DOI: 10.19783/j.cnki.pspc.201655

基于 Storm 架构的电力物联网流数据处理

张 磊[1]，刘辛彤[1]，蔡 硕[1]，刘红艳[1]，刘 蕾[2]

(1. 国网河北省电力有限公司信息通信分公司，河北 石家庄 050000；
2. 华北电力大学电气与电子工程学院，河北 保定 071003)

摘要：数据的采集、分析和处理是目前电力系统重点关注的问题之一。电力物联网系统中数据业务处理需求多样，尤其是流数据的处理对时延要求严格。然而，现有的数据处理方法形式单一，不能很好地满足低时延处理要求。因此，提出了基于 Storm 架构的电力物联网流数据处理方法。首先，基于 Storm 拓扑结构提出分布式流数据处理框架。进而采用流水线式的处理方式，从而达到缩短处理时间的效果。在数据接入后，采用循环队列和流转算子的方法。然后再通过边缘计算处理海量的数据，实现高效的协同工作。此外，通过支持向量机预测算法预测数据的发展趋势，采用清洗技术对脏数据进行处理，解决数据传输过程中的污染情况。仿真结果表明，与传统数据处理方法相比，所提流数据处理方法大大缩短了数据处理时间，满足了大量流数据处理的需求。

关键词：流数据处理；数据清洗；Storm 结构；支持向量机；电力物联网

Stream data processing of the power internet of things based on Storm architecture

ZHANG Lei[1], LIU Xintong[1], CAI Shuo[1], LIU Hongyan[1], LIU Lei[2]

(1. State Grid Hebei Electric Power Co., Ltd. Information and Communication Branch, Shijiazhuang 050000, China;
2. School of Electrical & Electronic Engineering, North China Electric Power University, Baoding 071003, China)

Abstract: Data collection, analysis and processing is one of the most important problems in power systems. There are various requirements for data business processing in the power internet of things (IoT) systems. In particular, there are strict requirements on delay for stream data processing. However, the existing data processing methods are single in form and cannot well meet the requirements of low-delay processing. Therefore, a stream data processing method of power IoT based on Storm architecture is proposed. First, a distributed stream data processing framework is proposed based on Storm topology. Then the pipeline-based processing method is adopted to shorten the processing time. Then data access, a circular queue and a flow operator are adopted. Then, through the edge computing processing of massive data, efficient collaborative work is achieved. In addition, the support vector machine prediction algorithm is used to predict the development trend of data, and the dirty data is processed by cleaning technology to solve the pollution in the process of data transmission. The simulation results show that, compared with the traditional data processing methods, the proposed method can greatly shorten data processing time and meet the needs of processing a large number of data streams.

This work is supported by the National Natural Science Foundation of China (No. 61971190) and the Science and Technology Project of State Grid Hebei Electric Power Co., Ltd. (No. SGHEXT00GCJS2000167).

Key words: stream data processing; data cleaning; construction of Storm; support vector machine; power internet of things

0 引言

随着电力系统信息化和智能化的不断发展，与电力系统相关的物联网数据规模也在不断增长，同时各类电力业务数据的处理需求日趋多样化，这给数据价值的挖掘带来了极大的挑战[1]。流数据作为电力物联网中一类典型的业务数据，对处理时延的要求更加严格。目前我国电网公司正在加强物联网在电力系统中的应用建设，电力物联网的发展如火如荼，因此，研究电力物联网中流数据的处理方法是十分必要的。

基金项目：国家自然科学基金项目资助(61971190)；国网河北省电力有限公司科技项目资助(SGHEXT00GCJS2000167)

虽然传统数据处理方法是多样的，例如文献[2]研究了通过人工智能方法进行数据处理，但是这些数据处理方式很难满足对数据的实时处理要求。由于当前对数据处理时延提出了更高要求，所以流数据处理方法受到越来越多的关注。目前，在多领域已对流数据处理方法进行了研究，如文献[3]研究了传感器网络中的实时数据处理方法。针对电力系统中的数据处理方法，文献[4]深入研究了基于数据挖掘的电力物联网多源业务体系，分别给出了对内、对外的数据业务体系架构；文献[5]从数据去噪、特征提取、模式识别、知识挖掘、数据存储、数据可视化等方面对电力数据进行了深入分析。然而，这些研究仍然存在处理方法形式单一，不能很好地满足低时延处理要求等不足，因此需要研究新的流数据处理方法满足电力物联网中大量数据快速处理的需求。

针对上述问题，本文提出电力物联网中分布式流数据处理框架 Storm 拓扑结构，以满足大量流数据处理的需求。进而采用流水线式的处理方式，达到缩短处理时间的效果，在数据接入后采用循环队列和流转算子方法。然后再通过边缘计算处理海量的数据，实现高效的协同工作。此外，文中采用 SVM 预测算法对数据的发展趋势进行预测，采用清洗技术处理脏数据，解决了数据传输过程中的污染状况。仿真结果表明，所提流数据处理方法大大缩短了数据处理时间，满足了大量流数据快速处理的需求。

