

电网企业网供负荷特性的分析和预测是电网调度运行方式预测工作的一个重要方面，准确把握网供负荷特性及其变化趋势是做好电网调度、运行方式调整等工作的重要基础，也是制定电网规划、安排设备检修的重要参考。尤其近年来新能源大规模接入、用户多样化的电力需求，电网负荷特性的预测难度大幅增加，一方面影响电网网供负荷特性指标增多，指标之间关联性进一步加强；另一方面影响负荷特性变化的因素更加复杂，一些气候因素如光照度、时长、气温、降雨等具有很大不确定性。

本文针对上述问题利用负荷曲线相似度的方法来对一定时期内的负荷曲线余弦特征值进行聚类，从而得到包含曲线相似特征属性的簇类。在对同时期的天气状况进行聚类，得到天气状况的簇类。将负荷簇类与天气簇类进行关联度分析，从而得到负荷曲线与天气状况的关联关系。从关联关系出发在来对负荷曲线进行分类训练产生算法模型，从而得到更精确的预测效果。

1 网供负荷特征分析

现有的网供负荷特征分析，大都从历史数据与天气数据人工筛选关联关系从而建立对应模型。从现有的筛选条件看，大都按照季节和天气情况的因素来区分网供负荷的特征曲线，但这种筛选方式本身人为的因素影响较大且如果数据集太大人工筛选的方式就显得不切实际，所以本文提出按照网供负荷特征曲线与天气及节假日进行关联分析，从而从算法上证明曲线特征与天气及节假日的关联关系[1]。

1.1 网供负荷特征选择及聚类

从负荷分析上看，每天第一个测量点的负荷必然与前一日最后一个测量点的负荷有所关联。确定了每天初始测量点的预测负荷，对于网供负荷的特征只需抽取每天负荷曲线的曲线相似度即可。对于负荷曲线相似度特征值的抽取，可采用余弦算法来计算各点的余弦值以及各同时间负荷节点的相似度。

1.2 负荷余弦值的抽取

余弦相似度是利用向量空间中两个向量夹角的余弦值来衡量两个向量的差异，余弦值越接近1就意味着两个向量的夹角接近于0度，也就是两个向量越相似。余弦计算的公式（1）：$\cos(\theta)=\frac{a^2+b^2-c^2}{2ab}$，假设a向量是$(x_1,y_1)$,b向量是（$x_2,y_2$），那么余弦定理公式（2）：$\cos(\theta)=\frac{a\bullet b}{\|a\|\times\|b\|}=\frac{(x_1,y_1)\bullet(x_2,y_2)}{\sqrt{x_1^2+y_1^2}\times\sqrt{x_2^2+y_2^2}}=\frac{x_1x_2+y_1y_2}{\sqrt{x_1^2+y_1^2}\times\sqrt{x_2^2+y_2^2}}$，如果向量a和b是个多维的向量，公式（2）的计算方法仍然正确，则公式（3）为：

$$\cos(\theta)=\frac{\sum_{i=1}^{n}(x_i\times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2}\times\sqrt{\sum_{i=1}^{n}(y_i)^2}} \quad (3)$$

1.3 负荷余弦特征值聚类

各负荷检测点的余弦特征值按天为单位进行特征抽取，以天为单位将负荷节点余弦特征值进行聚类[2]。对于输入的负荷特征值D=($x_1,x_2,\cdots,x_m$)，邻域参数(∈,MinPts)，输出为聚类组的划分C。

第一步，初始化核心对象集合 $\Omega=\phi$，初始化聚类的簇数 k=0，初始化未访问样本集合 $\Gamma=D$，簇划分 $C=\phi$；第二步，对于 J=1,2,…,m 按照如下步骤找到所有核心对象：利用距离度量算法，找到样本 x_j 的 ε－邻域子样本集 $N_\varepsilon(x_j)$，假设子样本集样本的个数满足 $|N_\varepsilon(x_j)| \geqslant MinPts$，那么就把样本 x_j 加入核心对象样本集合 $\Omega=\Omega \cup \{x_j\}$；第三步，假设核心对象集合 $\Omega=\phi$，那么就可以结束算法，否则转至下一步继续。

第四步，从核心对象集合 Ω 中随机选择核心对象 o，对当前的簇核心对象队列 $\Omega_{cur}=\{o\}$、对类别序号 k=k+1、当前簇样本集合 $C_k=\{o\}$ 进行初始化操作，更新未访问的样本集合 $\Gamma=\Gamma-\{o\}$；第五步，如当前簇核心对象对列 $\Omega_{cur}=\phi$，那么当前的类簇 C_k 生成完成，更新簇划分 $C=\{C_1,C_2,\cdots,C_k\}$，更新核心对象集合 $\Omega=\Omega-C_k$，结束本算法；第六步，在当前簇核心对象对列 Ω_{cur} 中取出核心对象 o'，通过邻域距离阈值 ε 找到所有的 ε－邻域子样本集 $N_\varepsilon(o')$，令 $\Delta=N_\varepsilon(o') \cap \Gamma$，更新当前簇样本集合 $C_k=C_k \cup \Delta$，更新未访问的样本集合 $\Gamma=\Gamma-\Delta$，更新 $\Omega_{cur}=\Omega_{cur} \cup (\Delta \cap \Omega)-o'$，转到第五个步骤。最后输出的结果：簇的划分为 $C=\{C_1,C_2,\cdots,C_k\}$。

