doi: 10.7690/bgzdh.2025.01.010

基于决策粗糙集的多源异构电网数据整合方法

郭　良，郑晓斌，段春明，王　卓，鲁　兵
（国网冀北电力有限公司工程管理分公司，北京 100070）

摘要：为解决多源异构电网数据整合耗时较长的问题，提出应用决策粗糙集的数据整合模型。通过纵向融合、横向融合 2 个环节，完成多源异构电网数据的融合处理，再应用并行化正向最大匹配去冗算法，删除融合数据中存在的重复冗余信息；依托决策粗糙集对电网数据进行属性约简，去除具有干扰作用的噪声条件属性，再从约简决策表内提取模糊分类规则，实现电网数据的分类整合管理；创建基于数据关联度的整合数据修复方案，完成数据整合模型的设计。实验结果表明：应用所提模型对 43 700 条多源异构电网数据进行整合处理，所需的数据整合时间为 5.9 s，符合实时性要求。

关键词：决策粗糙集；异构数据；电网；数据整合；正向最大匹配原则；关联度
中图分类号：TP311　　**文献标志码：**A

Data Integration Method of Multi-source Heterogeneous Power Grid Based on Decision Rough Set

Guo Liang, Zheng Xiaobin, Duan Chunming, Wang Zhuo, Lu Bing
(*Engineering Management Branch*, *Jibei Electric Power Co*., *Ltd*. *of State Grid*, *Beijing* 100070, *China*)

Abstract: In order to solve the time-consuming problem of multi-source heterogeneous power grid data integration, a data integration model based on decision rough set is proposed. Through vertical fusion and horizontal fusion, the fusion processing of multi-source heterogeneous power grid data is completed, and then the parallel forward maximum matching redundancy removal algorithm is applied to delete the redundant information in the fusion data; Based on decision rough set, the attribute reduction of power grid data is carried out to remove the noise condition attributes with interference, and then the fuzzy classification rules are extracted from the reduction decision table to realize the classification and integration management of power grid data. The integration data repair scheme based on data correlation is created to complete the design of data integration model. The experimental results show that the proposed model is applied to the integration of 43 700 multi-source heterogeneous power grid data, and the required data integration time is 5.9 s, which meets the real-time requirements.

Keywords: decision rough set; heterogeneous data; power grid; data integration; forward maximum matching principle; correlation degree

0　引言

电力系统生产运行的核心就是调度中心，该环节存在管理数据、业务数据、生产数据等多方面信息，这些信息是国家电网发展的依据[1]。尤其在电力大数据战略广泛推进后，调度中心的每个部门都处在独立的工作环境中，使得不同电力子系统之间的数据交互难度大幅度提升[2]，无法发挥电网数据的价值。为更好地利用这部分电网数据，需要对多源异构数据进行有效整合，将分散的电网数据融合、重组到一起，实现数据决策分析能力的提升。

目前，研究人员越来越重视数据整合的研究，各种数据整合模型也开始得到应用。文献[3]采用了分布式微服务架构，设计了数据整合模块与数据安全应用模块，2 个模块同步工作，快速部署，实现轻量级的数据整合；实验结果表明，该方法应用稳定性较差。文献[4]运用了虚拟化技术，研究一种基于服务器端的数据整合框架。设置虚拟机和物理节点充当数据整合管理者，控制服务器端执行数据整合命令，形成符合实际当前节点状态的数据整合方案，但该方法的数据整合耗时较长。文献[5]深入分析了电网数据的特点，明确多源电网数据融合的层次要求和基本原理，定义基于映射的数据融合方案，实现异构电网数据的整合处理，但该方法应用后整合数据的冗余度较高。