1　深度学习理论算法模型

深度学习是一种具有更多层次结构的人工神经网络，是一般人工神经网络的扩展。深度学习是在传统机器学习的基础上产生和发展的，其本质是包含多个隐含层的神经网络，通过非线性变换把低层次的特征组合成高层次的抽象特征，从而实现多层网络隐含层节点的数据特征表达。由于层次不同、权重共享不同，深度学习网络结构也存在较大差异。图 1 为典型的深度神经网络结构[6]。

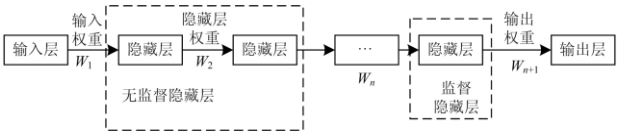

图 1　典型的深度神经网络结构

Fig. 1 Typical deep neural network structure

传统神经网络存在容易收敛到局部极小值的问题，深度学习算法的提出和发展解决了这一问题：① 大量样本数据；② 包含多个隐藏层的网络结构；③ 特征提取能力强于目标数据。其产生主要受两个方面的影响：① 由于通信、测量等技术的发展，数据积累量大，人工难以直观地发现数据中的特征规律；② 在大数据时代，由于并行异构计算的困难，难以解决计算难题。

循环神经网络多用于对序列数据的处理。在电力物联网中有大量数据产生，所以也会有多组数据序列，每组数据序列都是在不同时间和不同空间上组成的。这些数据序列的产生在时间和空间上都存在一定的联系。在对数据进行预测的时候可能会受到多种因素的影响，导致数据源参差不齐，人工很难对其做出正确判断，因此数据源整理也将成为一个很大的挑战。

2　数据处理结构

由于流数据具有连续不断、数据量大、规模及顺序无法预知等特点，单机处理难以满足海量数据处理的高效率要求，所以数据处理采用分布式计算模式，并充分考虑流式数据的处理特点[7]。根据多种处理方式的对比，递推计算模型可以更好地适应流数据处理特点。不同领域的数据其总体趋势和局部都存在差异。通过将复杂的处理流程划分为不同的处理逻辑单元，充分利用系统的计算资源，提高系统的通用性和复用性。一个典型的流数据处理流程如图 2 所示。可定义处理周期为

$$\Delta t = \max T_i + d,\ i \in [1,5] \tag{1}$$

式中：T_i 为执行第一个处理单元所需要的时间；d 为分段后由于单元间的数据传输、调度等造成的固定延迟。流水线模式为：在处理第一条数据的同时开始进行第二条数据的处理，即当第一条数据进行到识别模块的时候将第二条数据传送进来。由此可以计算出进行处理的时间为 $6\Delta t$，即 $T_2 = (2+5-1)\times\Delta t$，所以，处理 n 条数据的时间 $T_n = (n+5-1)\times\Delta t$。在流数据处理场景下由于数据源源不断地到来，即 $n \to \infty$；通过调整处理单元的逻辑与大小，使每个处理单元的时间近似相同，则对于未采用流水线模式的总时间 $5\times n\times\Delta t$。该技术将处理速度提高了近 5 倍。

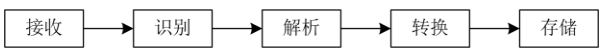

图 2　流数据处理流程图

Fig. 2 Flow data processing flow chart

3　数据分析方法

提出可靠合适的数据分析方法就是探索海量数据间规律的过程，也是解决数据处理问题的有效途径。在电力系统中需采集大量有关电表参数的数据，如何完整地呈现电表参数等情况，并有针对性地对这些数据进行分析和处理是值得探索的。

在电力系统监测中，由于各种因素会导致数据获取过程中存在多种问题，导致获取的数据也不准确，可靠性差，如传感器的异常故障以及数据的延迟。所以在分析电力系统数据时需要设计新的数据采集方案，对采集的数据进行处理分析，找出数据之间的关联，分析其关系，运用高效、高精度的检测算法，对数据进行精确检测，以应对电力系统中紧急情况的发生，这对电力监测来说至关重要。

3.1 数据接入

由于流数据是动态的，并且数量过大，可能会造成数据丢失或者部分数据未得到处理，同时数据量过大也会对数据接入处理架构带来一些难度，数据的接入问题也随之出现，因此需要提出合理的方法解决数据接入问题。本文在数据循环上采用循环队列，然后和流转算子相结合，不仅可以实现数据循环，也可以把算法逻辑循环起来。

