

doi: 10.7690/bgzd.2025.07.020

基于 Apriori 算法的供电公司营销数据挖掘系统设计

张 剑¹, 刘 畅², 杨 逸², 魏昕喆², 张 浩², 王 旭³

(1. 国网天津市电力公司, 天津 300010; 2. 国网天津城东供电分公司, 天津 300250;
3. 天津市普迅电力信息技术有限公司, 天津 300000)

摘要: 为解决供电公司营销数据量大, 影响数据频繁项集处理效率的问题, 设计一种基于 Apriori 算法的供电公司营销数据挖掘系统。硬件设计通过营销数据挖掘系统物理架构部署, 搭建系统硬件环境, 实现数据库信息的同步; 软件方面设计电力营销数据仓库, 采用 Apriori 算法通过映射剪枝处理频繁项集, 挖掘关联规则, 建立多维数据挖掘模型, 实现系统的数据挖掘功能。经实验论证分析, 结果表明: 该系统在电力负荷预测应用中的预测结果与实际值相差较小, 在最小支持度和事务数据量条件下, 数据挖掘执行时间分别在 2 和 10 s 以下, 具有较高的执行效率, 说明该系统是可行的。

关键词: Apriori 算法; 供电公司; 服务器; 营销数据挖掘系统; 关联规则; 数据仓库

中图分类号: TP311.12; TP391 **文献标志码:** A

Design of Power Supply Company Marketing Data Mining System Based on Apriori Algorithm

Zhang Jian¹, Liu Chang², Yang Yi², Wei Xinzhe², Zhang Hao², Wang Xu³

(1. State Grid Tianjin Electric Power Company, Tianjin 300010, China;
2. State Grid Tianjin Chengdong Power Supply Branch, Tianjin 300250, China;
3. Tianjin Puxun Power Information Technology Co., Ltd., Tianjin 300000, China)

Abstract: In order to solve the problem that the large amount of marketing data of power supply company reduces the processing efficiency of data frequent itemsets, a marketing data mining system of power supply company based on Apriori algorithm was designed. The hardware design deploys through the physical architecture of the marketing data mining system, builds the hardware environment of the system, and realizes the synchronization of database information; the software design of the power marketing data warehouse, uses Apriori algorithm to process frequent itemsets through mapping and pruning, mines association rules, establishes multi-dimensional data mining model, and realizes the data mining function of the system. The experimental results show that the difference between the predicted value and the actual value of the system in the application of power load forecasting is small. Under the conditions of minimum support and transaction data volume, the execution time of data mining is less than 2 s and 10 s respectively, which has high execution efficiency and shows that the system is feasible.

Keywords: Apriori algorithm; power supply company; server; marketing data mining system; association rules; data warehouse

0 引言

营销数据的及时性、准确性为供电公司增效提供了重要的保证, 也是确保数据质量、实现闭环管理、提高供电公司管理水平的重要基础。在当前电力市场环境下, 供电公司要打通业务系统壁垒, 以高质量数据支撑业务应用, 充分挖掘数据的价值, 需要加强对科学技术手段的应用^[1]。目前的电力营销数据系统已无法满足供电公司提质增效的要求, 为避免重复工作, 使一线员工减负增效, 需要提出与之协同发展的服务系统, 促进供电公司营销、配电等业务质量的提升。

文献[2]设计基于 C5.0 决策树算法的电力营销数据挖掘系统: 首先, 对数据挖掘中先进的 C5.0 决策树算法原理进行分析, 并通过引入信息熵对原有的属性选择方式进行改进, 提高了信息增益比率计算的速度; 然后, 根据设计的售电量关系模型进行对电厂管理信息系统中的数据进行挖掘, 但该系统在电力数据预测应用中的预测结果与实际值相差较大。文献[3]设计基于电力营销聚类分析的数据挖掘系统, 先运用聚类算法筛选数据, 再计算数据结构和相异度矩阵, 得出最接近的类距离, 在聚类分析框架下, 设计聚类分析数据挖掘算法流程, 为此

实现电力营销数据挖掘, 但该系统在进行供电公司营销数据挖掘时, 由于数据量大, 数据频繁项集处理时间较长, 导致系统执行效率较低。

Apriori 算法能够通过对频繁项目集的关联规则挖掘, 得到较为理想的最终目标。为此, 针对上述系统存在的问题, 笔者基于 Apriori 算法, 对供电公司营销数据挖掘系统进行设计, 对供电公司更好地利用数据, 获取更高价值具有重要意义。

