第49卷第2期 中国科学技术大学学报 Vol. 49, No. 2
2019年2月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Feb. 2019

文章编号:0253-2778(2019)02-0100-05

基于电力负荷曲线的设备识别方法

王子一[1,2],商 琳[1,2]

(1.计算机软件新技术国家重点实验室,江苏南京 210023;2.南京大学计算机科学与技术系,江苏南京 210023)

摘要:电力设备的负荷曲线随着时间而变化,其本质上是时间序列数据.为此提出了一种新的通过负荷曲线识别电力设备的方法,该方法在多个粒度划分出的负载曲线上使用卷积神经网络作为基分类器构造出一个集成学习器来提高分类精度.首先我们对原始数据进行不同粒度的划分,得到若干不同的新数据集.其次使用这些新的数据集训练不同的基学习器,并根据验证集上的精度得到不同基学习器的权重.将测试样本按照相同的粒度划分方式得到不同的测试数据集,使用不同的基分类器对这些测试数据集进行测试,得到对应的预测标签.最后对不同基分类器预测的标签进行加权,并选出权重最大的那个标签作为预测标签.在实际的电力负荷数据上将该模型与单个CNN模型进行对比,实验结果表明,该模型具有更高的设备识别精度.

关键词:力负荷曲线;粒度;时间序列;集成学习;分类

中图分类号:TP391 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2019.02.003

引用格式:王子一,商琳.基于电力负荷曲线的设备识别方法[J].中国科学技术大学学报,2019,49(2):100-104.
WANG Ziyi, SHANG Lin. Equipment identification from power load profile[J]. Journal of University of Science and Technology of China, 2019,49(2):100-104.

Equipment identification from power load profile

WANG Ziyi[1,2], SHANG Lin[1,2]

(1. *State Key Laboratory for Novel Software Technology, Nanjing* 210023, *China*;
2. *Department of Computer Science and Technology, Nanjing University, Nanjing* 210023, *China*)

Abstract: The power load profile of the equipment varies with time, and it is essentially time series data. A new ensemble learning method for identifying electrical equipment through load power profile is proposed, which uses convolution neural network (CNN) as base learner to train the multi-granular load profile to improve the accuracy of classification. First, the raw data with different granularities are divided and some different new data sets are obtained. Then, these new data sets were used to train different base learners and get the weight of different base learners according to the accuracy of validation sets. In the testing process, testing data are divided based on different granularities in the same way as the training data are fed into base learners and the final results are obtained by weighting the output of each base learner. The proposed model are compared with a single CNN model on the electrical equipment load data. The experimental results show that the proposed method has higher accuracy in the identification of electrical equipment.

收稿日期:2018-05-24;**修回日期**:2018-09-28
基金项目:国家自然科学基金(61672276)资助.
作者简介:王子一,男,1993年生,硕士生.研究方向:智能电力系统.E-mail:zywang@mail.nju.edu.cn
通讯作者:商琳,博士/副教授.E-mail:shanglin@nju.edu.cn

Key words: power load profile; granular; time series; ensemble learning; classification

0　引言

设备的监管除了人工维修,还有一些常用的技术,比如设备参数的检测和图像识别[1],除此之外设备检测系统还使用了很多技术分析设备产生的负荷曲线,用来保证设备的正常运转. 在某些情况下异常的发生只会导致负荷曲线出现偏移或者无规律的剧烈变化,人为的检测不易发现,这将会导致安全隐患. 如果能使用机器学习的技术对负荷曲线进行分析,这将有助于设备的管理.

负荷曲线本质上是一类时间序列数据,目前有很多针对时序数据分类设计的算法,比如基于动态时间扭曲(DTW)设计的 KNN 算法和根据 shapelet[2]设计的集成算法. 研究表明,当训练数据足够,基于 DTW 的 KNN 算法等价与基于欧式距离的 KNN 算法[3]. DTW 和 Shapelet 算法搜索的时间复杂度太高,这些传统的时序分类算法只适用于某些特定的数据集. 最近,有学者将 CNN 引入时序分类中,并取得了成功[4]. 尽管 CNN 已经在时序分类领域取得了成功,但是在实际应用中如果当训练数据的获取代价太大,即训练数据的大小受限的情况下,CNN 很容易在这样的数据集上过拟合,导致泛化性能降低. 虽然目前有一些技术可以缓解过拟合,但是最有效的办法还是增加训练数据的量.