在循环队列中通过全局变量定义一个大的数组，在数组中标志读、写两个位置，这样就可以实现一个循环队列的基本模型。将从 socket 缓冲区接收到的数据，缓存到队列中，将写指针向后移动，另外一个线程，操作读指针，不断跟随写指针，将数据取出，处理。

在算法逻辑的循环上采用流转算子，通过算子流转把某个数据需要处理的逻辑留在一个节点上，减少数据流转所用的时间，减少数据可靠性，保证开销，节约资源。可以采用 Storm 的分布式集群思想，它的每个节点均可以作为处理的中心或者节点，同时结合 Spark 的流转算子思路，进行逻辑流转。

3.1.1 Storm 结构

Apache Storm 是一个分布式的流数据处理框架，可以进行扩展，并且对数据有较高的包容能力，通过 Spout 获取数据，然后发送数据，系统中各个 Bolt 节点的作用是对发送后的数据进行处理。Storm 作为一种流处理技术，其提交运行的程序称为拓扑，拓扑结构由 Spout 和 Bolt 构成[8]，可以对电力系统的正常工作状态进行监控预测。Storm 拓扑结构如图 3 所示。

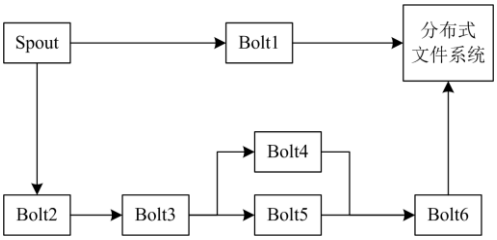

图 3 Storm 拓扑结构

Fig. 3 Topological structure of Storm

Spout 接收状态监测数据并形成元组，Bolt1 接收数据并存储到分布式存储系统中，Bolt2 对需要的监测数据进行筛选抽取，Bolt3 根据电力系统中收集到的数据特征将数据进行分类。然后将不同类别的数据分别存入不同的单位元组，如 Bolt4、Bolt5，在这些类别中选取一种类别作为正常数据，Bolt6 根据与正常数据的差异将数据分为正常、异常和故障三类，并将结果存入分布式文件系统中。此拓扑结构提取监测流数据中的一些特征，根据其偏离参考数据特征的程度，对电力系统状态进行评估。

通过不断增加数据量，对比数据集在 Storm 集群模式和单机模式下的运行时间，以此来验证 Storm 集群的吞吐能力。为提高实验结果的准确性，将各个数据取多次实验的平均值，测试结果如图 4 所示。

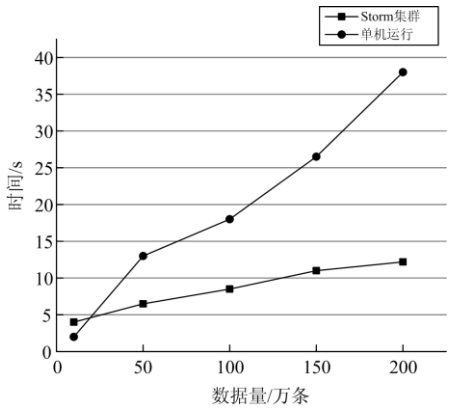

图 4 Storm 集群和单机模式运行时间对比图

Fig. 4 Storm cluster and stand-alone runtime diagram

由图 4 中的结果可以看出，运行时间与数据量的多少有关。当数据量过于小时，表明单机运行时间较短。造成这种结果的原因是 Storm 集群模式采用了分布式计算的方法，因此数据在每个节点间传输需要一定的时间。然而随着数据量的不断增加，Storm 集群处理数据所用的时间明显缩短，这时候集群的优势逐渐显现。Storm 集群具有拓展性，可以满足大量实时流数据的处理需求[9]。所以在流数据

处理过程中采用 Storm 结构可以提高数据处理速度。

Storm 架构具有以下几方面的优点：

(1) 分布式系统，可以进行横向拓展，为现有数据处理提供新思路；

(2) 运维简单，Storm 框架的部署非常简单，可降低整体复杂度；

(3) 高度容错，模块都是无状态的，随时宕机重启，即使一个模块出错也不会对整体造成太大影响；

(4) 无数据丢失，Storm 框架提出的消息追踪框架和复杂的事务性处理，能够满足更多级别的数据处理需求，实现对多级别数据进行处理。

Storm 具有实时连续性的分布式计算框架是其最大的优点，一旦运行起来数据会一直处于计算或等待计算的状态，这是非常高效的。而作为最佳的流式计算框架，使用 Storm 也需要考虑一些实际问题，例如数据入口需要使用消息队列，数据状态如何在流动中保存，以及怎样用分布式解决问题等。

3.1.2 数据预处理

在电力物联网的实际应用中，数据预处理消耗的时间也很多，这是因为数据要在一个高质量的基础上进行分析，这样可以避免错误数据带来的时间损耗，并且如果在数据存在误差的情况下对数据进行分析，不论分析算法多精确也会造成错误的结果。因此，高效率的数据预处理过程在数据分析中必不可少。