考虑到文献提出的模型应用到多源异构电网数据整合无法发挥良好的应用性能。笔者提出将决策

收稿日期：2024-07-06；**修回日期：**2024-08-19
第一作者：郭　良(1983—)，男，山西人，硕士。

粗糙集算法应用到整合过程中，避免多源电网数据的整合受到噪声信息的干扰，确保异构电网数据的高效整合。

1　多源异构电网数据整合方法

1.1　多源异构电网数据融合机制构建

由于电网调度中心内的电网数据来自多个电力子系统，每个子系统在运行过程中为符合阶段性和技术性要求，会采用不同的数据编码规则和保存格式，使得最终电网数据呈现出多源异构的特点[6]。对这些数据进行整合处理时，需要先建立多源异构电网参数融合机制，通过横向融合、纵向融合 2 步，将数据有机融合在一起，具体的数据融合原理如图 1 所示。

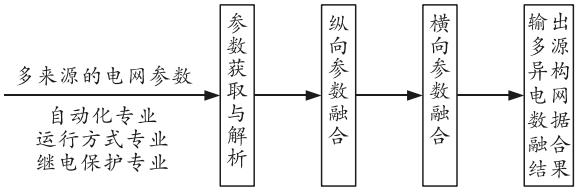

图 1　电网多源异构参数融合

根据上图可知：纵向融合是电网数据融合的第 1 个环节，主要是为消除同一部门的自动化设备、继电保护设备等设备的采集数据的差异性。实际操作过程中，先通过统一的 web 应用程序，自动获取每个数据源提供的电网参数，通过对不同来源的数据进行解析，再引入参数差异性计算公式，获取参数差异分析结果。

$$D=\left|X_1-X_2\right|\Big/\overline{X}\times 100\%。\tag{1}$$

式中：D 为参数差异度；X_1 为电网上级参数的数值；X_2 为电网下级参数的数值；$\overline{X}$ 为参数均值。

结合多源异构电网数据的复杂性，明确所有差异度较大的数据表达形式并不相同，笔者提出在纵向融合过程中，按照对应的转换规则，将多源数据表达形式变换为统一的有名值，完成纵向融合处理。

而后，再按照相似的操作步骤实现多源异构电网数据的横向融合，确保同一电力调度中心不同部门之间采集的数据有效融合在一起，去除电网数据的异构性，便于后续电网数据整合处理。

1.2　多源异构电网数据冗余处理

对于融合处理后的多源异构电网数据，应用一种并行化正向最大匹配去冗算法对多源异构电网数据进行冗余处理，去除原始电网数据中存在的冗余信息[7]。并行化正向最大匹配去冗处理的工作模式如图 2 所示，主要包括输入阶段、数据分解阶段、数据合并阶段和输出阶段。

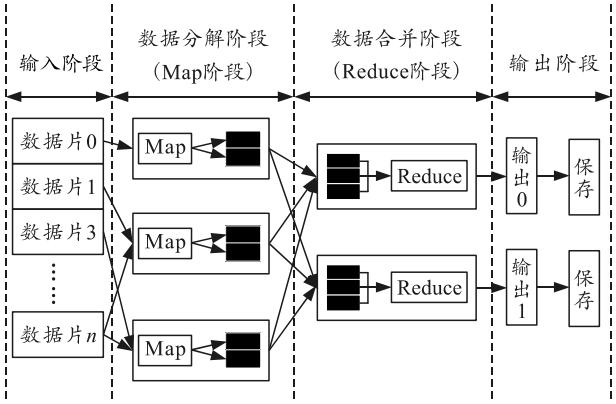

图 2　并行化正向最大匹配去冗机制

上图所示的去冗机制应用时，需要应用中央政府发布的电网通用模型描述规范，获取首字索引结构，形成合理的特征向量属性表，作为输入内容。

在数据分解阶段，采用滑动扫描的方式[8]，对输入数据进行全面分析，明确当前扫描的特征属性是否与特征向量属性表内容相匹配，若匹配则继续扫描，否则需要将其添加至特征向量属性表内。

匹配结束后，将数据分解结果保存在内存缓冲区，直到该区域数据到达临界值后，建立一个溢出文件，将缓冲区数据导入其中，遍历文件内的数据，将具有相同特征向量的数据发送至同一节点，实现分解数据的重组。在重组的过程中，Reduce 函数会自动对冗余特征向量进行不断过滤[9]，直到最后一个数据遍历完成，生成一个完成分区和排序的去冗数据文件。