集成学习通过构建并结合多个学习器来完成学习任务,并在工业界和数据挖掘竞赛上取得了巨大成功. 文献[5]提到,2015 年 Kaggle 比赛公布的 29 个最优解决方案中有 17 个是使用的 XGBoost. XGBoost 是一种基于 boosting 的集成学习算法,已经在工业界取得了巨大的成功. 另有数据表明 KDD 比赛从 2007 到 2016 年的第一名均是使用的集成学习算法.

本文提出了一种新的集成策略,来进行电力负荷数据的分类,从而实现电力设备的识别. 通过对原始样本进行不同粒度的划分得到不同的新数据集,达到对样本进行扰动的目的;然后对这些新的数据集,使用同种机器学习算法训练出多个基学习器. 接着,在验证集上验证基学习器的性能并将验证集上的精度作为基学习器投票的权重. 因为我们提出的这种集成策略,本质上是同质的集成,每个基分类器做预测的概率空间是相同的,所以不需要特别设计权重;最后使用相同的粒度划分测试样本,使用训练好的基分类器分别对划分好的测试样本进行预测,并对预测的结果进行加权投票,得到最后的预测结果.

本文的主要贡献如下:

(Ⅰ)提出了一种通过负荷曲线识别设备的方法;

(Ⅱ)设计了基于粒度的新的集成学习策略并提出了一种基学习器的权重计算方法.

1　相关工作

电力设备的异常运行使得工厂的工业生产存在一定程度上的安全隐患,给设备的管理和维护. 目前,智能电网运用越来越多的实时设备数据检测仪器,这些仪器的主要作用是根据已有的数据对设备用电状况进行估计或对参数进行监测. 由于一些隐患的发生,只是其负荷曲线发生偏移或者抖动而其负荷的值仍在正常范围内. 人工检修设备的难点主要有两个:一是工人数量有限很难兼顾所有的设备;二是一些隐患在发生前设备的参数仍是正常的,因此对设备的负荷曲线进行检测和分析对工厂的设备管理是非常必要的. 负荷曲线一个重要任务就是用于对设备进行识别. 如果能够以较高的精度识别某个负荷曲线属于何种设备,则通过负荷曲线的分析能进行设备预警.

设备的负荷曲线,本质上是一类时序数据. 目前时序的分类方法一般可以分为两类:基于距离计算和基于特征提取. 对距离计算的方法来说,其关键是计算两个时序之间的相似度,此类方法中最突出的工作是基于 DTW 的 KNN 算法. 由于时序数据之间往往存在局部的位移,基于欧氏距离的 KNN 很难处理时序的分类问题. DTW 算法可以调整时序数据之间的扭曲,DTW 度量的相似性能更好地反映时序之间的距离. 对于特征提取的方法,其关键是找到不同类时序间具有区别的特征,Shapelet 算法及其衍生出来的一系列算法是特征提取的一类经典算法. Shapelet 表示两类时序数据之间最具有差异性的子段,时序之间的相似性可以通过与 Shapelet 之间的距离来度量. 这类算法的时间复杂度太高,最近文献[4]将 CNN 引入到时序分类中提出了 MCNN,MCNN 通过对原始的时序数据进行不同尺

度和频率的变化提取更加丰富的特征，提高 CNN 的泛化性能. 由于 CNN 强大的特征提取能力，CNN 在时序分类领域取得了极大的成功. 除此之外，有学者针对时序数据设计了一些集成学习方法，在 BOSS[6] 中，作者按照不同的窗口大小对时序数据进行离散傅里叶变化得到基本特征；然后将其量化为符号，最后根据不同窗口的最近邻度量结果集成投票. COTE[7] 在时域、频域和 Shapelet 变化域等构造分类器；最后集成起来分类.

本文针对电力负荷运行曲线提出了一种基于集成学习的分类方法，通过对序列的分类来识别不同的设备. 不同于以上提到的针对时序数据设计的集成学习算法，本文提出的集成策略更简单有效. 我们使用 CNN 作为基分类器，通过粒度划分来得到不同的训练数据集. 集成学习的核心目标是设计“好而不同”的基学习器[8]，本文方法中 CNN 保证了基学习器的性能. 不同粒度的划分，可以看作是样本扰动，而对于 CNN 这类不稳定的基学习器来说，样本扰动对增加基分类的多样性是有效的实验结果也表明本文提出的集成学习算法是有效的.