1 营销数据挖掘系统物理架构部署

硬件是软件系统的基础支撑, 本文中系统基于原有营销数据核查应用 APP 部署方式, 依托于供电公司局域网, 采用分布式部署服务器, 通过分布式算法, 使系统处理性能呈指数级别的增长, 在服务器的选择上, 除了数据库服务器, 还需要进行备份服务器的配置^[4]。具体服务器配置与系统硬件部署情况如表 1 所示。

表 1 系统服务器配置与硬件部署情况

类型	配置信息	台数
数据库	8 核处理器、32 GB 内存、500 GB 以上硬盘(复用营销数据核查应用 APP 资源, 无需单独提供)	2
应用服务器	8 核处理器、32 GB 内存、500 GB 以上硬盘(复用营销数据核查应用 APP 资源, 无需单独提供)	4
缓存服务	Redis 集群 16 GB(复用营销数据核查应用 APP 资源, 无需单独提供)	1
分布式存储	DSS 容量 500 GB(复用营销数据核查应用 APP 资源, 无需单独提供)	1
网络	服务器间连接: 12 000 Mbps; 现场作业带宽: 60 M 总部出口带宽: 独享、双线、160 M	-
存储设备	型号: FC-SP 8 Gb; 接口类型: AL 单盘容量: 400 G; raid 50 阵列; 大于 12 个	3*N(N≥4)

由表 1 可知: 本文中系统基础的环境需求, 涉及空间数据分析, 笔者采用 Oracle 数据库。其中, 数据库文件由数据库服务器管理, 对于数据的备份工作则由备份服务器负责, 并与其他服务器进行信息的同步设置, 若遇突发状况, 备份服务器迅速做出响应, 保存数据一个时间点的数据, 待服务器修复完成后, 与主数据库进行信息的还原和对比。由于计算机损坏存在着不确定性, 因此对备份服务器的要求比主服务器的要求更为严格, 在备份数据处需要进行多次测试, 确保数据的安全, 避免数据的丢失对工作人员正常工作造成较大的影响。

2 营销数据挖掘系统软件设计

2.1 设计电力营销数据仓库

构建电力营销信息数据仓库是实现电力营销数据挖掘系统的先决条件, 设计电力营销数据仓库能

够将分散的电力营销全部原始业务数据进行收集和整理, 决定了数据查询的效率, 便于对现场作业的全面掌握和有效监管及指导, 将分散的代码表、目录电价等进行统一规范化的集中用电管理, 为原始数据的数据挖掘分析功能奠定基础^[5]。在数据仓库的设计中, 要求完成各方面复杂数据的处理, 实现对原始数据进行清洗、过滤等操作。

本文中数据仓库看作特殊的关系数据库, 由不同地区的供电单位数据库集成的数据为数据仓库的业务数据, 除此以外, 数据仓库中数据还包括系统外部的数据库来源, 这些数据需要通过网络连接的方式经过数据转换和净化接口才能进入数据仓库; 因此, 数据仓库负责网络连接的各項应用, 通过系统内的数据库以及相应的外界数据源, 进行电力营销数据的抽取, 具体电力营销数据抽取策略如图 1 所示。