在电力物联网中，未经过预处理的数据一般会包含错误数据，这些错误数据可能存在缺失、重复、冗余等问题。所以在数据传送到数据处理框架之前，应对数据进行预处理操作。本文提出的数据预处理方法具体流程如下：

(1) 在数据缺失属性比例很大的情况下(大于 25%)，删除此条收到的数据记录；对于缺失率较低并且对总体数据几乎无影响的数据，可以通过拉格朗日插值法补全缺失数据。

(2) 搜索样本中重复的数据，对重复的数据进行去除。

(3) 利用收集到数据的属性以及数据之间的相关性，删减与下一步分析无关的冗余样本属性。

通过上述数据预处理的三个方法，可以减少无效数据，有效防止产生错误数据，提高数据准确性与规律性，为数据收集整理打下良好的基础。

3.1.3 流数据处理结构

Spark Streaming 是通过 DStream 对连续的数据流在固定的时间上进行分割，以此得到分割的数据。这些数据可以转化为不变数据，系统通过各种算子对数据进行处理，最后输出处理结果，实现了对流数据的处理。

数据处理系统[6]分为 3 大部分，分别为计算框架、缓存框架和处理接口框架，其数据处理基本框架如图 5 所示。其中比较重要的是计算框架，该计算框架提供底层分布式，由多个节点组成集群，负责处理单机可靠性和多机可靠性；在处理接口处进行了创新，处理接口框架提供使用该边缘处理系统的各种接口，包括处理算子、接入终端和数据处理拓扑。

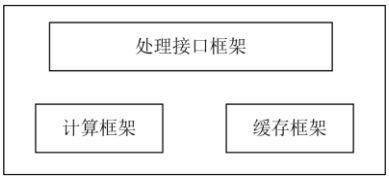

图 5 数据处理框架

Fig. 5 Data processing framework

在数据处理系统中最主要的是处理框架，数据的接入检测[10]以及数据的处理过程都是在该框架中。处理框架中包含循环队列和流转算子，实现数据及其逻辑的流转。

在循环队列模型中将从缓冲区接收到的数据缓存到队列中，将写指针向后移动，在另一个相邻的位置，操作读指针，不断跟随写指针，将数据取出、处理。在队列中的每一个节点上都带有流转算子，该算子通过对原始数据中产生的 RDD(Resilient Distributed Dataset, RDD)进行操作，将一个或多个 RDD 生成新的 RDD。在流数据处理过程中运用该方法使整个系统框架循环流动，在数据输入后，到达接收端开始进行处理，例如分类、清洗等。由于深度学习模型是静态的，所以可以通过该方法让静态的深度学习模型与流动的数据结合，让流数据更好地应用在深度学习模型中。

3.2 数据预测

数据预测可以对网络进行监控，识别网络中的噪声，或者发生故障时可以依据数据的合理范围进行判断，以此提高数据的质量，保证电力系统服务的可靠性。当数据存在噪声或传感器节点发生故障时，能够通过离群检测提高终端用户获取数据的稳定性，减少因错误数据的传输所产生的通信开销和其他不必要的问题。

3.2.1 支持向量机模型的构建

基于线性回归的 SVM 是 Vladimir N. Vapnik 在 20 世纪末提出的一种预测算法[11]。SVM 主要用于线性和回归。支持向量机基于统计学习理论，以结构风险最小化为原理，其核心思想是建立一个最优超平面，使正负比例在训练集中的间隔达到最大[12]。

采用 SVM 算法时，建立 SVM 模型，分析各部分的约束条件，定义损失函数 μ、松弛变量 ξ_i，给定样本数集[13]，并找到可实现样本分类的函数，这是 SVM 算法的重要步骤。

首先，在特征空间中，找到一个包围着目标样本点的超球体；然后，通过最小化由该超球体包围的体积[14]，使目标样本点尽可能地包围着超球体，将两种不同数据进行区分，以此达到区分两种数据类型的目的。先确定一个中心为 o，半径为 R 的最小球面。这个中心表示的是球体的中心，也是范围的参照点。

$$F(R,o,\xi_i)=R^2+D\sum_i\xi_i \tag{2}$$

并且使球面满足

$$(x_i-o)^{\mathrm{T}}(x_i-o)\leqslant R^2+\xi_i \tag{3}$$

其中，$i,\xi_i\geqslant 0$。满足上述条件就说明训练集中的数据都包含在球里面[15]。松弛变量 ξ_i 对数据点有一定的包含作用，以达到对模型的保护。D 为松弛变量的影响因子，可以调节松弛变量的大小，D 的大小也影响成本，D 变大，成本相应变大，因此需调小松弛变量。