1.3　基于决策粗糙集的数据整合模型构建

为便于电网数据的应用，笔者建立的数据整合模型以数据分类管理为核心。在电网数据融合和去冗完成后，引入决策粗糙集理论，以单输出分类为目标，建立电网数据类别标签集，通过模糊分类处理的方式，将不同主题的数据整合在一起。

笔者应用决策粗糙集算法，对待整合电网数据进行属性约简，将约简后的数据作为模糊分类的输入变量。采用离散化算法处理所有电网数据，结合聚类中心、最大隶属度原则，确定电网数据的连续值有序关系。基于离散化后的电网数据进行属性约简处理。在决策粗糙集算法应用时，先针对电网数据定义决策表五元组：

$$S=\left[A,\varphi\cup F,V,H\right]。\tag{2}$$

式中：S 为决策表五元组；A 为论域；φ 为条件属性；F 为决策属性；V 为属性的值域；H 为信息函数。

针对离散化处理后的电网数据进行分析，建立可用于约简的属性子集，且该子集满足以下 2 个条件：

$$\left.\begin{aligned}&\left|P_g^{(\alpha,\beta)}(F)\right|\geqslant\left|P_\varphi^{(\alpha,\beta)}(F)\right|\alpha,\beta\\&\forall\varepsilon\in g,\ \left|P_{g-\{\varepsilon\}}^{(\alpha,\beta)}(F)\right|\leqslant\left|P_g^{(\alpha,\beta)}(F)\right|\end{aligned}\right\}。\tag{3}$$

式中：g 为属性子集；ε 为属性；P 为正域；α,β 为阈值。

此外，对于决策属性和条件属性来说，前者对后者的依赖度可以表示为：

$$Y_\varphi^{(\alpha,\beta)}(F)=\left|P_\varphi^{(\alpha,\beta)}(F)\right|\Big/|A|。\tag{4}$$

式中 Y 为依赖度。

某个电网数据属性对于整个属性子集的重要度可以表示为：

$$\chi_{(g,F)}^{(\alpha,\beta)}(\varepsilon)=Y_{g\cup\varepsilon}^{(\alpha,\beta)}(F)-Y_g^{(\alpha,\beta)}(F)。\tag{5}$$

式中 χ 为重要度。

式(5)计算出的重要度结果，描述了电网数据某个特征属性的重要程度。笔者运用贪心式启发搜索策略，作为决策粗糙集约简实现的搜索模式，该约简搜索模式主要包括 2 个环节：1) 在重要度计算结果的引导下，应用前向贪心搜索策略遍历所有属性集，去除重要度低于预先设定阈值的属性；2) 运用后向搜索策略[10]，从空集开始分析每个电网数据特征属性的依赖度，剔除依赖度较低的属性值，完成离散电网数据的属性约简处理。

而后，建立一个基于决策粗糙集的模糊分类模型，以某个样本点作为起始点，引出横坐标，并以该样本的隶属度作为纵坐标，绘制隶属度函数拟合图像。根据聚类结果建立多个模糊子集，结合曲线拟合策略明确不同隶属度函数对应的参数值，生成与每个输入变量相符的模糊子集。根据电网数据属性约简结果，建立一个不包含冗余信息的属性集决策表，从而生成如下所示决策规则：

$$\begin{aligned}R^k:&(\varepsilon_1,\lambda_1)\wedge(\varepsilon_2,\lambda_2)\wedge\cdots\wedge(\varepsilon_m,\lambda_m)\rightarrow\\&(d,w_1)d_1^k\wedge\cdots\wedge(d,w_m)d_m^k。\end{aligned}\tag{6}$$

式中：R^k 为第 k 条决策规则；λ 为属性值；m 为待整合电网数据的属性数量；d 为置信度；w 为属性权重。

为简化数据整合步骤，笔者提出对式(6)所示的决策规则前件进行离散化处理，获取模糊分类规则，此时模糊规则后件的置信度可表示为：

$$\bar{d}_o^k=\sum_{s=1}^n\boldsymbol{\eta}^{(s)}\alpha^k(\boldsymbol{\eta}^{(s)})\Big/\sum_{s=1}^n\alpha^k(\boldsymbol{\eta}^{(s)})。\tag{7}$$

