基于数据挖掘技术的负荷曲线对故障反应相似性的研究

林济铿[1]，罗萍萍[2]，曹绍杰[3]，C. M. MAK[4]，K. M. YUNG[4]
（1. 天津大学电气与自动化工程学院，天津市 300072；2. 上海电力学院，上海市 200437）
（3. 香港城市大学，香港；4. 香港中华电力公司，香港）

摘要：各种故障都对电力系统负荷的变化有显著影响，深入地理解和掌握这些影响是十分有益的。文中把数据挖掘技术应用于 CLP(Chinese Lighting Power)公司的数据库，分析了在故障影响下母线负荷曲线的聚类，获得了与 CLP 系统中的特定变电站 AAA 母线具有相似负荷变化曲线的母线组或区域，从而可以更好地估计和抑制未来故障对这些负荷所造成的损失(例如对它们采取一致的策略等)，有利于系统的安全稳定运行。

关键词：数据挖掘；故障；负荷曲线；聚类分析
中图分类号：TM714

0 引言

电力系统的负荷总是处于不断变化和波动之中。各种系统扰动，尤其是故障引起的扰动，会显著影响各母线的负荷。对于各种故障，如果能发现负荷总是受到相似影响的母线组或区域，对进一步理解和掌握潜在故障影响下的系统运行状况有一定的帮助。例如在分析系统的安全性时，这些节点的负荷可以采用结构相同或相似的模型，从而有助于系统调度人员或规划人员做出更好的决策。

数据挖掘或信息挖掘是指在数据中发现有效的、异常的、有潜在使用价值或最终可被理解模式的过程，主要用于从大量数据中挖掘出潜在有用信息(定性的或定量的)。数据挖掘依赖于各种方法和技术，如决策树分析、聚类分析、统计分析、可视化分析和专家系统方法等。数据挖掘已经被应用于许多实际领域，比如：大范围通信网[1]，核电站的生产过程[2]，水、火电厂专家系统规则的获取[3]，热电厂的监视和优化过程[4]，变压器的故障诊断[5]，相似用户负荷曲线的获取[6]，以及电力系统总体数据的挖掘[7]等。

本文应用聚类分析和统计分析，根据一段时间内系统所发生的各种故障，确定具有相似故障反应负荷曲线的区域或母线组。应指出，文献[6]也有相似的研究，但其主要讨论正常情况下负荷的变化和聚类，对故障情况并未涉及。

本文根据对 CLP(Chinese Lighting Power)公司电力系统数据库中负荷曲线的聚类分析，在系统出现大的故障时，找出了与特定变电站 AAA 的母线具有相似负荷变化曲线的母线组或区域。

收稿日期：2004-07-02；修回日期：2004-08-30。

1 聚类分析的基本方法

常用的数据挖掘方法有多种[8,9]，如概念/分类描述、联合分析、分类及预测、聚类分析、例外分析、演化分析等。其中，聚类分析在各个领域应用较为广泛。

聚类分析由若干模式组成，而模式是一个度量的向量或多维空间中的一个点。聚类分析以相似性为基础，类与类之间的模式相似性要小，类内的模式相似性要大。根据类内相似性最大化、类间相似性最小化的原则，对模式进行分类。这样形成的每一个类视为一类物体的聚类，从中可推导出许多有用的规则。

事实上，聚类算法中很多算法的相似性都是基于各种形式的距离测度。根据距离测度定义的不同和研究问题初始条件的不同，有以下几类方法：分裂法、层次法、基于密度的方法、基于网格的方法、基于模型的方法。其中，层次法是对给定的数据集进行层次的分解，直到某种条件满足为止，具体又可分为“自下向上”和“自上向下”两种分解方案。

本文将聚类分析中的层次法结合统计分析应用于 CLP 系统的负荷曲线相似性和特异性的挖掘。

2 负荷曲线的聚类分析

不同性质的电力用户对系统故障扰动的反应是不一样的。因此，从历史数据中找出对系统故障具有统计意义上相似反应的母线组或区域有一定的实际意义。例如在安全性分析时，对这些负荷采用相

同或相似的负荷模型,在进行负荷控制时,对这些负荷采取相同或相似的措施等,从而有利于制定出对系统运行更加有效的运行和调度计划。已经发现[10],CLP 系统的一个 11 kV 变电站 AAA 对故障扰动十分敏感。本文利用聚类分析和统计分析找出与该变电站密切相关的其他变电站负荷,通过对这些用户的进一步调查分析,更好地估计和抑制未来故障对这些负荷造成的损失,从而有利于系统的安全经济运行。