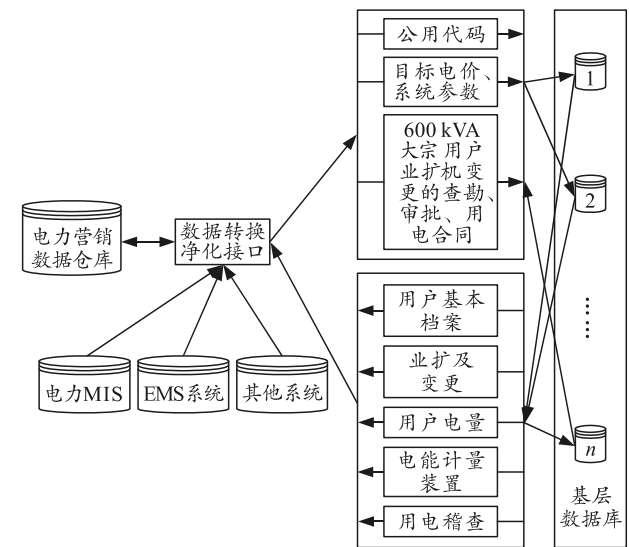


图 1 电力营销数据抽取策略

由图 1 可知: 对于数据库互联的数据, 可通过定时复制技术收集, 网络连接中的应用可以通过 DBMS 来实现对应的支持, 令客户端得以形成所需求的 SQL 代码, 在服务器完成预期的执行操作。采用事务复制, 保证所有变化按照生成时的顺序, 发送到订购数据库中, 复制时间的设置可以为每天一次, 对于不同的数据库设置不同的复制时间, 合理分配以避免出现冲突状况^[6]。本文中的数据仓库结构设计为 3 层, 中间层服务器能够通过该层级将多个维度数据的控制程序的处理转变为标准关系操作。顶层为负责数据查询、报告、分析和数据挖掘功能的工具, 为配套的查询功能, 通过客户展示层实现指标的基础分析等操作。

2.2 基于 Apriori 算法的关联规则挖掘

笔者利用 Apriori 算法进行关联规则的挖掘, 首先对 2.1 节设计的电力营销数据库进行扫描, 将所有数据集记作 $R=\{R_1, R_2, \dots, R_n\}$, X 、 Y 为项目集, 其中项目的数目为 n , 事务数据库集合记为 $D=\{T_1, T_2, \dots, T_n\}$, 其中有 $T_i=\{R_{i1}, R_{i2}, \dots, R_{in}\}$, T_i 中的任意元素 $R_{ij}(j \in [1, n]) \subseteq R$, 则称 T_i 为数据库中的事务^[7]。接下来对支持度进行计算, 具体公式为:

$$\text{supp}(A \Rightarrow B) = \frac{|\{T: A \cup B; T, T \in D\}|}{|D|}。 \quad (1)$$

式中 $A \Rightarrow B$ 的形式可表示为关联规则, A 为关联前件, B 为关联后件。Apriori 算法利用数据迭代, 通过频繁项集的挖掘, 进一步在候选项集的基础上, 实现对关联规则的挖掘。扫描到的项目的集合即为第 1 个候选项集, 电力营销数据由于来源较广, 数据量大, 为提高数据处理效率, 避免数据缺失和数据资源的浪费, 笔者进行了 Apriori 算法步骤的设计, 先通过映射进行剪枝, 缩短了数据频繁项集处理时间, 提高了系统执行效率, 然后设计目标事物库 D 的存储结构, 定义目标数据库中具有重复事务的个数, 具体公式为:

$$V_i = \left\{ \begin{array}{l} V_i + 1, \text{ if } T_{S_j} = T_{T_j} (S \neq T) \\ 1, \text{ if } T_{S_j} \neq T_{T_j} (S \neq T) \\ J = 1, 2, \dots, n; S, T = 1, 2, \dots, m \end{array} \right\}。 \quad (2)$$

式中 S_j 为目标数据库中事务包含的项集数, 若 V_i 为 1, 则说明重复事务不存在, 若 V_i 为 +1, 则表示存在重复事务。在映射阶段, 将事务库中的项集读入为关键键入, 认定数值为 1 作为输出值, 在剪枝时, 以项集为关键键入, 数值 1 作为输入值。各项集出现的概率不同, 统计各项集的出现概率, 根据最小值尺度, 判断是否为频繁项集, 若 k 频繁为空, 可以结束整个迭代过程, 由此通过映射来压缩目标事务库, 实现对频繁项集的预先瘦身和剪枝。接着计算置信度, 具体公式为:

$$\text{conf}(A \Rightarrow B) = \frac{|\{T: A \cup B; T, T \in D\}|}{|\{T: A; T, T \in D\}|}。 \quad (3)$$