式中：$\bar{d}$ 为模糊规则后件的置信度；O 为电网数据类别；s 为数据；n 为多源异构数据集；$\boldsymbol{\eta}$ 为输入向量。

笔者依托“胜者为王”的理念，确定电网数据的类别判别函数：

$$o=\arg\max_o^N\{z_o(\boldsymbol{\eta})\}。\tag{8}$$

式中：z 为判别函数；N 为电网数据类别总数量。通过上述计算，将待处理的多源异构电网数据按类进行整理，完成电网数据的基本整合处理。

1.4 电网整合数据修复方案创建

通过上述处理完成电网数据的初步整合处理，为更好地发挥数据价值，笔者应用数据关联度分析理论，建立一个数据修复方案，对整合后的多源异构电网数据进行修复，使得每个管理类别内的电网数据具有一致性。实际操作过程中，对于 2 个整合处理后的电网数据类来说，其特征向量可以表示为：

$$\left.\begin{aligned}\boldsymbol{\varpi}_J&=(W_{J1},W_{J2},\cdots,W_{J\partial})\\\boldsymbol{\varpi}_I&=(W_{I1},W_{I2},\cdots,W_{I\partial})\end{aligned}\right\}。\tag{9}$$

式中：J, I 为 2 个系统；$\boldsymbol{\varpi}$ 为领域概念；$\boldsymbol{W}$ 为特征向量；∂ 为特征向量数量。

式(9)所示的 2 个特征向量之间的关联度，可以定义为：

$$\psi(\boldsymbol{\varpi}_J,\boldsymbol{\varpi}_I)=\cos(\boldsymbol{\varpi}_J,\boldsymbol{\varpi}_I)。\tag{10}$$

式中：ψ 为关联度；cos 为余弦函数。

式(10)所示的关联度计算结果，描述了相同类别内的电网数据属性关联程度，余弦值靠近 1 时，表明当前选定的 2 个特征向量之间存在较低的关联度，需要分别进行保存。而对于少部分属性不一致的情况，表示此时数据整合结果不合理，将不符合要求的数据误归纳至对应的集合中，数据修复处理正是以处理这种情况为核心。实际处理过程中，需要针对一个特征子集制定一个目标值，基于此对不一致属性值进行不断更新，直到余弦值计算结果接近 1。

除此之外，整合数据修复方案实施时，需要满足 3 项规则：1) 数个特征向量的属性描述出现重复情况后，需要任选其一保留下来，避免出现整合数据重复的问题。2) 在属性描述不一致的情况下，需要依托于关联度计算结果，保留新采集的电网数据，

舍弃原有数据，保证整合后的数据具有更大的利用价值。3) 当出现某个特殊的特征向量时，直接进行保留。通过上述处理，对基于粗糙决策集整合后的数据进一步处理，这也是数据整合模型的最后一个组成模块。

2　实验

笔者针对电网数据整合问题进行研究，提出一种结合粗糙决策集的整合模型，为体现该模型的应用性能，需要结合文献[2]模型、文献[3]模型进行实验对比分析，以数据整合时间为评价指标，体现笔者设计模型的优越性。

2.1　实验准备

以某城市的电网为例，截选 2022 年 5 月 1 日到 2022 年 9 月 30 日的电网数据，充当本次实验所需的基础数据。而经过调查可以发现，这些实验数据来自电力调度中心调度运行管理系统和安全管理系统，且除了异构数据外还具有部分损坏数据，用以验证所提整合模型的合理性。实验数据基本情况如图 3 所示。

