文章编号:1671-4598(2019)03-0173-04 DOI:10.16526/j.cnki.11—4762/tp.2019.03.035 中图分类号:TP338.6 文献标识码:A

分布式 SVR 在短期负荷预测中的研究

张志禹，侯 凯，李晨曦
(西安理工大学 自动化与信息工程学院，西安 710048)

摘要：准确的负荷预测，可以合理安排机组启停，降低发电成本，特别是短期负荷预测对电力系统控制、运行和规划都有重要意义；传统的预测方法不能及时准确地反映需求响应，在 Hadoop 环境下利用分布式支持向量回归机（Support Vector Regression，SVR）实现负荷预测，同时使用基于均匀设计的自调用 SVR（UD—SVR）方法进行参数寻优，进一步提高文章实现的分布式 SVR 算法精度；通过真实的电力负荷数据集验证该算法，实验数据来自我国西部某地级市连续 424 天的真实用电量数据；结果表明，文章改进后的算法用于短期电力负荷预测是可行的，不仅预测准确度又在原有基础上明显提高，并且随着数据量的增加，计算速度也大幅提高，减小了负荷预测时间。

关键词：负荷预测；Hadoop 平台；支持向量机；参数优化

Research on Distributed SVR in Short—term Load Forecasting

Zhang Zhiyu，Hou Kai，Li Chenxi
(College of Automation and Information Engineering，Xi'an University of Technology，Xi'an 710048，China)

Abstract：Accurate load forecasting can reasonably arrange the start and stop of the unit and reduce the cost of power generation，especially short—term load forecasting is of great significance for power system control，operation and planning. However，the traditional forecasting methods can not reflect the demand of users timely and accurately. Load forecasting using distributed support vector regression (SVR) in Hadoop environment，at the same time，the self—call SVR (UD—SVR) method based on uniform design is used to optimize the parameters，and the accuracy of the distributed SVR algorithm implemented in this paper is further improved. The algorithm is validated by a real power load data set. The experimental data comes from real electricity consumption data for 424 consecutive days in a prefecture—level city in western China. The results show that the improved algorithm is feasible for short—term power load forecasting. Not only the prediction accuracy is improved on the original basis，but also the calculation speed is greatly improved with the increase of data volume，which reduces the load forecasting time.

Keywords：power load forecasting；Hadoop platform；SVR；parameter optimization

0 引言

随着计算机技术的飞速发展以及移动互联网科技的不断进步，网络已经渗透进人们生活的方方面面，与此同时互联网所产生的信息也发生了爆炸式的增长，随着智能电网的发展，电力数据资源急剧增长。以云计算为代表的新一代 IT 技术在电力系统中的应用更加广泛[1]。智能电网的主要目的就是通过获取更多相关信息来优化电能的生产、分配和消耗[2]。从本质上来说，智能电网是当今的大数据技术在电网系统上的具体应用。电网在实际的运行、检修和管理等过程中都将会产生海量数据，这些都是典型的大数据特点。由此可见，大数据技术即将成为我国智能电网未来发展的新方向[3]。

Hadoop 是由 Apache 基金会开发的分布式系统基础架构，是一个大数据处理的框架，它能够部署在任何普通 PC 机上，并可以在多种平台上运行，具有良好的可靠性和可扩展性。Hadoop 的分布式系统架构，既保证了数据的安全又有效缓解了运算时间过长等问题[4]。传统方式处理大规模数据时，大多选择高性能计算，硬件器材往往非常昂贵，并且对大规模数据进行有效分割以及计算任务的合理分配都需要经过开发人员繁琐复杂的编程才能实现，而 Hadoop 的 HDFS（Hadoop Distributed File System）系统与 MapReduce 编程框架很好的解决了这些问题[5]。

本文将机器学习中的 SVR 算法[6]以及能够处理海量数据的 Hadoop 平台相结合并应用到电力负荷预测的领域中。针对传统算法面对海量高维数据时单机运算资源不足以及运算时间成指数级增长的缺陷，引入了 MapReduce 编程框架、HDFS 分布式文件系统等分布式计算技术，通过改进负荷预测算法，以提高负荷预测算法的计算速度；使用基于均匀设计[7]的自调用 SVR（UD—SVR）[8]方法进行参数寻优，在 SVM 算法中加入参数寻优[9]进一步提高预测算法的准确性。最后，通过实验验证了本文实现的 Hadoop 平台分布式 SVR 的算法性能与 UD—SVR 参数寻优方法效果。

