基于组合支持向量回归的排污企业生产识别

靳　旦,唐　伟
(国网四川省电力公司电力科学研究院,四川 成都　610041)

摘　要:在电网大数据结合环保监管的背景下,研究了以企业用电和纳税数据为特征的企业生产状态识别方法。实际输入特征通常存在异常和部分缺失的情况,采用回归分析解决数据异常和部分缺失的问题,提高了分析结果的鲁棒性。建立支持向量回归模型来识别污染企业生产状况,通过网格搜索选择多个支持向量回归组合模型来识别污染企业生产状况,增加了模型的泛化能力,提高了分类精度。最后,实际测试结果验证了所提出的基于组合支持向量回归的排污企业生产识别方法的精确性和适用性。

关键词:组合支持向量回归;网格搜索;生产识别
中图分类号:TP701　**文献标志码**:A　**文章编号**:1003-6954(2020)03-0029-04

Identification of Production for Pollution Emission Enterprises Based on Ensemble Support Vector Regression

Jin Dan, Tang Wei
(State Grid Sichuan Electric Power Research Institute, Chengdu 610041, Sichuan, China)

Abstract: Under the background of big power data and environmental supervision, the identification of enterprise production status is studied, which is characterized by enterprise electricity consumption and tax payment. Regression analysis is adopted to solve the problem of data anomaly and partial missing, which improves the robustness of analysis results. Support vector regression model is established to identify the production status of pollution emission enterprises, and multiple support vector regression models are selected through grid search to identify the production status of pollution emission enterprises, which increases the generalization ability of the model and improved the classification accuracy. Finally, the simulations confirm the accuracy and applicability of the proposed method based on ensemble support vector regression (e-SVR).

Key words: ensemble support vector regression, gird search, identification of production

0　引　言

信息技术和互联网技术的快速发展使得大数据分析和人工智能等新技术应用应运而生,在建设电力物联网的背景下,大数据和人工智能新技术与能源行业相结合引发了越来越多研究人员的关注[1]。电力大数据平台在全面管控电网营销服务和企业安全生产等方面获得显著成效[2-3]。一种应用电网用电信息采集系统数据以实现城市大气污染排放在线管控的方法被提出。该方法基于大数据手段,将电网企业用电信息采集系统电量与环保管控数据进行批量化关联分析,无需新增硬件设备,实现了城市企业大范围在线监管,其中,排污企业对于环保措施的响应程度评判是电力数据在环保应用的一个重要功能。

支持向量机(support vector machine,SVM)是在分类、回归和其他学习任务方面广受欢迎的一种机器学习方法,在计算机视觉、自然语言处理、神经成像、生物信息学等领域已有成功的应用[4]。支持向量机一般分为3类:支持向量分类(support vector classification, SVC)、支持向量回归(support vector regression, SVR)和一类支持向量机(one-class support vector machine, 1-SVM)[5]。其中:根据分类特征,支持向量分类可分为两值分类和多值分类;支持向量回归用于处理数据回归问题;支持向量机还可实现一种特殊的一类分类问题,有学者将其称为一类支持向量机(one-class support vector machine,

1-SVM)，在实际中通常应用于异常值检测[6]。

下面研究以企业用电和纳税信息为特征的排污企业生产状态识别方法，考虑到实际输入特征异常和部分缺失的情况，采用回归分析解决数据异常和部分缺失的问题，提高了分析结果的鲁棒性；通过网格搜索选择多个支持向量回归组合模型来识别污染企业生产状况，增加了模型的泛化能力，提高了分类精度。

1 支持向量机数学模型

从模式分类中可分离模式的情况下了解支持向量机是如何工作的可能是最容易的。给定可线性或非线性分离的训练样本，支持向量机通过非线性核函数映射，生成一个超平面作为决策曲面，使得正例和反例之间的隔离边缘被最大化。

