电力营销数据冗余检测与剔除方法

李　凯[1]　王晓军[2]　石多梅[1]
（1. 甘肃同兴智能科技发展有限责任公司　2. 国网甘肃省电力公司市场营销事业部）

摘要：为提高对电力营销数据冗余检测的精度，并将冗余数据全部剔除，提高数据质量，开展电力营销数据冗余检测与剔除方法设计研究。获取电力营销数据，并纠正电力营销数据编码。对电力营销数据进行循环冗余校验，引入 ISODATA 算法，剔除冗余数据。通过实例证明，新的方法可以将电力营销数据中的冗余数据全部找出并剔除，有效提高电力营销数据质量。

关键词：电力；营销；冗余检测；剔除；数据

0 引言

电力营销数据作为电力企业管理与决策的重要支撑，其准确性、完整性和高效性日益受到重视。电力营销数据涵盖用户用电信息、电费结算、市场趋势分析等多个方面，是电力公司优化服务、提升运营效率、制定市场策略的关键资源。然而，在实际的数据采集、传输与存储过程中，由于设备故障、数据传输错误、重复录入等多种因素，往往会产生大量的冗余数据，这些数据不仅占用了宝贵的存储空间，还可能干扰数据分析结果的准确性，影响管理决策的有效性。

近年来，学术界与工业界在数据冗余检测与剔除领域进行广泛的研究与实践，为电力营销数据的清洗与优化提供了重要参考。例如，谢绒娜[1]等人提出基于标签的数据流转控制策略冗余与冲突检测方法，通过引入标签机制，有效识别并解决了数据流在传输过程中的冗余与冲突问题，为提升数据处理的效率与准确性提供了新思路。然而，该方法主要聚焦于数据流转层面的控制，对于已存储于数据库中的历史冗余数据检测与清理尚显不足。

另一方面，商俊燕[2]等人利用 XGBoost 算法在无线传感器网络冗余数据检测中的应用，展示了机器学习在数据冗余识别方面的巨大潜力。通过构建高效的分类模型，该方法能够精准识别并剔除网络中的冗余数据，显著提升了数据质量。然而，电力营销数据具有复杂多变、规模庞大的特点，直接应用该方法可能面临计算资源消耗大、模型训练周期长等挑战，且需要针对电力营销数据的特定场景进行深度优化。

综上所述，现有研究成果在数据冗余检测与剔除方面已取得了一定进展，但仍存在一些不足，特别是在针对大规模、高复杂度的电力营销数据场景时，需要更加高效、智能且适应性强的解决方案。因此，本文将开展电力营销数据冗余检测与剔除方法研究。

1 电力营销数据编码纠正

优化电力营销数据编码的准确性，探索 CRC 技术在数据编码校正中的应用，以加速数据传输并保障

其质量。循环冗余校验（CRC）作为一种高效的线性编码机制，广泛应用于检测数据传输中的错误[3]。实施 CRC 编码方案涉及两个核心组件： CRC 校验位与有效数据段，合并构成总长度为：

$$N = K + R \tag{1}$$

式中，N 为二进制序列；K 为有效信息数据位；R 为 CRC 校验码位。将详细阐述 CRC 校验码的生成流程，旨在通过这一技术提升数据处理的精确性和效率。

在电力营销的数据传输中，有效信息通过数据多项式 $M(x)$ 来表示。为了生成 CRC 校验码，将 $M(x)$ 向左移动 R 位，以模拟数据传输中的扩展过程。这一移动操作后，数据多项式的右侧将自然形成 R 位的空位，这些空位正是用来填充 CRC 校验码的区域[4]。设定 R 次多项式，并运用 $M(x)$ 与该多项式相除，得到的余数作为校验码，其公式为：

$$\frac{M(x)}{F(x)} = Q(x) + \frac{R(x)}{F(x)} \tag{2}$$

式中，$F(x)$ 为 R 次多项式；$Q(x)$ 为 $M(x)$ 与多项式相除后的商；$R(x)$ 为 $M(x)$ 与多项式相除后的余数。电力营销数据在传输的过程中，将有效数据与余数进行叠加，其构成的数据块在电力营销运行环境中传输，该数据块的码元可表示为：

$$S(x) = M(x) + R(x) \tag{3}$$

式中，$S(x)$ 为电力营销中接收端接收到的各个码元多项式。在按照上述内容操作时，要求 $S(x)$ 与 $F(x)$ 之间应整除，不产生余数，若产生余数，则说明在传输码元时出现错误，需要重新按照上述步骤完成操作[5]。在电力营销的数据传输中，尽管 $S(x)$ 与 $F(x)$ 之间的整除性理论上应保证数据的正确性，但实际中仍可能因码元传输错误而导致问题[6]。为减少此类错误的发生，一种有效策略是增加余数（校验码）的位数，即提升 $R(x)$ 校验码多项式的长度。这样做能够增强校验的严格性，从而进一步降低数据传输中的错误率。