基于约束条件，可以用 Lagrange 乘子法进行求解。

$$\begin{aligned}L(R,o,\alpha_i,\xi_i)=R^2+D\sum_i\xi_i-\sum_i\alpha_i\{R^2+\xi_i-\\(x_i-2ox_i+o^2)\}-\sum_i\gamma_i\xi_i\end{aligned} \tag{4}$$

因为 $\alpha_i\geqslant 0$ 且 $\gamma_i\geqslant 0$，且 α_i 和 γ_i 两个参数表示在拉格朗日算子中的影响变量，对参数求导并令导数等于 0，得到

$$\sum_i\alpha_i=1,\ o=\frac{\sum_i\alpha_i x_i}{\sum_i\alpha_i}=\sum_i\alpha_i x_i \tag{5}$$

并且有

$$D-\alpha_i-\gamma_i=0 \tag{6}$$

将上述公式代入拉格朗日函数，得到

$$L=\sum_i\alpha_i(x_i\cdot x_i)-\sum_{i,j}\alpha_i\alpha_j(x_i\cdot x_j) \tag{7}$$

将超球面的中心用支持向量来表示，则判定新数据是否属于这个类的判定条件为

$$(z\cdot z)-2\sum_i\alpha_i(z\cdot x_i)+\sum_{i.j}\alpha_i\alpha_j(x_j\cdot x_i)\leqslant R^2 \tag{8}$$

先假设含有 n 个训练集样本对 $\{(x_i,y_i),i=1,2,\cdots,n\}$，$x_i$ 是第 i 个样本的输入值，y_i 是第 i 个样本对应的输出值，然后在高维特征空间中建立一个线性回归函数，如式(9)所示。

$$f(x)=k\phi(x)+b \tag{9}$$

其中，$\phi(x)$ 是一个非线性映射函数。

μ是一个损失函数，可以根据μ的值来判断根据回归函数得到的预测值 $f(x)$ 和实际值y之间的关系[16]。

$$L(f(x),y,\mu)=\begin{cases}0, & y-f(x)|\leqslant\mu\\ |y-f(x)-\mu|, & |y-f(x)|>\mu\end{cases} \tag{10}$$

将松弛变量 ξ_i、ξ_i^* 加入，可以得到

$$\begin{cases}\min\dfrac{1}{2}\|k\|^2+D\sum_i^n(\xi_i+\xi_i^*)\\ \text{s.t.}\begin{cases}y_i-k\phi(x_i)-b\leqslant\mu+\xi_i,\ i=1,\cdots,n\\ -y_i+k\phi(x_i)+b\leqslant\mu+\xi_i^*\\ \xi_i,\xi_i^*\geqslant 0\end{cases}\end{cases} \tag{11}$$

其中，D 越大表示对训练误差大于μ的样本包容性越大，μ规定了回归函数的误差要求，μ越小表示回归函数的误差越小。

可见，将 SVM 预测方法运用在流数据的处理中，在面对大量数据的时候可以对后续数据进行判断，从而做好应对措施，避免一些故障的发生。

3.2.2 支持向量机模型的优化

对于输入向量的 m 维数，支持向量机优化为

$$\begin{aligned}\min k=\frac{1}{2}\|\omega\|^2+\\ \frac{1}{2}D\sum_{i=1}^n e_i^2-\sum_{i=1}^n\lambda_i(k^{\mathrm{T}}\phi(x_i)+b-e_i-y_i)\end{aligned} \tag{12}$$

式中：ω 的范数表示间隔距离的倒数；e_i 为初始误差；y_i 是一个矢量误差；λ_i 是拉格朗日乘子；b 是调节误差包容性的参数。

3.2.3 支持向量机模型的求解

引入 Lagrange 函数，对上面构建的模型进行求解：

$$\begin{cases}\max[-\dfrac{1}{2}\sum_{i=1}^n\sum_{j=1}^n(o_i-o_i^*)(o_j-o_j^*)K(x_i\cdot x_j)-\\ \qquad\sum_{i=1}^n(o_i+o_i^*)\varepsilon+\sum_{i=1}^n(o_i-o_i^*)y_i]\\ \text{s.t.}\begin{cases}\sum_{i=1}^n(o_i-o_i^*)=0\\ 0\leqslant o_i\leqslant m\\ 0\leqslant o_i^*\leqslant m\end{cases}\end{cases} \tag{13}$$

可以得到

$$k^*=\sum_{i=1}^n(o_i-o_i^*)\phi(x_i) \tag{14}$$

式中：o_i、o_j、o_i^*、o_j^*分别表示两个分类的支持向量到该超平面的距离，o_i、o_j 为一组，o_i^*、o_j^*为一组；K 表示一个核函数；m 表示最大的距离，即可区分的界限。在流数据处理过程中应用了 SVM 模型，与平常的预测模型不同，该模型中加入了一个流转算子，将预测固定的数据变成预测变化[17]的数据，并且在输入端的数据都是不断更新变化的。