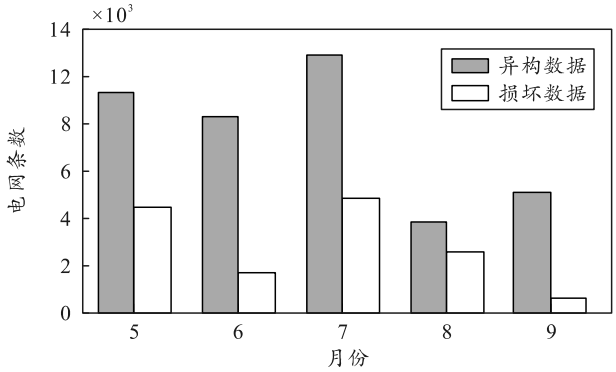

图 3　实验数据统计

本次实验基于 Matlab 软件实现，将采集的多源异构电网数据导入仿真模拟软件内，形成实验文件夹，作为数据整合测试的基础。本次实验搭建的测试平台如图 4 所示，主要包括实验数据显示器和实验控制终端 2 个核心设备，且这 2 个设备分别与一个服务器相连接。如图 4 所示，先将实验数据输入至实验数据显示器，再由此发送至实验控制终端，支持数据整合实验的过程实现。

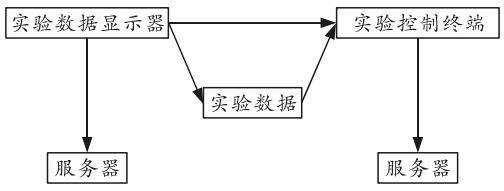

图 4　实验测试平台搭建结果

2.2　整合结果

考虑到笔者设计数据整合模型的关键内容，在于应用粗糙决策树算法进行电网数据分类整合管理。操作过程涉及一个关键参数就是阈值，其取值直接影响数据属性约简性能。在电网数据整合实验开始之前，采用常见的分类算法获取不同阈值取值条件，数据分类精度变化情况，如图 5 所示。

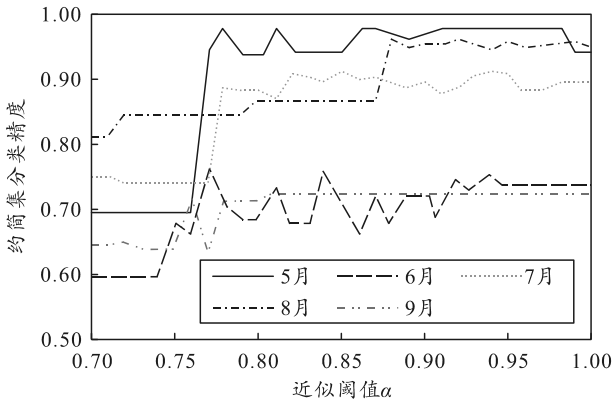

图 5　阈值与约简集分类精度的关系

根据上图可知：阈值的取值变化，会对电网数据约简属性集的分类精度产生一定影响，但这种影响的规律性并不明显，阈值取值的增加不一定会使得约简集分类精度提升。为保证电网数据整合效果更好，确定最高分类精度对应的阈值分别为 0.77、0.77、0.82、0.89 与 0.81，作为每个月份的多源异构电网数据整合处理过程中应用的阈值。

结合电网业务划分情况，将电网数据整合归类到 9 个主题下，按照笔者研究内容进行数据去冗、修复处理后，确定每个主题下包含的数据量如表 1 所示。

表 1　电网数据整合主题

大类	主题	数据来源	数据量
电网设备	电网设备	中心调度运行管理系统 中心调度安全管理系统	10 254
电网运行	一次能源	中心调度运行管理系统 中心调度安全管理系统	2 576
	量测	中心调度运行管理系统 中心调度安全管理系统	7 213
电网调度	调度指令票	中心调度运行管理系统 中心调度安全管理系统	3 197
	设备检修单	中心调度运行管理系统 中心调度安全管理系统	3 250
	电力电量平衡	中心调度运行管理系统 中心调度安全管理系统	4 750
	事故信息	中心调度运行管理系统 中心调度安全管理系统	5 510
	日志	中心调度运行管理系统 中心调度安全管理系统	3 500
	人员管理	中心调度运行管理系统 中心调度安全管理系统	3 500