2.1 聚类分析

2.1.1 建立聚类树

采用自下向上的聚类算法[11],对事先不知道聚类数目的问题来说,它比均值更加适用[9]。每一负荷曲线 $W_j(j=1,2,\cdots,M)$ 用一个序列对 $(y_{ji},t_i)(i=0,1,\cdots,N)$ 表示,这里 y_{ji} 表示第 j 条曲线的第 i 个值,t_i 表示故障发生后所经过的 16 s 间隔数,M 是所考虑的负荷曲线数,N 是所考虑的总间隔数。其过程如下:

1)以 M 个聚类开始,即每一聚类只包括 1 个元素(负荷曲线)。

2)计算任意 2 个元素之间的距离测度,距离测度公式为[12]:

$$d_{ij}=\frac{1}{N}\sqrt{\sum_{n=1}^{N}\left(\frac{y_{in}}{y_{i\max}}-\frac{y_{jn}}{y_{j\max}}\right)^2} \tag{1}$$

或

$$d_{ij}=\frac{1}{N}\sqrt{\sum_{n=1}^{N}\left(\frac{y_{in}}{y_{i\max}}-\frac{y_{in}}{y_{j\max}}-\Delta\right)^2} \tag{2}$$

式中:

$$\Delta=\frac{1}{N}\sum_{n=1}^{N}\left(\frac{y_{in}}{y_{i\max}}-\frac{y_{jn}}{y_{j\max}}\right)^2$$

3)形成相异矩阵 $\boldsymbol{C}$:

$$\boldsymbol{C}=\begin{bmatrix} c(1,1) & c(1,2) & \cdots & c(1,s) \\ & c(2,2) & \cdots & c(2,s) \\ & & & \vdots \\ & & & c(s-1,s) \\ & & & c(s,s) \end{bmatrix} \tag{3}$$

式中:元素 $c(i,j)$ 表示聚类 i 和 j 之间的距离;$c(i,j)$ 为按式(1)或式(2)定义的 d_{ij};$1,2,\cdots,s$ 为当前聚类序列号。

若式(3)中最小的元素是 $c(e,f)$,即聚类 e 和 f 是最相似的,将它们合并在一起形成一个新的聚类,称为聚类 P;其他聚类不变,合称为 Q 集。

4)更新相异矩阵 $\boldsymbol{C}$。聚类 P 中至少有 2 个元素,Q 集至多有 $m-2$ 个聚类。因为聚类 P 是由多个聚类融合而成,所以聚类 P 到 Q 中每个聚类的距离为聚类 P 中的每一元素到 Q 中相应聚类距离的平均值。若更新距离后的 $\boldsymbol{C}$ 中最小元素记为 $c(i,j)$,表示当前阶段聚类 i 和聚类 j 最相似,下一次就应合并它们;它们可以一个是 P 中的聚类,另一个为 Q 中的聚类,或者两个都是 Q 中的聚类。继续这一过程,这些聚类集在合并的过程中变得越来越大,直至形成包括所有负荷曲线的单个聚类。

2.1.2 把一个聚类树划分成聚类组

利用基于距离测度的距离指标线(下文称相似性距离指标线,小于这个指标,意味着属于同一聚类,是相似的),横向分割聚类树,所得到的每一子树即为一个个聚类(满足相似性要求),这些聚类组成了满足相似性距离指标的聚类集。

2.2 对故障具相似负荷变化反应的母线聚类

由于 CLP 电力系统的 11 kV 变电站 AAA 对系统故障最敏感,因此,可以找出在 CLP 系统故障时,其负荷特性与变电站 AAA 的母线在统计学上最相似的母线组。

对于 D 个故障,应用上述聚类分析产生了 D 个聚类集,每个聚类集对应 1 个故障。在 D 个聚类集中,统计每一母线与变电站 AAA 的母线属于统一聚类的次数。在这个由大到小的与变电站 AAA 母线属于统一聚类的次数列表中,排在最高位的相应母线自然应拥有与变电站 AAA 的母线最相似的故障反应。