2 网供负荷特征与天气因素关联分析

在进行网供负荷特征与天气因素关联分析前，需要对天气数据进行非结构化数据到结构化数据的转换，将天气状态，风向等文字信息转化成数值格式。结构化的天气数据利用算法进行聚类，得到天气数据的聚类模型。

表1 数据的二元表示

T	C1	C2	C3	D1	D2	D3
1	1	0	0	0	1	1
2	0	1	0	1	0	0
3	0	1	1	0	0	1

网供负荷特征的聚类模型和天气数据的聚类模型进行关联分析，形成一组 {C,D} 的关系，C 为网供负荷特征在聚类模型的簇值，D 为天气数据在聚类模型中的簇值。问题的定义：表1表示部分网供负荷特征与天气数据的记录，其中每一行都是一天的负荷特征与天气情况，每一列对应一个簇值。令 $I=\{i_1,i_2,\cdots,i_d\}$ 是数据中的所有项的集合，而 $T=\{t_1,t_2,\cdots,t_n\}$ 是所有事务的集合。每个事务 t_i 包含的项都是 I 的子集。在关联分析中有两个概念——支持度和置信度，支持度指给定的项集在事务 T 中同时出现的概率，置信度指出现 C_i 的事务 T 中项集 C_j 也同时出现的概率。支持度如公式（4），置信度如公式（5）。

$$s(X \to Y)=\frac{\sigma(X \cup Y)}{N} \quad (4)$$

$$c(X \to Y)=\frac{\sigma(X \cup Y)}{\sigma(X)} \quad (5)$$

构建 FP-tree：通过构建 FP-tree 是一颗前缀树，用来压缩数据中的信息，其按支持度降序排列，支持度越高的频繁项离根节点越近，从而使更多的频繁项可以共享前缀。表2中，$C_1,C_2,\cdots,C_n,D_1,D_2,\cdots,D_m$ 表示网供负荷与天气数据的簇值。首先，对该数据集进行一次扫描，计算每一行记录的各记录的支持度，然后按支持度排列，保留频繁项集，假设这里设置支持度阈值取2，从而得到 $<(D_2:3),(C_1:2)>$。表2的第三列展示了排序后的结果。

FP-tree 的根节点为 null，然后对数据进行第二次扫描，从而构建 FP-tree。为了方便对整个 FP-tree 进行遍历，建立一张项的头表。表的第一列按照降序排列频繁项。第二列指向该频繁项在 FP-tree 中节点位置的指针。FP-tree 中的每个节点的指针指向相同名称的节点。

表2 关联分析数据表

T	Items	Frequent Items
100	C_1,C_2,D_1	C_1,D_1
200	C_1,C_3,D_2,D_3	D_2,C_1
300	C_1,C_3,D_2	D_2
400	C_2,C_3,D_3	D_2,C_2

从 FP-tree 中挖掘频繁模式：从头表的底部开始挖掘 FP-tree 中的频繁模式。在 FP-tree 中以 C_1 结尾的节点链只有一条，为 $<(D_2:3),(C_1:2)>$，尽管 D_2 出现了3次，但是它与 C_1 同时出现只有1次，所以在 FP-Tree 中记为 $<(D_2:1),(C_1:1)>$。将 C_1 的前缀节点链 $<(D_2:1)>$ 称为 C_1 的条件模式基。将 C_1 的条件模式基作为新的事务数据，每一行存储 C_1 的一个前缀节点链，根据构建 FP-tree 的过程，计算每一行记录中各数据的支持度，然后按照支持度降序排列，仅保留频繁项集，建立一颗心的 FP-tree，这个树称之为 C_1 的条件 FP-Tree。

3 基于 LSTM 的网供负荷预测

3.1 LSTM 模型搭建

在网供负荷及天气数据聚类完成后，按簇类来抽取网供负荷数据集及相同日期的天气情况，节假日情况数据等进行模型的训练。

LSTM 的模型结构[3]：LSTM 模型中在每个序列索引位置 t 时刻向前传播除了隐藏状态 $h^{(t)}$，还需要传送一个称为细胞状态为 $C^{(t)}$。LSTM 中包含门控结构，其中包含遗忘门、输入门和输出门。遗忘门的主要作用是以移动的概率控制是否遗忘上一层隐藏细胞的状态。例如上一序列的隐藏状态 $h^{(t-1)}$ 和本序列数据 $x^{(t)}$，通过连接激活函数得到遗忘门的输出 $f^{(t)}$，由于激活函数的特性使得 $f^{(t)}$ 的输出值在 [0，1] 之间，因此这里的输出 f^{(t)} 代表了遗忘上一层隐藏细胞状态的概率。其公式为 $f^{(t)}=\sigma(W_f h^{(t-1)}+U_f x^{(t)}+b_f)$。