2 基于粒度划分的集成学习方法

2.1 样本的扰动

集成学习中，数据扰动一般通过采样来实现. 比如 Bagging 里的自助采样，Boosting 里的序列采样. 与这些采样方法不同，我们通过对原始样本不同粒度的划分达到样本扰动的目的. 电力负荷数据中，一条负荷数据包含 96 个值的记录. 若对一个数据进行 1/2 粒度的划分，将得到 0-47，48-95 两个等长的新数据，如图 1 所示. 若进行 1/3 粒度的划分，将得到 0-31，32-63，64-95 三个等长的数据. 1/4，1/5，…，$1/n$ 粒度的划分以此类推.

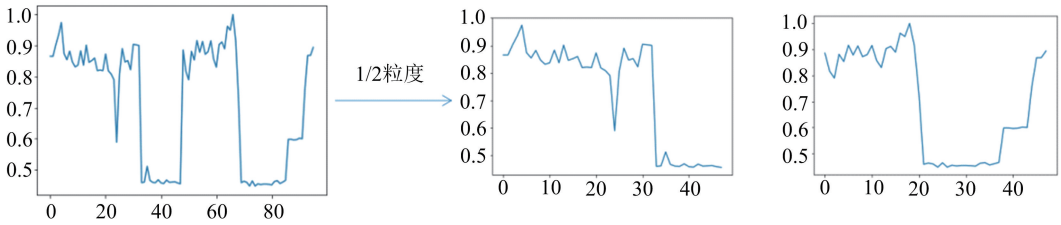

图 1 电力负荷数据的 1/2 粒度划分

Fig. 1 1/2 grained partition of power load profile

按照以上的规则，将所有的样本进行 1/2 粒度，可以得到 2 个新的数据集，进行 1/3 粒度划分可以得到 3 个新的数据集. 我们可以得到 N 个不同的数集按照我们所需要的任意划分方式. $N=\sum_{i=1}^{n} n$，其中 n 表示 n 中粒度划分方式，依次从 1，1/2，1/13，…，$1/n$，1 粒度表示不对样本进行划分. 与自助采样和序列采样不同，本文构造的新数据集并不包含相同的样本. 与随机森林随机[9]选择属性不同，对非结构化的数据比如语音、图像、时序数据等，其原始数据的属性并不明显，往往需要更进一步的抽象来手工提取特征，比如图像中的边缘信息，颜色分布直方图，时序数据中傅里叶变化[10]的系数，离散小波变换[11]的系数等. 随机森林里随机挑选属性的思想很难应用在这些非结构化数据上. 除此之外，本文的粒度划分方式保留了空间信息，有利于进一步的研究分析.

2.2 基学习器的选择

集成学习的关键是设计出“好而不同”[8]的基学习器. “好”容易度量，对于分类问题，我们可以认为一个基学习器分类精度高就是“好”. 对于回归问题，我们可以认为最小均方误差越小越“好”. “不同”也可以说多样性并不是很容易衡量，一般来说有 4 种策略来增强多样性：数据样本扰动、输入属性扰动、输出表示扰动、算法参数扰动[12]. 我们使用了数据样本扰动的策略来增强基学习器的多样性，对常见的基学习器可以根据它们是否对样本扰动敏感将它们分为“稳定的基学习器”和“不稳定的基学习器”[12]. “稳定基学习器”对数据样本扰动并不敏感，比如支持向量机、朴素贝叶斯、k 近邻学习器. “不稳定基学习器”对数据样本扰动敏感，样本的变化可以导致学习器显著的变化，比如决策树和神经网络等. 由于 CNN 已经被证明在时序分类领域取得了成功[4]，我们使用简单的 3 层卷积的 CNN 作为基学习器. 不用刻意针对每个数据集去调整每个 CNN 的超参数，这样时间复杂度会比较高. 在实验部分，我们只在没有进行任何粒度划分的原始数据集上调整 CNN 的超参数. 在其他的数据集上，我们均使用与之相同的超参数.