2.3 数据处理和结果解释

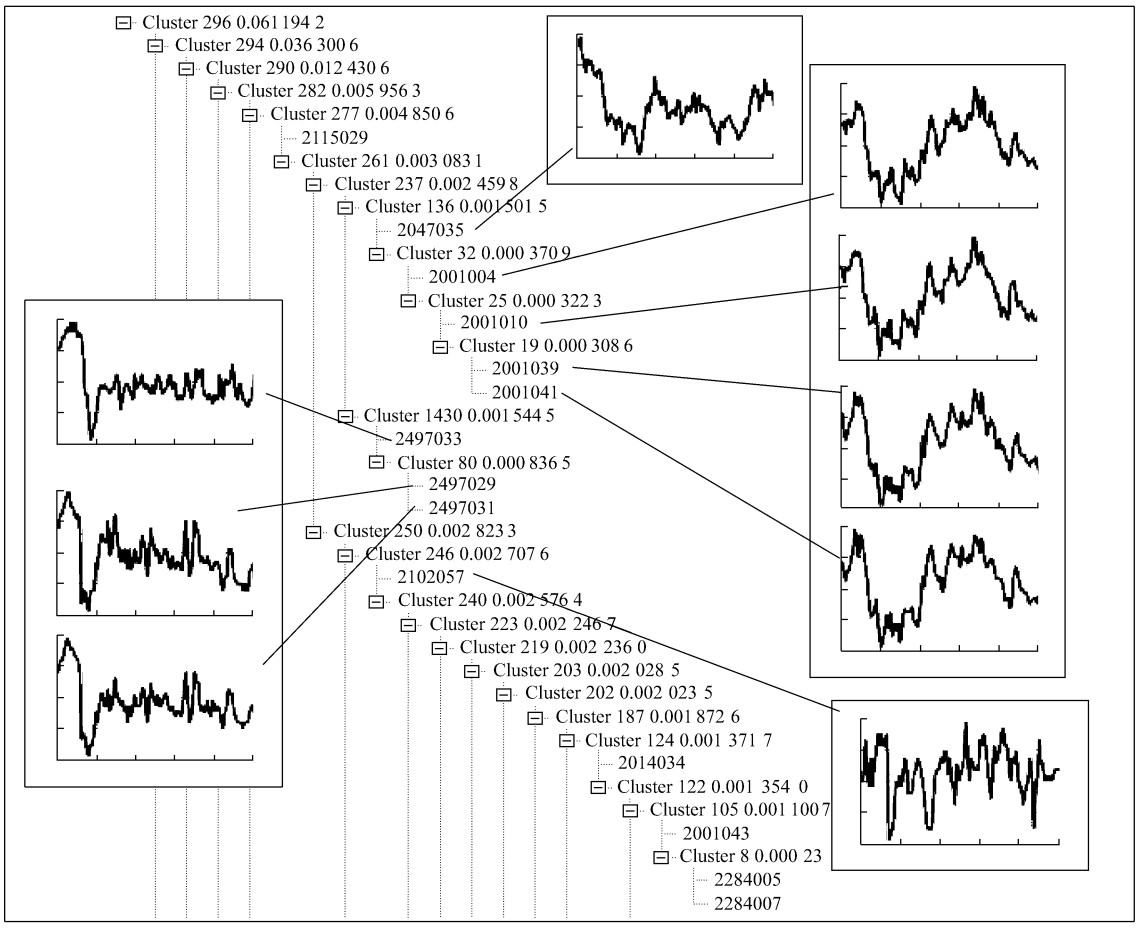
2.3.1 源数据

在 CLP 系统的数据库中,2000 年~2002 年 400 kV 网络共发生了 30 多个故障。由于各种原因,只有 15 个故障的数据是完备的,本文的分析即基于这 15 个故障的数据记录。每个故障的数据包括 300 多个 11 kV 变电站变压器的二次侧负荷曲线,曲线以 16 s 为记录间隔,记录时间为故障前 3 min到故障后 27 min。

2.3.2 负荷曲线聚类树

对于 15 个故障中的每一故障,均按照上述聚类树建立方法建立一个 11 kV 负荷曲线聚类树,树的一部分如图 1 所示,部分母线的负荷曲线也显示其中。从子树的叶节点可以看出,标号为 2001039 的负荷曲线与标号为 2001041 的负荷曲线十分相似,但是沿着叶节点到根节点这种相似性越来越小,比如标号为 2497033 的负荷曲线和标号为 2497029 的负荷曲线。位于不同子树的负荷曲线不会十分相似,即使它们同为更大子树的子树。因为在聚类树建立过程的某一阶段,这两个子树之间的平均距离是它们与其他聚类之间以及其他聚类互相之间最小的,因此,这两个子树(聚类)合并成一个更大的子树