2 电力营销数据循环冗余校验

在电力营销中，确立循环冗余校验码后，将其应用于数据传输管理，利用除法与余数原理来识别子系统间及与数据库通信时的数据错误[7]。在此框架下，码元 $S(x)$ 作为二进制数据多项式，其系数限定为 0 或 1。

CRC 在系统中实施循环检测，校验过程要求原始数据多项式 $M(x)$ 在长度上需超过 $S(x)$，并且预设 $M(x)$ 的首位（高位或低位）为1。若 $M(x)$ 的阶数为 B，则在 $S(x)$ 尾部附加 B 个 0 以扩展其长度，记作 $S+B$，表示扩展后的码元位数。此时，$x'S(x)$ 表示扩展后的多项式。

接下来，从 $x'S(x)$ 中剔除那些位于 $M(x)$ 定义域内且值不大于 1 的位串部分，剩余部分即构成 $Y(x)$，也即 CRC 校验码数据块。

在明确具体冗余校验内容后，引入 BP 神经网络，利用其实现对循环冗余校验的优化。BP 神经网络结构除基础的输入输出层外，还包含一个或多个隐含层，其中主要的数据处理与特征提取工作多在隐含层中进行。为提高效率，通常设计一个包含多个神经元的单一隐含层，以简化计算复杂度。各层之间的连接紧密程度通过权值来量化，这些权值在神经网络的学习过程中不断调整优化。在优化 BP 神经网络的权值及实现误差反向传播时，采用最速下降法，其核心在于根据误差函数的负梯度方向来调整权值系数。误差函数 E 被定义为期望输出与实际输出之间差异的平方和，以此量化网络性能，并引导权值向减小误差的方向调整。假设有 n 个样本，每个样本的误差为 e_i，则误差函数 E 可表示为：

$$E = \frac{1}{2n}\sum_{i=1}^{n} e_i^2 \tag{4}$$

式中，e_i 为期望输出 y_i 与实际输出 $\hat{y}_i$ 之差，即：

$$e_i = y_i - \hat{y}_i \tag{5}$$

式中，实际输出 $\hat{y}_i$ 为神经网络对第 i 个样本的

预测输出，是输入 x_i 通过网络后得到的。在 BP 神经网络中，权值更新采用梯度下降法。梯度下降法的基本思想是沿着误差函数关于权值的负梯度方向更新权值，以减小误差。假设有一个权值 ω，其更新规则可以表示为：

$$\omega_{new}=\omega_{old}-\eta\frac{\partial E}{\partial\omega} \tag{6}$$

式中，η 为学习率，用于对更新补偿的控制参数；ω_{new} 为更新后的权值；ω_{old} 为更新前的权值；$\partial E/\partial\omega$ 为误差函数关于 E 权值 ω 的偏导数，即梯度。假设 q 为数据并行处理位数，在经过（$n+q$）/q 个时刻后，实现循环冗余校验。采用 BP 神经网络优化的循环冗余校验技术，能够强化电力营销系统中各子模块与数据库间数据传输的效率与质量，确保电力企业获取精准的用户信息，并提升信息传输过程中的安全性与可靠性。

3 冗余数据剔除

在检测出冗余数据后，引入 ISODATA 算法，将冗余数据剔除。ISODATA 算法在去除样本冗余数据方面的核心策略是通过迭代聚类来优化数据集的代表性。该算法初始假设每个样本点均为潜在的聚类中心，随后基于用户定义的最小距离阈值，将相似数据点聚集成簇。在聚类过程中，算法动态调整聚类中心，使其更好地反映簇内数据的分布特性。完成聚类后，算法通过计算每个数据点到其所属聚类中心的距离，来评估数据点的代表性。具体操作为保留每个簇中距离聚类中心最近的元素作为有效数据，而将同一簇内的其他数据视为冗余并予以剔除。