3.3 预测结果

某小区的电力负载数据是不断变化[18]的，通过采集该小区的电力负载数据，以及记录对应的当天天气情况，分析天气对电力负载的影响。根据每天的电力消耗情况，提前预测用户用电情况，预测不同天气对电力负载的影响。

将标准化后的平均温度作为影响值，相应的电力负荷值作为模型输出[19]，进而对模型进行仿真，预测每天的电力负载情况并与真实值进行比较。其中通过对模型中的 D 值和 ξ 值进行调整，对预测模型进行优选，其预测值与真实值的对比见图 6。

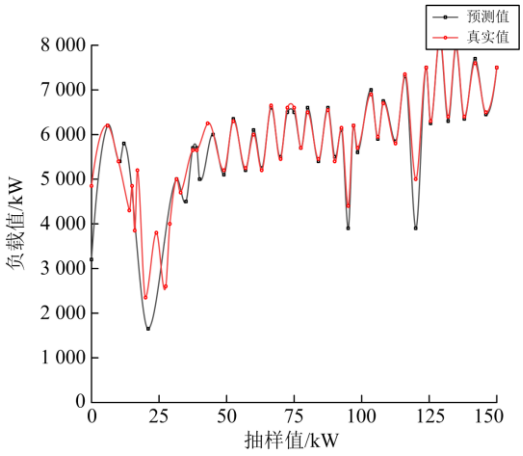

图 6 每天的电力负载预测值与真实值比较

Fig. 6 Comparison of predicted value and actual value of daily power load

由于电力负载也会受到天气环境的影响，通过调查冬天和夏天两个季节的实际负载[20]，可以根据预测值对用户的用电情况做出判断，提前采取防护措施。所以需要判断实际情况与 SVM 预测结果的差距，以此验证 SVM 模型是否合适，是否还需要做出调整。图 7 和图 8 分别对应夏季和冬季中某一天的预测结果。

从图 8 可以看出，正如预期的那样，冬季的误差开始有点大，因为相比夏季来说，白昼时间和天气因素的动态变化有点大。并且，从两个图都可以看出，所需的预测值都要比实际值大一点，以防止电力负载过高，预测没有达到要求，导致供电系统发生故障，用户用电出现问题。

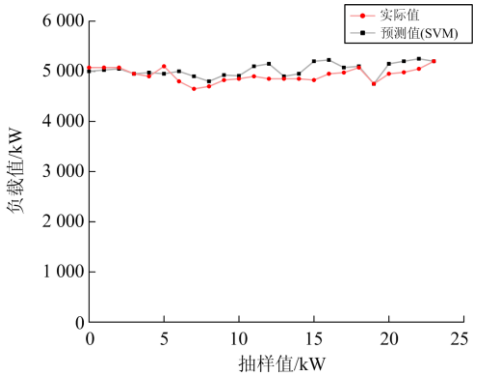

图 7 夏季某一天的实际需求和预测需求

Fig. 7 Actual and forecasting demand of a day in Summer

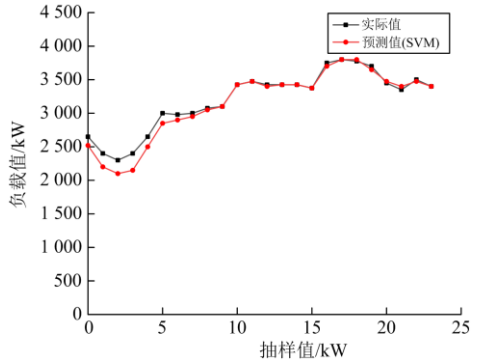

图 8 冬季某一天的实际需求和预测需求

Fig. 8 Actual and forecasting demand of a day in Winter

4 数据清洗

数据清洗主要是对输入的数据进行再检查、再处理和过滤的过程，主要通过数据抽取、数据转换、数据加载三个阶段来完成，目的是删除不合格数据，保留有用数据，以提高数据清洗的效率。

分布式数据清洗系统以 Map Reduce[21]为设计核心。在 Map Reduce 中，通过 Map 和 Reduce 两部分对数据进行处理，Map 端读取 HDFS 分布式文件系统中的文件，然后对读取的文件进行分割，在此过程中不同的分割片段执行不同的任务，再经过 shuffle 阶段进入到 Reduce 阶段进行整合操作，具体操作步骤如下[22]：