考虑训练样本$\{(x_i, y_i), i=1,2,3...N\}$，$x_i$为输入模式的第$i$个样例，$y_i$为对应的期望响应，用于分离的超平面形式的决策曲面方程为

$$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+\boldsymbol{b}=0 \tag{1}$$

式中：$\boldsymbol{x}$为输入向量；$\boldsymbol{w}$为权值向量；$\boldsymbol{b}$为偏置。对于一个给定的权值向量$\boldsymbol{w}$和偏置$\boldsymbol{b}$，支持向量机的目标就是找到一个特殊的超平面，这个超平面的分离边缘最大。支持向量机是一个二次规划问题，数学推导如下。

当样本中$y_i=+1$和$y_i=-1$代表的两类模式是线性可分时，式(1)可以为

$$\begin{aligned}\boldsymbol{w}^{\mathrm{T}}x_i+\boldsymbol{b}\geqslant 0 \quad \text{当 } \mathrm{y_i}=+1\\ \boldsymbol{w}^{\mathrm{T}}x_i+\boldsymbol{b}\leqslant 0 \quad \text{当 } \mathrm{y_i}=-1\end{aligned} \tag{2}$$

考虑使式(2)等号成立的那些点，也就是距离超平面最近的两类点，只要成比例地调整$\boldsymbol{w}$和$\boldsymbol{b}$的值就能保证这两类点的存在，且对分类结果没有任何影响。设2个超曲面为H_1、H_2。

$$\begin{aligned}H_1:\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+\boldsymbol{b}=1\\ H_2:\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+\boldsymbol{b}=-1\end{aligned} \tag{3}$$

超曲面H_1到原点的距离为$|1-\boldsymbol{b}|/\|\boldsymbol{w}\|$，超平面$H_2$到原点的距离为$|-1-\boldsymbol{b}|/\|\boldsymbol{w}\|$。所以，$H_1$和$H_2$之间的距离为$2/\|\boldsymbol{w}\|$。因此，要使分离间隔最大就是使$\|\boldsymbol{w}\|$最小，为一个二次规划问题。

对于非线性问题，可以通过非线性变换转化为高维空间的线性问题。因此，对于非线性分类，首先，采用一个映射φ将数据映射到一个高维空间。此时，在高维特征空间中就可对输入数据进行线性分类，映射回原空间后就成了输入数据的非线性分类。支持向量机采用了一个核函数$K(x,y)$代替高维空间的内积运算，避免高维空间的复杂运算。为使得所有样本都能被分离超平面正确分类，增加模型的鲁棒性，可采用松弛变量解决这个问题，因此优化问题为

$$\min\left(\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{l}\xi_i\right) \tag{4}$$

式中：$\boldsymbol{w}$是权值向量；C为惩罚因子；ξ_i为松弛变量。

约束为

$$\begin{aligned}\boldsymbol{w}^{\mathrm{T}}\Phi(x_i)+\boldsymbol{b}\geqslant 1-\xi_i \quad i=1,\cdots\cdots,l\\ \xi_i\geqslant 0 \quad i=1,\cdots\cdots,l\end{aligned} \tag{5}$$

式中：$\Phi(x_i)$为核函数。

2 排污企业生产识别

2.1 排污企业生产识别整体流程

基于组合支持向量回归的排污企业生产识别的输入数据为企业的用电量和税收值，与企业生产状态有强相关性。对输入数据进行归一化处理可直观迅速判断企业的基本运行情况，同时大量简化了计算。规则化后的输入数据存在负值或缺失的情况，这是异常的输入数据。通过对企业用电量和税收值的历史回归分析，可校正负值的输入数据，预测缺失的输入数据，提高了分析结果的鲁棒性。

将输入数据分为3部分：训练数据、验证数据和测试数据。输入数据用来训练支持向量回归的超参数，不同的超参数对应一个支持向量回归模型。通过网格搜索可以确定多个支持向量回归模型，验证数据用来筛选已确定的支持向量回归模型，得到最优的支持向量回归模型集用于组合回归判断，可提高单一模型的精度。最后，将最优的支持向量回归模型集来测试历史数据。排污企业生产识别整体流程如图1所示。