1 HADOOP 平台

1.1 HDFS

HDFS 是 Hadoop 分布式文件系统（Hadoop Distribu-

收稿日期:2018-08-28； **修回日期**:2018-09-25。
基金项目:国家自然科学基金资助重大项目(41390454)。
作者简介:张志禹(1966-)，男，山西朔州人，博士，硕士生导师(教授)，主要从事阵列信号处理、大数据和人工智能方向的研究。

ted File System）的缩写。HDFS 是 Hadoop 的核心子项目[10]。HDFS 具有很强的容错性，能够保证在一个或者若干个出现故障后，集群仍旧能够正常运行；它能够有效提高网络带宽的利用率，减少网络阻塞的风险；同时可以提供高吞吐量的数据访问。

在 HDFS 中文件的存储、处理和备份的基本逻辑单元是 block（块），默认一个 block 的大小是 64 M，每一个 block 都有自己的副本，默认副本数为 3。节点冗余一般是通过拷贝来实现的。默认的 3 个副本的存储方式为，在同一个机架的不同节点，各保存一份，在不同的机架上再保存一份。这样的好处是，当本节点出现故障时，优先使用同一机架的另一个节点作为替代，原因是在同一个机架内带宽较大，速度较快；当整个机架出现故障时，使用保存在另一个机架上的备份，这样使得文件不会丢失。这样就使得 HDFS 既安全可靠有很高的容错性，又保证高效性，提高带宽利用率。

1.2 MapReduce

MapReduce 是由主/从结构组成，即 JobTracker 与 TaskTracker[11]。对大规模数据集进行复杂的并行化运算都可以抽象成 map 函数和 reduce 函数。具体的底层实现被封装起来了，用户只要自己编写 map 函数和 reduce 函数来实现功能即可。图 2 展示了一种 MapReduce 的操作过程，MapReduce 中数据转换成如下键值对形式：

Input→Map（K 1，V1）→list（K2，V2）→Reduce（K2，list（V2））→list（K3，V3）→Output

首先，从 HDFS 中读取数据，数据被分割成数据 split；其次，数据会被转化为 key/value 格式，MapReduce 会为每一个 map 任务分配一个 split 进行并行运算；其中 shuffle 是一个 map 端的排序过程，最终 reduce 端以前对数据进行合并，再把 reduce 端输出的结果，保存到分布式系统所指定的位置。

2 预测模型

短期负荷预测对电力系统的安全运营有着极为重要的作用，例如是制定阶梯电价的重要依据，对电网的安全经济运行提供了保证。本文所用的负荷数据来自我国西部某地级市 2015 年 1 月 1 日至 2016 年 2 月 28 日，共 424 天的真实用电量数据，一天 24 小时内，每 15 分钟采样一次，每天 96 个采样点，负荷的单位是兆瓦。其中 2015 年全年的历史负荷数据作为训练样本，将 2016 年的数据当做未知数据，对不同时间的负荷值进行预测。

在电力系统的负荷预测的角度，影响因素有很多，是一个多变量的预测问题[12]，为了提高 SVM 算法的推广能力，需要兼顾各种因素。依据文献[13]和生活经验，t 时刻的负荷量可以用下面的公式表示：

$$Y(t) = N(t) + W(t) + S(t) + r(t) \quad (1)$$

其中：Y（t）是 t 时刻的负荷总量，N（t）是受到历史负荷的影响所产生的负荷分量，W（t）是受到受到气象因素影响所产生的负荷分量，S（t）是受到特殊事件影响所产生的负荷分量，r（t）是不可控事件所产生的负荷分量。通过以下实验将讨论各个因素对负荷预测问题的影响。

2.1 历史负荷因素 N（t）

它与天气、节假日等其他因素无关，主要指的是所预测地区养成的用电习惯，在负荷预测中有延续性和类推性的原则，也就是说，我们可以根据先前发生的事件来推测后续事件发生的趋势，从而对后续事件做出预测。