2.3　权重计算以及预测

N 个不同的数据集可以训练出 N 个不同的基学习器，由于每个基学习器的能力并不一样，使用相同的权重进行投票，并不是合理的选择. 我们给 N 个不同的基学习器计算出 N 个不同的权重，第 i 个基学习器权重根据其在验证集上的测试精度作为其权重，对于验证集上精度低于随机乱猜的基学习器，我们将该基学习器丢掉. 假设，N 个基学习器在验证集上的精度都大于随机乱猜的精度，我们可以得到 $W_1, W_2, \cdots, W_N$，N 个不同的基学习器权重. 最后加权投票时的权重并不直接使用这些权重，我们设定了一个投票权重的置信度值 $C_i = W_i - 1/c$. c 表示数据集中类别数，$1/c$ 表示类别平衡下的随机乱猜的精度，我们使用 C_i 作为第 i 个基分类器最后投票的权重. 相对于 W_i，使用 C_i 作为基分类器的权重，使得最后集成学习器的结果能够更加偏向好的基学习器.

对于一个新的测试样本，我们按照训练集上的粒度划分方式对测试样本进行相同的划分，得到了 N 个新样本. 将这 N 个新样本分别送到 N 个训练好的基分类器中进行预测并得到 N 个结果，每个结果赋予该基分类器的投票权重. 最后，将 N 个结果中预测相同类的进行加权，并选出权值最大的那个类作为最终的预测结果.

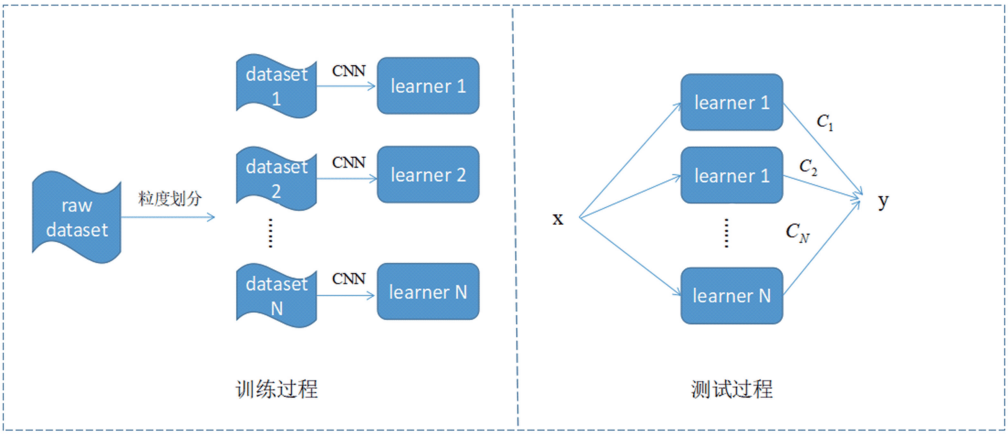

图 2　我们模型的训练和测试过程

Fig. 2　The training and testing process of our model

3　实验和结果

3.1　数据集

本文将在一个实际的工厂电力设备负荷记录的数据集上测试我们的算法，这些数据来自华东地区某省某工厂 10 个不同电力设备一段时间的记录值. 所有的值都已经做了归一化处理，由于收集的负荷记录中包含缺失值，我们使用向前和向后补齐的方式补齐所有缺失值. 向前补齐就是使用后面一个时刻的值补齐前面的缺失值，向后补齐就是使用前面一个时刻的值补齐后面一个时刻的缺失值.

我们将数据按照 5∶2 划分成训练集和测试集，又将训练集分成了 5 中不同的规模，分别是 500-size，1000-size，1500-size，2000-size，2500-size，表示数据集中的样本数量. 我们在这 5 个不同规模的数据集上做了实验.

3.2　Baseline

我们在 5 个不同规模的数据集上均与单个的 CNN 进行比较. 本文并不与传统的分类算法如 KNN，SVM 等做比较，这是因为这些算法与 CNN 在时序数据分类上有比较大的差距. 最后，使用精度作为评价标准. 因为在我们的数据集中十类样本的数量是平衡的，所以精度可以衡量分类模型的好坏.

3.3　实验和结果

我们在 5 个不同规模的数据集上，与单个 CNN 比较精度，只在不进行任何划分的数据集上对 CNN 进行调参，在其他进行粒度划分后的数据集上均使用相同的参数. 由于本文算法本质上是 Bagging 的，所以天然适合并行训练其余的 N−1 个基分类器.