通过置信度验证数据的合法性, 若置信度较高, 则说明关联前件可以较为准确地推出关联后件, 根据最小置信度, 生成数据的关联规则, 实现对关联规则的挖掘^[8]。

2.3 建立多维数据挖掘模型

根据挖掘的关联规则, 构建多维数据挖掘模型,

实现本文中系统的电力营销数据挖掘功能。首先进行模型的变量选择, 模型的变量选择就是对数据挖掘空间的选择, 由于各变量之间的关系较为复杂, 所以必须将影响较弱的变量进行剔除。同时, 不合理的变量选择还会直接影响结果的准确性和可靠性, 在充分考虑上述问题后, 笔者选择神经网络方法来确定变量。首先对单隐层前向神经网络的输出估计进行表示, 具体为:

$$\hat{y}(H, W) = f_2(w_0^2 + \sum_{j=1}^l w_{ij}^{(1)}(w_{i0}^{(1)} + \sum_{j=1}^u w_{ij}^{(1)} h_j) w_i^{(2)})。 \quad (4)$$

式中: l 为隐藏节点数; j 为输入变量个数; $H=(h_1, \dots, h_u)^T$; $w_i^{(1)}=(w_{i0}^{(1)}, w_{i1}^{(1)}, \dots, w_{iu}^{(1)})^T$; 为各个输入与隐层节点 i 之间的连接权重; W 为全部网络权值; $f_2(\bullet)$ 为输出层转换函数; $f_1(\bullet)$ 为隐层转换函数^[9]。根据不同的 $f_1(\bullet)$ 和 $f_2(\bullet)$ 的选取形式, 在 $f_1(h)=1/(1+e^{-h})$, $f_2(h)=h$ 时, 得到简化的公式为:

$$\hat{y}(H, W) = w_0^2 + \sum_{i=1}^l f_1(w_{i0}^{(1)} + \sum_{j=1}^u w_{ij}^{(1)} h_j) w_i^{(2)}。 \quad (5)$$

利用贡献分析法进行计算决定系数, 具体公式为:

$$E = \frac{\text{Var}(\hat{y})}{\text{Var}(y)} = \frac{\sum_{i=1}^l w_i^{(2)} \text{Cov}(f_1(H^T w_i^{(1)}), \hat{y})}{\text{Var}(y)}。 \quad (6)$$

从方差分析角度理解决定系数, 得到第 i 个隐层节点对输出的贡献系数和第 j 个输入对第 i 个隐层节点的贡献系数, 具体为:

$$\left. \begin{array}{l} G_i = \frac{\text{Cov}(f_1(H^T w_i^{(1)}), \hat{y})}{\text{Var}(y)} w_i^{(2)} \\ U_{ij} = \frac{\text{Cov}(H^T w_i^{(1)}, h_j)}{\text{Var}(H^T w_i^{(1)})} w_{ij}^{(1)} \end{array} \right\}。 \quad (7)$$

式中: $i=1, 2, \dots, l$; $j=1, 2, \dots, u$, 最终得到各个输入变量对输出的总体贡献, 具体为:

$$Q_j = \sum_{i=1}^l U_{ij} G_i。 \quad (8)$$

式中: G_i 为第 i 个隐层节点对输出的贡献系数; U_{ij} 为第 j 个输入对第 i 个隐层节点的贡献系数。通过上述各式计算完成数据挖掘空间的选择^[10]。对第 i 维的贡献系数进行求解, 根据各维的贡献分量大小, 确定主聚焦维, 由此完成模型构建, 实现系统的数据挖掘功能。

3 实验论证分析

3.1 实验环境搭建与数据准备

为验证设计系统的有效性,笔者对系统进行了功能与性能测试,模拟真实运行数据,在与实际环境数据量接近的数据下实现测试。在实验开始前模拟真实环境,按照系统要求进行服务器的配置,测试客户端采用 Windows 系统的台式机,使用一台用户机器运行访问脚本,利用 Linux 机制下的守护进程对系统不停的发出请求,查看系统响应时间以及数据挖掘计算的平均时间。性能测试工具为 mercury 公司 winrunner、loadrunner。

在系统功能测试中,利用笔者设计的数据挖掘系统进行挖掘任务,并选取文献[2]设计的基于 C5.0 决策树算法的数据挖掘系统和文献[3]设计的基于电力营销聚类分析的数据挖掘系统为对照系统,在相同实验条件下依次进行挖掘处理,对比不同系统的数据挖掘处理功能。功能测试中将各系统应用于电力负荷预测中,所选取的数据是负荷数据和天气数据,利用这些数据进行关联规则挖掘分析,得到天气和负荷预测之间的关联关系。数据为 2022 年 6 月 1 日至 2022 年 6 月 8 日的数据,所用数据如表 2 所示。