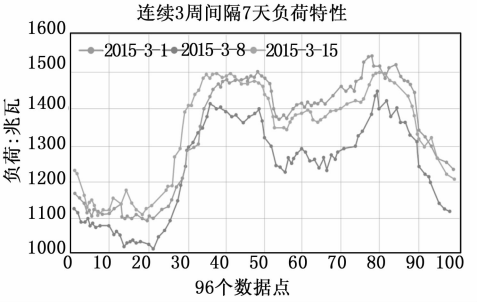

图 1 每周的相似日负荷特性

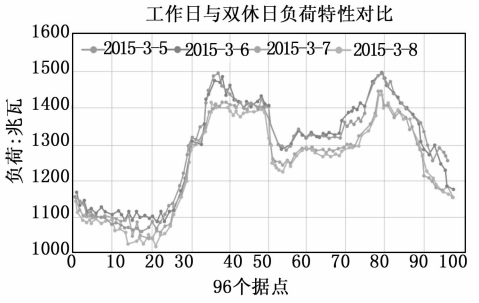

图 2 工作日与双休日的负荷数据特性对比

图 1 中，3 月 2 日，9 日，16 日之间分别间隔 7 天，也就是说他们的星期属性相同，由于相隔时间较远以及气温等其他因素的影响，因此这 3 天的负荷值，在数值上差别相较于连续 3 天的差值更大，但是在波动的走势上却极为相似，也就是说连续几周内，星期属性相同的时候，负荷特性具有相似性。

图 2 中，3 月 5 日至 8 日，共四天的时间内，包括连续两天的工作日，和连续两天的休息日。可以看出，连续两天的工作日，即 5 日和 6 日，它们的负荷量极为相似；连续两天的休息日，即 7 日 8 日，它们之间的负荷特性极为相似，同时，6 日和 7 日两天虽然也是连续的时间，但是曲线之间的距离就相对比较远，也就是说工作日和双休日之间的负荷特性差别较大。因为历史数据对待预测的负荷值相关性很大，因此，历史数据是不可或缺的属性值。

2.2 气象因素 W（t）

由气象因素产生的负荷分量是 W（t），影响负荷的气象因素有很多，比如温度、湿度、降水量等等，天气的突变会引起负荷的剧烈变化。在诸多气象因素中，温度对负荷的影响最为显著，在短期负荷预测中，将气温数据当做重要的属性考虑进去，早已成为人们在研究负荷特性时的

共识。

当温度值一定时，负荷总会在一个特定的范围内变化，表明气温与负荷之间相关性很强。

2.3　特殊事件因素 S（t）

在 S（t）因素的作用下，将会导致负荷量明显偏离区域用电习惯的轨迹，特殊事件常指重大节假日、纪念日、电力系统限电、自然灾害等等。此类事件往往可以根据人为修正和优化模型中的属性得到改进。以 2015 年除夕和春节作为特殊事件为例，图 3 展示了除夕前后的负荷曲线。

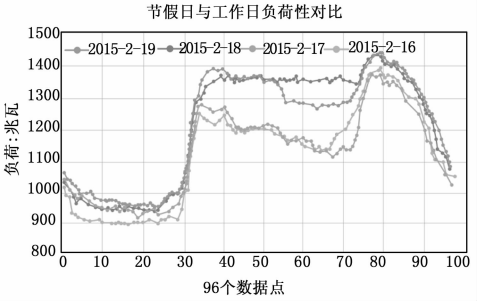

图 3　节假日与负荷的关系

图 3 中，2 月 18 日是除夕，可以看出假期前的最后两个工作日，即 16 日和 17 日，它们的负荷特性类似，18 日和 19 日，这两天假期的负荷特性相似，但是工作日和假期之间的负荷特性差别较明显，因此影响负荷的因素中还应该包括节假日因素。

在负荷 L（t）中提取出 N（t）、W（t）和 S（t）后剩余的残差，即不可控因素 r（t），它是随机负荷序列，这些随机序列因其随机性从而往往难以预测，因此很难把它们量化到属性值中。