(1) 从 HDFS 中读取文件，将这个文件进行分割，划分成多组 Key/Value 键值对。

(2) 在 Map 里将 Value 值计算出来，然后进行统计。

(3) Combiner 对每个分区 Map 所对应的 key 值进行聚合，将 Map 端的输出作为 Combiner 的输入。

(4) Partition 针对分片进行处理，将 Combiner 统计出的 key 进行分区。

(5) Reduce 完成最后数据汇总并将其存入数据仓库。

将分布式系统与传统系统清洗时间进行对比，其结果如图 9。

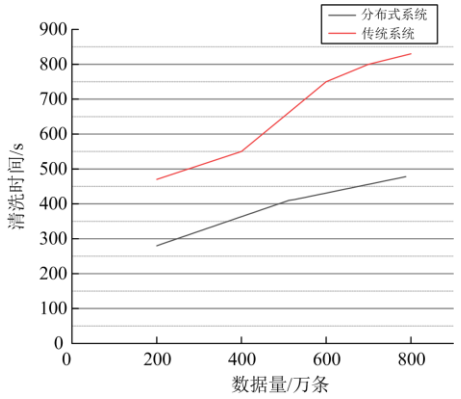

图 9 分布式系统与传统系统清洗时间对比图

Fig. 9 Distributed system versus traditional cleaning time diagram

如图 9 所示，随着数据的增长，两种清洗方式所用的时间折线图斜率逐渐变缓，但分布式清洗数据所用时间明显少于普通系统所用时间，所以分布式清洗数据方式有一定的优势。对于数据清洗的理论研究尚未成熟，但在大数据盛行的时代，数据清洗的研究仍有广阔的发展前景，针对电力物联网[23]数据清洗仍将是研究重点。

5 结论

基于流数据分析的电力物联网信息终端，可以有效地对电力设备及周边环境进行实时监测，并收集可靠的数据信息，为相关电力人员提供可靠的参考依据，保证服务用户的用电质量，但是在实际运行中存在的难题还需进一步解决。随着互联网的应用增加，人们对信息时代的迫切渴求，以及设备的更新，流数据处理技术不断发展，相关算法不断完善，在深度学习基础上和流数据技术的驱动下，电力物联网的发展在中国市场十分广阔。

参考文献

[1] SABARISH B A, SHANMUGAPRIYA S. Improved data discrimination in wireless sensor networks[J]. Wireless Sensor Network, 2012, 4(4): 117-119.

[2] ZHANG C. Intelligent internet of things service based on artificial intelligence technology[C] // IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), March 26-28, 2021, Nanchang, China: 731-734.

[3] OSBOME M A, ROBERTS S J, ROGERS A, et al. Towards real-time information processing of sensor network data using computationally efficient multi-output gaussian processes[C] // International Conference on Information Processing in Sensor Networks, April 22-24, 2008, St. Louis, MO, USA: 109-120.

[4] 王晓冰, 宋宝同, 王方敏. 基于数据挖掘的电力物联网多源业务体系研究[J]. 机电信息, 2020, 27: 29-31, 33.
WANG Xiaobing, SONG Baotong, WANG Fangmin. Research on multi-source business system of power internet of things based on data mining[J]. Mechanical and Electronic Information, 2020, 27: 29-31, 33.

[5] 方静, 彭小圣, 刘泰蔚. 电力设备状态监测大数据发展综述[J]. 电力系统保护与控制, 2020, 48(23): 176-186.
FANG Jing, PENG Xiaosheng, LIU Taiwei. Development trend and application prospects of big data-based condition monitoring of power apparatus[J]. Power System Protection and Control, 2020, 48(23): 176-186.

[6] WU Shanshan, NING Xin, GUO Shen, et al. Discussion on application of distribution internet of things in new industry form[J]. High Voltage Engineering, 2019, 45(6): 1723-1728.

[7] 于洪飞. 工业物联网技术的应用及发展[J]. 电子技术与软件工程, 2019(8): 20-22.
YU Hongfei. Application and development of industrial internet of things technology[J]. Electronic Technology & Software Engineering, 2019(8): 20-22.

[8] 甄凯成. 面向物联网应用的数据接入和存储系统的设计与实现[D]. 合肥: 中国科学技术大学, 2019.
ZHEN Kaicheng. Design and implementation of data ingestion and storage system for internet of things applications[D]. Hefei: University of Science and Technology of China, 2019.

[9] BATYUK A, VOITYSHYN V. Apache storm based on topology for real-time processing of streaming data from social networks[C] // IEEE Data Stream Mining & Processing, August 23-27, 2016, Lviv, Ukraine: 345-349.

[10] WANG X, ZHOU Z, HAN P, et al. Edge-Stream: a stream processing approach for distributed applications on a hierarchical edge-computing system[C] // IEEE/ACM Symposium on Edge Computing (SEC), November 12-14, 2020, San Jose, CA, USA: 14-27.

[11] 施珮. 基于无线传感器网络的水质数据流异常检测与预测方法[D]. 无锡: 江南大学, 2020.
SHI Pei. Outlier detection and prediction of water quality

data streams using wireless sensor networks[D]. Wuxi: Jiangnan University, 2020.