表 2 实验数据

日期	时刻	最高气温/℃	最低气温/℃	实时气温/℃	负荷量/kWh
06.01	00:20	20	8	31	49 352.21
06.01	01:30	20	8	30	50 421.04
06.01	02:40	20	8	32	52 989.36
⋮	⋮	⋮	⋮	⋮	⋮
06.01	23:30	15	8	30	66 877.45
06.01	23:50	15	8	30	71 611.55

根据各系统对表 2 数据的挖掘处理,得出受天气因素影响的电力负荷预测结果,进行结果对比分析。

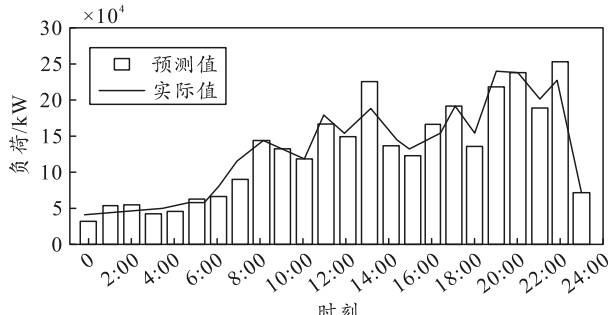
在性能测试中,测试不同系统在不同最小支持度和不同事务数据量情况下的数据挖掘执行时间,发送挖掘任务请求,给出最小支持度和最小置信度,在确定全局频繁项集后,分析各系统的执行效率。

3.2 功能测试结果

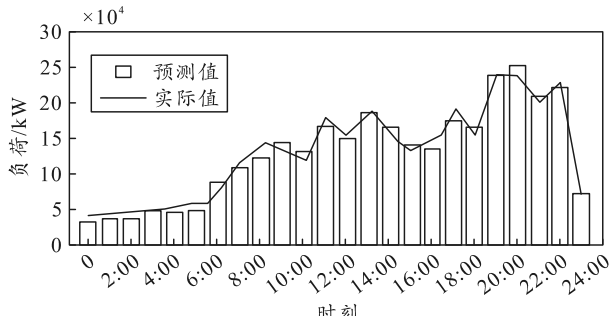
由不同系统数据挖掘处理进行的电力负荷预测结果如图 2 所示。

根据图 2 可以看出:天气变化对负荷值具有一定的影响,根据天气变化特征,电力负荷也出现了一定的规律变化。根据预测结果可知,基于 C5.0 决策树算法的数据挖掘系统在负荷预测应用中的预

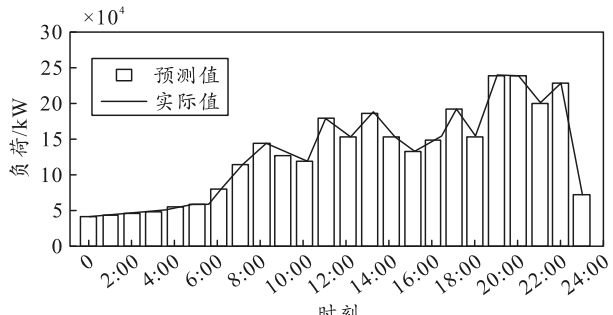
测结果与实际值相差更大,本文中系统应用得到的预测结果误差值最小;由此可见,本文中系统在具体功能应用中具有良好的效果。