2.4　预测负荷建模

综上所述，通过阅读资料和分析历史数据，最终将输入样本的属性值确定如下：

（1）自预测之日起前 7 天的同一个时间点的的负荷值：$N_1=\{n_1, n_2, \ldots\ldots, n_7\}$。

（2）自预测的时间点起最近的前 7 个时间点的的负荷值：$N_1=\{n_8, n_9, \ldots\ldots, n_{14}\}$。

（3）预测日的星期属性，0 表示周内，1 表示周末。

（4）预测日的特殊事件属性，0 表示工作日，1 表示特殊日。

（5）预测日的温度属性，$W=\{T_h, T_l\}$，T_h 表示当日最高温度，T_l 表示当日最低温度。

3　UD－SVR 参数寻优

在大数据集情况下，单机 SVR 算法处理大数据集时耗时巨大，为了进一步提高分布式 SVR 训练精度，本文采用了 UD－SVR 方法进行参数寻优。在众多实验点中选取其中最具有代表性的若干个点，这些被选取的点以其在所有实验点中的“均匀分布”程度作为标准，均匀设计只考虑试验点在范围内的均匀散布[14]，通过提高试验点“均匀分散”的程度，使试验点具有更好的代表性，能用较少的试验获得较多的信息[15]。

SVR 算法有（c，g，p）三个参数需要优化选择，共有 729 组（c，g，p）参数，若将这所有 729 组全部进行 SVR 计算，则耗时巨大，可行性较低。本文用均匀设计法，从这 729 组参数中选择最具代表性的 27 组参数进行参数计算，得到的结果作为训练样本再利用 SVR 的预测功能对全部 729 组参数进行预测，最终预测结果中误差最小值所对应参数为最优参数，误差选择使用均方差 MSE：

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(Y-M)^2 \tag{2}$$

其中 Y 表示真实值，M 表示实际值。

表 1　27 组参数 SVR 对应 MSE

	MSE	C	G	P
1	9.80E－04	0.5	0.0625	0.00390625
2	9.10E－04	2	1	0.0078125
3	0.0010653	2	0.015625	0.0625
4	0.001716696	128	0.0078125	0.125
5	9.88E－04	16	0.0078125	0.00390625
6	9.92E－04	64	0.00390625	0.03125
7	0.00102503	16	0.125	0.0625
8	0.007442132	0.5	0.0078125	0.25
9	9.33E－04	1	0.25	0.03125
10	0.001061167	1	0.00390625	0.015625
11	0.007765813	128	0.0625	0.25
12	0.001914247	4	0.0625	0.125
13	0.129285749	64	0.25	1
14	0.129340946	1	0.125	1
15	0.129360979	8	0.00390625	1
16	0.024413315	2	0.03125	0.5
17	9.87E－04	4	0.015625	0.0078125
18	0.056415381	4	1	0.5
19	9.11E－04	8	0.25	0.015625
20	9.12E－04	64	0.125	0.015625
21	9.49E－04	32	0.03125	0.0078125
22	0.002667686	0.5	0.5	0.125
23	0.010185466	16	0.5	0.25
24	0.001348564	32	1	0.0625
25	0.012695333	32	0.015625	0.5
26	0.001241302	64	0.5	0.0625
27	9.70E－04	8	0.03125	0.03125

具体参数寻优步骤可分为：

（1）依据均匀设计表对 729 组参数进行 27 组参数的选择。

（2）对 27 组参数进行单机 SVR 的训练，使用 5 折交叉验证，得到的 27 组 MSE。

（3）将 27 组 MSE 与 27 组参数组合为新的训练样本，使用留一法对 729 组参数进行 SVR 训练，得到的 MSE 最小值所对应的（c，g，p）即为新样本的最优参数。

(4) 使用 3 中得到的最优参数重复 3 中实验，最后得到的 MSE 最小值所对应的 (c，g，p) 即为单机 SVR 的最优参数。

4 实验结果与分析

本节设计三组实验用来检验分布式 SVR 模型的预测性能，实验对比了包括训练时间、预测值与实际值的误差、训练加速度等，综合分析评判所建立的分布式 SVR 模型与单机 SVR 性能之间的差异。

4.1 训练时间

对于训练时间的比较，首先选用不同大小的数据集 (1 万行，5 万行，10 万行，15 万行，20 万行)，数据集大小从 0.6 M 至 12 M，分别对单机 SVR 算法、分布式 SVR 算法进行训练预测模型时间的对比实验。实验结果如图 4 所示。