(聚类),即使它们之间的距离还比较大。例如在图1中,136号和143号是2个子树,但它们同属于子树237;标号为2047035的负荷曲线属于子树136,标号为2497033的负荷曲线属于子树143,但它们并不是很相似。



整数2001041,2001039等表示每个负荷曲线的标号,单词Cluster后的第1个整数表示该子树中所含的子树数,后边的小数表示该子树中元素之间的距离

图1 部分聚类树(包括一些负荷曲线)

Fig. 1 Part of the cluster tree (with some profiles included)

2.3.3 与变电站AAA的母线对故障有相似反应的母线组

获得了15个故障的聚类树之后,用最大元素(变压器的负荷曲线)间距离的1/8作为相似性距离指标线,把每一聚类树转化成聚类集。只关注包括AAA变电站H1,H2号变压器的负荷曲线的15个聚类,并且计算每一元素在该15个聚类中出现的次数(除了AAA自己的负荷曲线),位于前3位的变压器负荷与变电站AAA的负荷在故障中具有统计意义上的负荷反应相似性。这一结果示于表1中,其中第1个变压器母线是作为参考的AAA变电站中的变压器母线。

从表1可以看出,位于表中前3位的变压器母线(除了变电站AAA自己的变压器母线之外)的次数很大,最多为10次,最少为5次,说明这些变压器的负荷与AAA变电站的负荷在统计意义上有相似的故障反应。同时,BBB变电站的次数最大,表明BBB变电站对故障的反应与AAA变电站有统计意义上的相似性。但是,BBB变电站最主要的负荷成分是电梯和空调(位于香港的酒店集中区),而AAA变电站的主要负荷是电动机(位于集装箱基地),这个新发现与CLP公司专家的运行经验有一定的偏差。基于此发现,CLP公司启动了一项独立的研究,进一步分析变电站AAA与变电站BBB的负荷特性,同时建立它们的负荷模型。

经过分析,这种相似性是合理的。BBB变电站的供电对象是高级酒店区,且每一高级酒店基本都是智能大楼,中央空调及各种电梯是每天24 h运行,加上其他的小部分负荷,构成了此地区的主要负荷,基本是电动机群负荷。AAA变电站的供电对

象是码头,主要是各种升降机和起重设备,其主要负荷基本也是电动机群。香港的集装箱码头号称是亚太区最为繁忙、效率最高的集装箱码头,基本也是每天 24 h 作业,因此,同是电动机群负荷,且基本都是每天 24 h 运作,导致对故障的反应具有相似性。

表 1 出现在次数列表前 3 位的变压器母线组
Table 1 Top three groups of transformers appearing in the list

变压器母线	次数	变压器母线	次数
AAATFH1011P12	15	AAATFH2011P21	15
AAATFH3011P30	11	BBTXH1	8
BBBTXH1	10	AAATFH3011P30	8
CCCTFH1011P12	9	AAATFH1011P12	
CCCTFH2011P21	9	FFFTXH2011	6
DDDTFH3011P30	9	EEETXPD1011	6
EEETXH6011	9	EEETXH6011	6
GGGTXH1011	9	HHHTXH2	6
JJJTAITFH3011	9	DDDTFH3011P30	6
KKKTXH1011	9	LLLTXH1	6
HHHTXH2	9	KKKTXH1011	6
EEETXPD1011	9	MMMTFH1011	6
NNNTFPD3011	9	NNNTFPD3011	6
PPPTXH2011	9	QQQTXH2	6
RRRTXH2011P10	9	MMMTFH3011	6
SSSTFH3011	8	JJJTFH3011	5
TTTTFM1011	8	QQQTXH1	5
AAATFH2011P21	8	VWWTFH2011P14	5
MMMTFH1011	8	WWWTXH2011	5
UUUTXH1011	8	XXXTXH2011	5

依据 CLP 公司电力系统工程师的运行经验,证实在表 1 中居高位的母线组对故障的反应是相似的。

以上采用最大元素(变压器的负荷曲线)间距离的 1/8 作为相似性距离指标线,对于本研究问题是适用的。本文用最大元素(变压器的负荷曲线)间距离的 1/6,1/10 或 1/12 作为相似性距离指标线,也获得了相似的结果。

3 结语

本文把数据挖掘技术运用于 CLP 公司的数据库,分析了故障情况下负荷曲线的聚类分析。通过对负荷曲线的聚类分析,找到了 CLP 系统中在故障情况下与变电站 AAA 母线负荷反应曲线最为相似的母线组,对它们采取一致的策略,从而更好地估计和抑制未来故障对这些负荷造成的损失。在对系统进行安全性分析时,若得不到详细、准确的负荷模型,对这些负荷可采用相同或相似的负荷模型,有利于系统的安全稳定运行。

参 考 文 献

[1] SASISEKHARAN R, SESHADRI V. Data Mining and Forecasting in Large-scale Telecommunications Network. IEEE Expert, 1996, 11(1): 37—43.