对比上表所示的数据整合结果和图 3 所示的实验数据统计图可以发现：二者显示的多源异构数据量相差无几，表明所提数据整合模型应用后，可以有效去除实验数据集内的破坏数据，并按照主题对电网数据进行划分，得到合理的数据整合结果。

2.3 模型性能对比

本次实验过程中，还同时应用了基于人工智能的数据整合模型、基于微服务的数据整合模型，分别针对实验数据进行处理，记录不同模型面对不同电网数据规模的整合时间，绘制如图 6 所示。

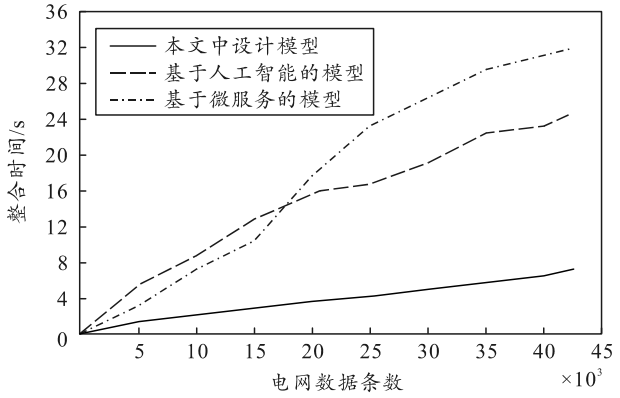

图 6 不同模型的数据整合时间对比结果

根据上图可知：随着电网数据量的提升，模型的数据整合时间逐步提升，但基于粗糙数据集的整合模型，整合时间提升幅度明显更低。当所有实验数据整合完成后，笔者设计模型的整合时间为 5.9 s，其他 2 种方法的最终整合时间分别为 24.3、31.1 s。综上所述，笔者研究数据整合模型的应用，极大地提高了电网数据的整合效率。

3 结束语

智能电网的不断发展，使得电力系统运行产生的数据量成倍增长。为更好地处理越来越多的多源异构电网数据，笔者提出应用粗糙数据集算法，设计一种新的数据整合模型，通过融合、去冗、分类、修复等多个环节，实现多源异构电网数据的高效整合管理。

参考文献：

[1] 刘文君, 董明, 徐元孚, 等. 电力设备运行状态大数据标签体系与关键技术[J]. 中国电力, 2022, 55(1): 126−132.

[2] 许洪强, 蔡宇, 万雄, 等. 电网调控大数据平台体系架构及关键技术[J]. 电网技术, 2021, 45(12): 4798−4807.

[3] 杨舒, 苏放. 基于微服务的分布式数据安全整合应用系统[J]. 计算机工程与应用, 2021, 57(18): 238−247.

[4] 刘孙发, 林志兴. 基于虚拟化技术的服务器端数据整合系统设计研究[J]. 现代电子技术, 2020, 43(2): 77−79, 83.

[5] 覃松涛, 黄超, 田君杨, 等. 电网多源大数据融合方法的研究与应用[J]. 电子器件, 2021, 44(2): 480−485.

[6] 杨漾, 敖知琪, 刘佳, 等. 面向数字电网的基于容器技术的边缘计算数据处理机制[J]. 南方电网技术, 2021, 15(5): 98−103.

[7] 孙利宏. 基于 Hadoop 的智能电网时序大数据处理方法[J]. 计算机仿真, 2020, 37(12): 67−71.

[8] 钟雅婷, 林艳梅, 陈定甲, 等. 多组学数据整合分析和应用研究综述[J]. 计算机工程与应用, 2021, 57(23): 1−17.

[9] 穆肃, 崔萌, 黄晓地. 全景透视多模态学习分析的数据整合方法[J]. 现代远程教育研究, 2021, 33(1): 26−37, 48.

[10] 崔金栋, 王胜文, 辛业春. 区块联盟链视角下智能电网数据管理技术框架研究[J]. 中国电机工程学报, 2020, 40(3): 836−848.