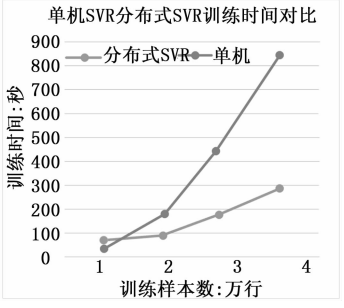

图 4 单机 SVR，分布式 SVR 训练时间

分析可知，在实验样本数据较少的情况下，Hadoop 环境下由于存在一些集群的通讯协作开销和 Reduce 端的 SVR 训练，训练时间比单机训练时间长，随着训练样本的增大，单机 SVR 的训练时间呈指数级增长，而 Hadoop 平台分布式 SVR 训练时间增长斜率很小。

4.2 预测模型准确率

对于预测模型准确率的比较，训练数据选用原数据集中 2015 年一整年的数据，数据大小 7.96 M 占用 Hadoop 集群 4 个 block 块。选用 2016 年 1 月 1 日 0 时 0 分—2016 年 1 月 7 日 23 时 45 分 (共一星期) 的 672 个数据点作为预测数据集，分别对单机 SVR 算法、分布式 SVR 算法进行预测实验。因篇幅有限，图 5 列出了前三天的预测效果图。

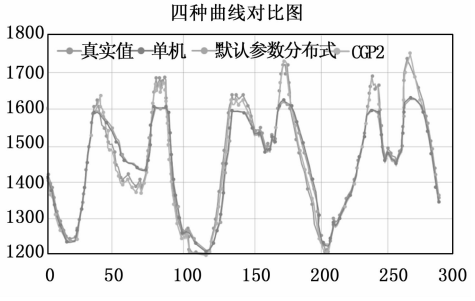

图 5 4 种曲线对比图

从图中可得使用默认参数的单机 SVR 与分布式 SVR 效果相似，而使用了参数寻优后的分布式 SVR 与真实值最为接近，效果最好。表 2 显示了使用平均相对误差 (MAPE) 和均方差 (MSE) 来衡量预测值与真实值误差。

表 2 672 个数据点平均相对误差、均方差

误差 / 算法	MAPE	MSE
单机 SVR	0.0189	1317.19
分布式 SVR	0.0215	1703.54
UD−SVR 分布式 SVR	0.0064	317.76

从中可以看到默认参数的单机 SVR 预测值稍好于默认参数的分布式 SVR，经过参数寻优的分布式 SVR 的预测效果最优。

4.3 加速比

为了充分测试 Hadoop 平台的并行 SVR 算法的性能变化与集群中 Map 任务的数量的关系，本文采用了加速比来衡量该并行算法在训练时间上的提升速率。

计算加速比方法如下：

$$P = T_1/T_2 \tag{3}$$

其中：P 表示加速比，T_1 为单机 SVR 算法的训练时间，T_2 为 Hadoop 平台分布式 SVR 算法的训练时间。

实验结果如表 3 所示。

表 3 加速比对比实验

Map 任务个数	1 个(2MB)	2 个(4MB)	3 个(6MB)	4 个(8MB)
加速比	0.44	1.92	2.46	3.00

当 Hadoop 集群中的节点为 1 个 Map 任务时，由于集群中的节点间的网络通讯和 Reduce 端的 SVR 训练，使得 Hadoop 平台的并行 SVM 算法的训练时间超过了单机 SVR 算法。然而，随着集群中计算 Map 任务数目的增加，并行算法的加速比逐渐提高。

5 结束语

本文在 Hadoop 平台实现了分布式 SVR 算法，在大数据集情况下，使用了 UD−SVR 方法进行了 (c，g，p) 参数寻优。通过 3 组实验对比了单机 SVR 和分布式 SVR 算法在训练时间和预测模型准确度上的效果。实验表明分布式 SVR 在保证准确度不明显降低的情况下大幅度缩短了训练时间，而使用了 UD−SVR 参数寻优后的分布式 SVR 的预测效果极佳。

参考文献：

[1] 宋亚奇，周国亮，朱永利. 智能电网大数据处理技术现状与挑战 [J]. 电网技术，2013，37 (4)：927－935.

[2] 刘 旭，罗滇生，姚建刚，等. 基于负荷分解和实时气象因素的短期负荷预测 [J]. 电网技术，2009，33 (12)：94－100.

[3] 李振元，李宝聚，王泽一. 大数据技术对我国电网未来发展的影响研究 [J]. 吉林电力，2014，42 (1)：10－13.

(下转第 182 页)