[2] LEECH W J. A Rule Based Process Control Method with Feedback. ISA Transactions, 1987, 26(2): 73—80.

[3] MANUEL M L, GUILLERMO R O. Obtaining Expert System Rules Using Data Mining Tools from a Power Generation Database. Expert System with Application, 1998, 14(1, 2): 37—42.

[4] OGILVIE T, SWIDENBANK E, HOGG B W. Use of Data Mining Techniques in the Performance Monitoring and Optimisation of a Thermal Power Plant. In: Proceedings of IEE Colloquium on Knowledge Discovery and Data Mining. London (UK): 1998. 7/1—7/4.

[5] STEEl J A, MCDONALD J R, ARCY C D. Knowledge Discovery in Databases: Applications in the Electrical Power Engineering Domain. In: Proceedings of IEE Colloquium on IT Strategies for Information Overload. London (UK): 1997. 8/1—8/4.

[6] PITT B D, KIRSCHEN D S. Application of Data Mining Techniques to Load Profiling. In: Proceedings of the 21st 1999 IEEE International Conference Power Industry Computer Applications PICA '99. Santa Clara (CA): 1999. 131—136.

[7] MADAN S, SON W K, BOLLINGER K E. Applications of Data Mining for Power Systems. In: Proceedings of IEEE 1997 Canadian Conference on Electrical and Computer Engineering, Engineering Innovation: Voyage of Discovery. Newfoundland (Canada): 1997. 403—406.

[8] FAYYAD U, SHAPIRO G P, SMYTH P. Advances in Knowledge Discovery and Data Mining. Menlo Park (CA): AAAI Press; Cambridge (MA): The MIT Press,1996.

[9] HAN J, KAMBER M. Data Mining: Concepts and Techniques. San Francisco (CA): Morgan Kaufmann Publishers, 2001.

[10 TSO S K, LIN Jikeng, HO H K et al. Data Mining for Detection of Sensitive Buses and Influencial Buses in a Power System Subjected to Disturbances. IEEE Trans on Power Systems, 2004, 19(1): 563—568.

[11] KAUFMAN L, ROUSSEEUW P J. Finding Groups in Data: An Introduction to Cluster Analysis. New York: John Wiley & Sons Inc, 1990.

[12] VAN WIJK J J, VAN SELOW E R. Cluster and Calendar-based Visualization of Time Series Data. In: Proceedings of IEEE Symposium on Information Visualization. Vienna (Austria): 1999. 4—9.

林济铿(1967—),男,博士,副教授,研究方向为电力系统稳定性分析及控制、电力市场、人工智能在电力系统中的应用。E-mail: mejklin@dt.com.cn

罗萍萍(1969—),女,硕士,讲师,研究方向为人工智能在电力系统中的应用、继电保护。

曹绍杰,男,博士,讲座教授,研究方向为工业系统评价、工业自动化及智能控制。

Study on Similarity of Load Profiles Following Disturbances Based on Data Mining

LIN Ji-keng[1], *LUO Ping-ping*[2], *S. K. TSO*[3], *C. M. MAK*[4], *K. M. YUNG*[4]

(1. Tianjin University, Tianjin 300072, China)

(2. College of Electrical Engineering of Shanghai, Shanghai 200437, China)

(3. City University of Hong Kong, Hong Kong)

(4. The CLP Power, Hong Kong, Hong Kong)

Abstract: Various disturbances including faults have conspicuous influence on variations of system load. It is very useful to acquire better understanding of the influence. This paper applies data-mining techniques to the CLP Power database to analyze the cluster of load profiles in various buses in response to disturbances. The bus group or region, whose load profiles are very similar with that of the special station (station AAA) at CLP system are achieved. Therefore, it could be more accurately to assess and reduce the loss and influence to these loads resulted from the future disturbance (such as adopting same measure or strategy to prevent or deal with the disturbance for these loads), which is very helpful to enhance the security and stability operation of power system.

Key words: data mining; disturbance; load profile; cluster analysis