(a) 基于 C5.0 决策树算法的数据挖掘系统



(b) 基于电力营销聚类分析的数据挖掘系统

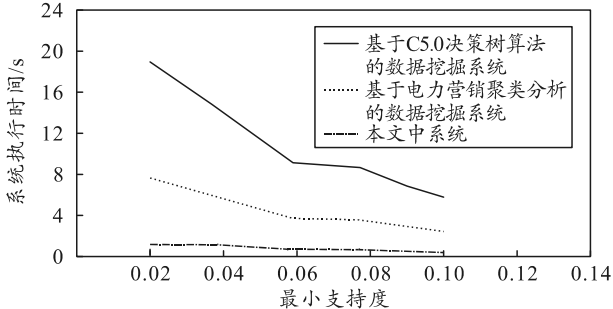


(c) 本文中系统

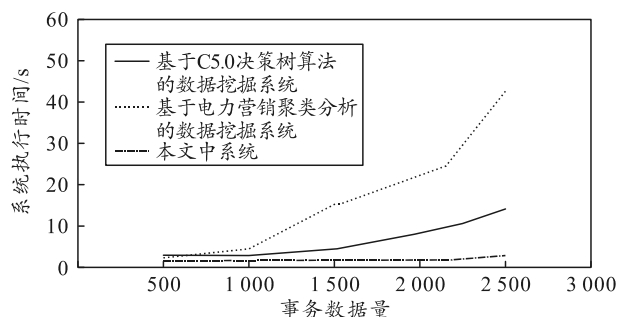
图 2 负荷预测结果对比

3.3 性能测试结果

设置最小支持度分别为 0.02、0.04、0.06、0.08 和 0.10;事务数据量分别为 500、1 000、1 500、2 000、2 500,对不同系统的数据挖掘执行时间进行分析,具体如图 3 所示。



(a) 最小支持度不同情况



(b) 事务数据量不同情况

图3 不同系统执行时间对比

图3清晰地展现了不同系统在不同实验条件下的数据挖掘执行时间。图3(a)中,当目标事务数据库中的事务数固定时,在最小支持度不断增大的情况下,本文中系统的执行时间呈小幅下降趋势,但执行时间始终在2s以下,用时最短,基于电力营销聚类分析的数据挖掘系统的执行时间有了大幅减少,但整体执行时间仍较其他系统时间长。图3(b)中,当最小支持度固定时,在事务数据量不断增大的情况下,不同系统的执行时间均有不同程度的增加,其中基于C5.0决策树算法的数据挖掘系统的执行时间变化最大,执行效率明显变差,本文中系统的执行时间较为稳定,均在10s以下,证明本文中系统所受影响较小,具备更好的执行效率。

4 结束语

在我国电力改革体制不断完善与深入下,目前的供电企业电力营销数据管理的信息系统已不能满足当下电力营销工作的要求,如何科学合理地制定与调整发电计划和电价对电力市场环境下的电力营销十分重要,为此笔者从供电公司的实际电力营销工作出发,采用Apriori算法,建立多维数据挖掘模型,设计了营销数据挖掘系统。笔者建立的多维数据挖掘模型可在短时间内预测电力负荷,在国家电

网优化电力配置的过程中具有一定的现实意义。在今后的研究中,笔者还将不断优化与完善系统设计,进一步提高供电公司营销领域的智能化,为营销服务质量的改善与用电的安全提供有力保证,提高该系统对电力营销方案设计的参考价值。

参考文献:

- [1] 江疆,梁盈威,彭泽武,等.基于大数据的电力系统数据集成管理平台设计[J].电子设计工程,2020,28(14):163-166.
- [2] 卜晓阳,蔡岩,王宗伟,等.基于C5.0决策树算法的电力营销数据挖掘[J].微型电脑应用,2022,38(1):23-26.
- [3] 臧玉魏,谢连科,张永,等.基于电力营销聚类分析的数据挖掘算法研究[J].信息技术,2020,44(4):56-59,64.
- [4] 程江洲,闫冉阳,冯梦婷,等.基于ACT-Apriori算法的电网故障诊断方法研究[J].电子测量技术,2021,44(24):32-39.
- [5] 王华满.基于数据挖掘技术的体育训练模式决策支持系统设计及应用[J].电子设计工程,2021,29(5):114-117,123.
- [6] 蔡榕,周长江,杨毅,等.基于数据挖掘的电力仪表图像智能监控系统设计[J].自动化与仪器仪表,2020(3):117-119,123.
- [7] 马强.基于关联规则的漏洞信息数据挖掘系统设计[J].现代电子技术,2020,43(5):82-85,91.
- [8] 朱孟周,黄强,孙志明,等.基于数据挖掘的电网作业安全管理系统设计[J].自动化与仪器仪表,2021(6):141-144.
- [9] 张凌,田珂,张巍,等.智能电网环境下电力营销管理系统的优化设计[J].计算技术与自动化,2020,39(4):138-141.
- [10] 熊蕾,彭吉琼,李铭,等.基于深度学习的基层网络数据个性化挖掘算法[J].计算机仿真,2022,39(1):318-321,332.