为了简单起见我们只训练 10 个基分类器，即我们只进行 1/2，1/3，1/4 这三种粒度划分. 我们在 5 个不同规模数据集上数据如表 1 所示. 除此之外，我们也展示了 10 个基分类器在这 5 个分类器上的精度，如表 2 所示.

表 1 五个不同规模训练集下的测试精度

Tab. 1 The testing accuracy under five different size training sets

	500	1000	1500	2000	2500
CNN	0.533	0.62	0.626	0.749	0.806
Our Model	0.595	0.653	0.664	0.798	0.837

从表 1 可以看出,随着训练集规模的增加,模型的泛化性能越来越好. 10 个基分类器的在 5 个不同规模下的集成效果要显著的好于单个 CNN 分类器. 从表 2 可以看出,一个较好的基分类器与若干稍弱的基分类器结合起来可以构造出一个更强的集成学习器.

表 2 五个不同规模训练集下十个基分类器的测试精度

Tab. 2 The testing accuracy of ten basic classifiers under five different size training sets

Size	0	1	2	3	4	5	6	7	8	9
500	0.533	0.495	0.536	0.536	0.474	0.524	0.488	0.436	0.493	0.5
1 000	0.62	0.54	0.578	0.593	0.51	0.559	0.528	0.493	0.582	0.546
1 500	0.626	0.581	0.616	0.59	0.554	0.585	0.555	0.525	0.562	0.57
2 000	0.749	0.66	0.733	0.648	0.651	0.703	0.587	0.612	0.66	0.657
2 500	0.806	0.728	0.774	0.723	0.7	0.725	0.65	0.713	0.72	0.69

4 结论

本文提出了一种粒度集成的方法用于电力设备负荷曲线的识别. 通过对负荷曲线进行不同粒度的划分,达到数据样本扰动的目的提高集成学习的多样性. 本文选取 CNN 作为我们的基学习器,一是因为 CNN 在时序分类领域取得的成功,二是因为 CNN 属于不稳定的基学习器,样本稍微扰动就会导致 CNN 显著的不同. 通过基学习器的权重给出了每个基分类器最后投票的置信度值. 最后我们在实际的电力数据集上验证了本文算法的有效性.

参考文献(References)

[1] QIU J, WANG H F, LIN D Y, et al. Nonparametric regression-based failure rate model for electric power equipment using lifecycle data[J]. IEEE Transmission on Smart Grid, 2015, 6(2): 955-964.

[2] YE L X, KEOGH E. Time series shapelets: A new primitive for data mining[C]// 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France: ACM, 2009: 947-956.

[3] DING H, TRAJCEVSKI G, SCHEUERMANN P, et al. Querying and mining of time series data: Experimental comparison of representations and distance measures[J]. Proceedings of the VLDB Endowment, 2008, 1(2): 1542-1552.

[4] CUI Zhicheng, CHEN Wenlin, CHEN Yixin. Multi-scale convolutional neural networks for time series classification[C]// Conference on RR, http://arxiv.org/abs/1603.06995 [2016-022].

[5] CHEN T Q, GUESTRIN C. XGBoost: A scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA: ACM, 2016: 785-794.

[6] SCHÄFER P. The BOSS Is Concerned With Time Series Classification in the Presence of Noise[M]. Kluwer Academic Publishers, 2015.

[7] BAGNALL A, LINES J, HILLS J, et al. Time-series classification with COTE: The collective of transformation-based ensembles [J]. IEEE Transactions on Knowledge & Data Engineering, 2015, 27(9): 2522-2535.

[8] ZHOU Z H. Ensemble Methods: Foundations and Algorithms[M]. Taylor & Francis, 2012.

[9] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[10] AGRAWAL R, FALOUTSOS C, SWAMI A N. Efficient similarity search in sequence databases[C]// International Conference on Foundations of Data Organization and Algorithms. Springer, 1993: 69-84.

[11] CHAN K P, FU W C. Efficient time series matching by wavelets[C]// Proceedings of the International Conference on Data Engineering. Sydney: IEEE, 1999: 126-133.

[12] 周志华. 机器学习[M]. 北京:清华大学出版社, 2016.