此过程通过迭代优化实现，迭代参数的设置是关键。其中，K 代表初始考虑的样本总数，而两个重要的距离阈值 θ_N 和 θ_S 则用于指导聚类过程。θ_N 设定为相邻样本间欧氏距离的几何平均值，这一值反映了数据集整体的紧密程度。而 θ_S 的设定则更为灵活，它结合了相邻样本间欧氏距离的算术平均值与一个可调节的冗余剔除显著度因子 α，$\alpha\in[0,1]$，通过调整 α 的值，可以平衡数据保留的精度与冗余剔除的严格程度。初始时，假设每个样本点都是一个聚类中心。因此，如果有 N 个样本点，则初始聚类中心数量为 $K=N$。样本间距离通常使用欧氏距离计算。对于两个样本点 X_i 和 X_j，其欧氏距离为：

$$d_{ij}=\sqrt{\sum_{k=1}^{D}(X_i^k-X_j^k)^2} \tag{7}$$

式中，d_{ij} 为两个样本点 X_i 和 X_j 的欧氏距离；D 为样本的维度。对于上述论述中的 θ_S，可以通过相邻样本间欧氏距离的算术平均值加冗余剔除显著度得到：

$$\theta_S=\frac{1}{N(N-1)/2}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}d_{ij}+\frac{\alpha}{2} \tag{8}$$

将每个样本点分配到最近的聚类中心所属的簇。更新聚类中心为簇内所有样本点的均值。对于每个簇，保留距离聚类中心最近的样本点，删除其他样本点。上述论述中，ISODATA 算法通过迭代聚类与冗余数据剔除步骤，旨在保留最具代表性的数据样本，同时减少数据集中的冗余信息，优化数据质量。

4 应用效果分析

为验证上述方法的应用可行性，以某电力企业的营销数据为例，利用本文上述操作方法对冗余数据进行检测和剔除。通过对实例应用性能的量化评价，实现对该方法应用可行性的证明。分别设置 5 组数据样本，样本中数据均为该电力企业日常营销过程中产生的数据。在每组数据中人为增加不同数量的冗余数据，并将 5 组样本的具体情况记录如表 1 所示。

表 1　5 组数据样本情况记录表

序号	组别	正常数据量/条	冗余数据量/条
(1)	第一组	15263	8569
(2)	第二组	12526	8425
(3)	第三组	10245	8457
(4)	第四组	15236	9635
(5)	第五组	10236	8659

在完成对上述 5 组数据样本的冗余检测和剔除后，记录每组检测到的冗余数据量以及剔除后样本中剩余数据量，将得到的结果记录如表 2 所示。

表 2　冗余检测与剔除结果记录表

序号	组别	检测到冗余数据量/条	剔除后剩余数据量/条
(1)	第一组	8569	15263
(2)	第二组	8425	12526
(3)	第三组	8457	10245
(4)	第四组	9635	15236
(5)	第五组	8659	10236

从表 2 中数据可以看出，上述方法可以将数据样本中的冗余数据全部检测出来，并且剔除后剩余数据均为正常数据，可以将所有冗余数据全部剔除。因此，通过上述得出的结果证明，本文上述提出的方法具备极高的实际应用可行性。

5 结束语

本文围绕“电力营销数据冗余检测与剔除方法”展开研究，针对电力营销数据的特点与需求，提出一种综合的数据冗余检测与剔除方法。通过实际应用验证，本文所提方法显著提升了电力营销数据的质量，为电力企业的数据分析、业务决策提供了更加可靠的数据支持。

未来，随着大数据、人工智能等技术的不断发展，电力营销数据的冗余检测与剔除技术也将持续演进。期待通过持续地研究与创新，不断优化和完善相关算法与模型，进一步提升数据处理效率与准确性，为电力行业的数字化转型与智能化升级贡献更多力量。同时，也希望本文的研究能够激发更多学者与从业者的关注与探讨，共同推动电力营销数据冗余检测与剔除技术的深入发展。

参考文献

[1] 谢绒娜，范晓楠，李苏浙，等．基于标签的数据流转控制策略冗余与冲突检测方法 [J]. 网络与信息安全学报，2023，9（5）：21-32.

[2] 商俊燕，丁辉，胡学龙．基于 XGBoost 的无线传感器网络冗余数据检测算法 [J]. 传感技术学报，2022，35（11）：1568-1572.

[3] 崔亚洲，曹敬立，王玉君，等．基于电力营销大数据技术的反窃电检查应用分析 [J]. 自动化技术与应用，2024，43（5）：131-134，162.

[4] 孙磊．基于 K-means 算法的电力营销稽查异常数据监测方法 [J]. 信息技术与信息化，2024(6)：150-153.

[5] 计军恒，王建华，白留星，等．一种基于多源异构数据融合分析的电力营销管控平台设计 [J]. 四川水力发电，2024，43（2）：136-140.

[6] 李慧翔，刘博．探究基于改进决策树的电力营销数据挖掘方法 [J]. 电气技术与经济，2024(3)：209-211.

[7] 李佳凝．基于改进深度学习的电力营销数据异常识别研究 [J]. 电气技术与经济，2024（2）：212-214.

（收稿日期：2024-08-08）