输出门负责处理当前序列位置的输入，其由两部分组成，第一部分通过激活函数，输出值为 $i^{(t)}$，第二部分使用了 tanh 激活函数，输出值为 $a^{(t)}$，将两者相乘再去更新细胞状态：$i^{(t)}=\sigma(W_i h^{(t-1)}+U_i x^{(t)}+b_i)$，$a^{(t)}=\tanh(W_a h^{(t-1)}+U_a x^{(t)}+b_a)$。

在讨论 LSTM 的输出门之前，先查看下 LSTM 的细胞状态。遗忘门和输入门的输出都会对细胞状态 $C^{(t)}$ 有影响。$C^{(t)}$ 是一部分由 $C^{(t-1)}$ 和遗忘门输出 $f^{(t)}$ 的乘积，另一部分是由输入门 $i^{(t)}$ 和 $a^{(t)}$ 的乘积，公式为 $C^t=C^{(t-1)}\otimes f^{(t)}+i^{(t)}\otimes a^{(t)}$。

输出门的隐含状态 $h^{(t)}$ 的更新由两部分组成，一部分是 $\sigma^{(t)}$ 由上一序列的隐藏状态 $h^{(t-1)}$ 和本序列数据 $x^{(t)}$ 以及通过激活函数得到，另一部分是由状态 $C^{(t)}$ 和 tanh 激活函数组成，公式为：$o(t)=\sigma(W_o h^{(t-1)}+U_o x^{(t)}+b_o)$，$h^{(t)}=o^{(t)}\otimes\tanh(C^{(t)})$。

3.2 基于余弦相似度的网供负荷预测

按如下流程进行预测：在聚类模型中对预测日的天气数据和节假日数据的进行分类，取得预测日天气和节假日的聚类簇值；使用聚类的簇值从网供负荷与天气数据的关联关系中找到对应的网供负荷的簇值，从网供负荷的簇值找到对应的 LSTM 算法模型，使用此模型对预测日的网供负荷进行预测[4-5]。随机抽取5天的负荷进行对比测试，结果如表3。抽取一天的网供负荷曲线与预测负荷曲线进行对比（图1）。

表3 负荷预测准确度对比

样本序号	新方法准确度 %	经验分类方法准确度 %
1	96.49	94.16
2	96.32	93.82
3	95.98	94.24
4	96.67	93.97
5	96.11	94.09

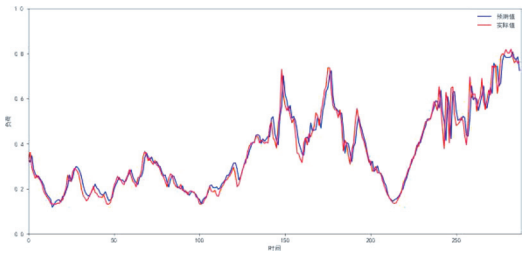

图1 负荷实际值与预测值曲线

4 结语

本文从寻找负荷曲线与天气状况的关联关系出发，提出了采用余弦相似度算法来聚合负荷曲线，再与天气聚类寻找网供负荷与天气状况的关联关系。从而可以科学的对负荷曲线按照天气状况进行分类训练，再利用 LSTM 算法对网供负荷的不同簇类生成不同的算法模型。本文虽然对负荷曲线利用余弦相似度的方法进行了聚类，达到了预想的效果，但是针对天气状况，节假日状况尚无有效的方式抽取特征值进行聚类。从现状来看对于天气状况的聚类效果较差，从而也影响了关联分析的准确性。今后的工作需要对天气状况、节假日状况需要寻找更有效的特征抽取算法来生成有效的聚类。

参考文献

[1] 刘艳红，倪秋龙，黄民翔．多小水电地区网供负荷预测研究 [J]. 浙江电力，2015,12.
[2] 范金骥．基于 ARMA 与 ANN 模型组合交叉方法的电网日负荷预测 .[J]. 浙江电力，2018,8.
[3] 罗育辉，蔡延光，等．基于最大偏差相似性准则的 BP 神经网络短期电力负荷预测算法 [J]. 计算机应用研究，2018,11.
[4] 刘建军．电力系统负荷预测综述 [J]. 中国科技信息，2016,16.
[5] 程宇也．基于人工神经网络的短期电力负荷预测研究 [D]. 浙江大学，2017.