[12] 贾鹏, 杨炼鑫, 唐一鸣. 基于 SVM 算法在电力负荷预测中的研究[J]. 科技视界, 2020, 31: 14-16.
JIA Peng, YANG Lianxin, TANG Yiming. Research on power load forecasting based on SVM algorithm[J]. Science & Technology Vision, 2020, 31: 14-16.

[13] 席雅雯, 吴俊勇, 石琛. 融合历史数据和实时影响因素的精细化负荷预测[J]. 电力系统保护与控制, 2019, 47(1): 80-87.
XI Yawen, WU Junyong, SHI Chen. A refined load forecasting based on historical data and real-time influencing factors[J]. Power System Protection and Control, 2019, 47(1): 80-87.

[14] ZHENG Y, ZHONG G, LIU J, et al. Visual texture perception with feature learning models and deep architectures[C] // 2014 6th Chinese Conference on Pattern Recognition (CCPR), November 17-19, 2014, Changsha, China: 401-410.

[15] JABER A S, SATAR K A, SHALASH N A. Short term load forecasting for electrical dispatcher of Baghdad city based on SVM-PSO method[C] // 2018 2nd International Conference on Electrical Engineering and Informatics (ICon EEI), October 16-17, 2018, Batam Indonesia: 140-143.

[16] 康瑞, 齐林海, 王红. 基于流式计算的暂态电压扰动并行实时监测技术[J]. 电力系统保护与控制, 2020, 48(2): 129-136.
KANG Rui, QI Linhai, WANG Hong. Parallel real-time monitoring technology for transient voltage disturbance based on flow calculation[J]. Power System Protection and Control, 2020, 48(2): 129-136.

[17] 徐杰彦, 许雯旸, 褚渊. 区域尺度住宅建筑日用电负荷模型构建方法研究[J]. 中国电力, 2020, 53(8): 29-39.
XU Jieyan, XU Wenyang, CHU Yuan. Research on constructing method of daily electrical load model of residential building at regional scale[J]. Electric Power, 2020, 53(8): 29-39.

[18] 李福东, 曾旭华, 魏梅芳. 基于聚类分析和混合自适应进化算法的短期风电功率预测[J]. 电力系统保护与控制, 2020, 48(22): 151-158.
LI Fudong, ZENG Xuhua, WEI Meifang. Short-term wind power forecasting based on cluster analysis and a hybrid evolutionary-adaptive methodology[J]. Power System Protection and Control, 2020, 48(22): 151-158.

[19] 江元, 杨波, 郑黎明, 等. 工业物联网中基于机器学习方法的预测技术[J]. 物联网技术, 2020, 10(2): 45-48.
JIANG Yuan, YANG Bo, ZHENG Liming, et al. Prediction technology based on machine learning method in industrial internet of things[J]. Internet of Things Technologies, 2020, 10(2): 45-48.

[20] 李宝树, 张凤佳, 沈杨杨. 面向电网的边缘算力优化与分布式数据存储处理模型研究[J]. 广东电力, 2020, 33(9): 92-99.
LI Baoshu, ZHANG Fengjia, SHEN Yangyang. Research on power grid oriented edge computing power optimization and distributed data storage and processing model[J]. Guangdong Electric Power, 2020, 33(9): 92-99.

[21] 唐俊熙, 曹华珍, 高崇, 等. 一种基于时间序列数据挖掘的用户负荷曲线分析方法[J]. 电力系统保护与控制, 2021, 49(5): 140-148.
TANG Junxi, CAO Huazhen, GAO Chong, et al. A new user load curve analysis method based on time series data mining[J]. Power System Protection and Control, 2021, 49(5): 140-148.

[22] 赵闻, 张捷, 李倩. 面向电网数据采集系统的多业务资源分配算法[J]. 广东电力, 2020, 33(11): 60-65.
ZHAO Wen, ZHANG Jie, LI Qian. Multi-service resource allocation algorithm for power grid data collection system[J]. Guangdong Electric Power, 2020, 33(11): 60-65.

[23] CHEN H Y, WANG X Z, LI Z H, et al. Distributed sensing and cooperative estimation/detection of ubiquitous power internet of things[J]. Protection and Control of Modern Power Systems, 2019, 4(2): 151-158. DOI: 10.1186/s41601-019-0128-2.

收稿日期：2020-12-31；　修回日期：2021-08-19

作者简介：

张　磊(1985—)，男，硕士，高级工程师，研究方向为网络、安全、语音交换；E-mail: 18003214162@126.com

刘辛彤(1991—)，女，硕士，中级工程师，研究方向为语音交换、数据网络；E-mail: 452953765@qq.com

蔡　硕(1990—)，女，硕士，中级工程师，研究方向为卫星通信、视频会议系统。E-mail: caishuo11111@163.com

（编辑　魏小丽）