2.2 排污企业数据来源

排污企业的主要数据有企业类型、企业注册地、企业纳税、企业用电等，取自于不同的机构。其中，企业类型和企业注册地来自四川省工商局，企业纳税历史数据来自于四川省税务局，企业用电历史数据来自国网四川省电力公司用电信息采集系统和营

图 1　排污企业生产识别总体流程

销系统。将历史数据分为 3 部分: 训练数集、验证数集和测试数集。其中,训练数据用来确定模型的参数; 验证数据用来做模型验证,选定预测误差小的超参数组合,提高总模型的精度; 最后,测试数据用来做模型测试及分析结果。

2.3　输入特征选择与正则化

支持向量机的输入特征选择为企业用电量和纳税值。这两个特征与企业生产密切相关,输出数据为企业开工判据。

输入数据为企业用电量和纳税值,输出数据为企业开工判据,其中,输入数据的幅值远远大于输出数据的幅值。为了降低运算难度,输入数据的大小被规则化,企业用电量被企业配电变压器容量整除,纳税值被最大纳税值整除,则企业用电量和纳税值规则化后的范围均为 [0,1]。

$$\text{企业用电量}^{*} = \frac{\text{企业用电量}}{\text{企业配电变压器容量}}$$

$$\text{税收值}^{*} = \frac{\text{税收实际值}}{\text{最大税收值}} \tag{6}$$

实际情况中存在输入特征值为负或者缺失的异常情况,为处理这种异常值情况,通过历史数据回归建立起企业用电量和纳税值的关系,如图 2 所示。

图 2 给出了企业历史的税收值和用电量,存在异常与缺失,图 3 对缺失与异常做了标记,通过回归分析,可对异常输入特征进行修正,并填补缺失的特征,如图 4 所示。

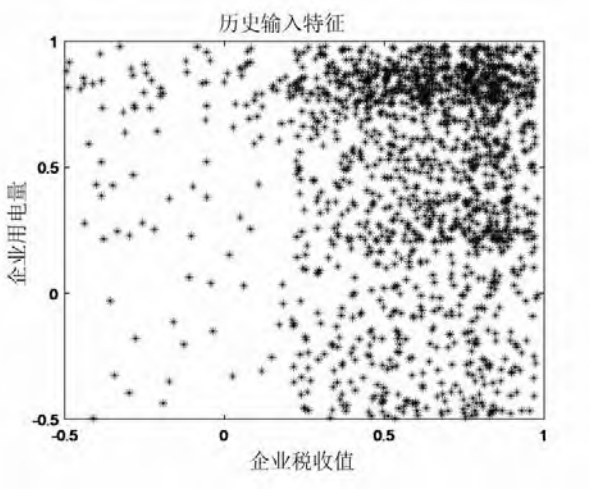

图 2　企业历史输入特征(存在特征异常与缺失)

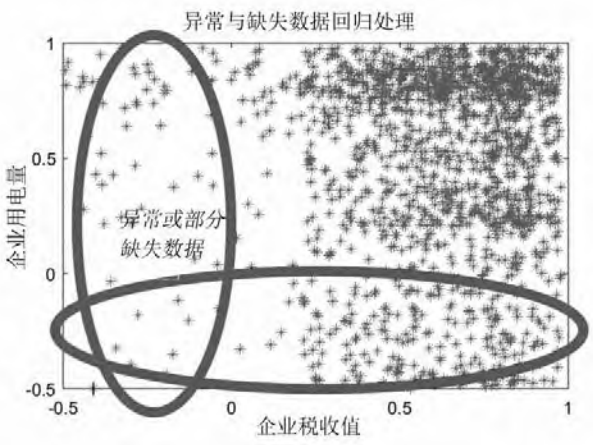

图 3　异常与缺失特征辨识

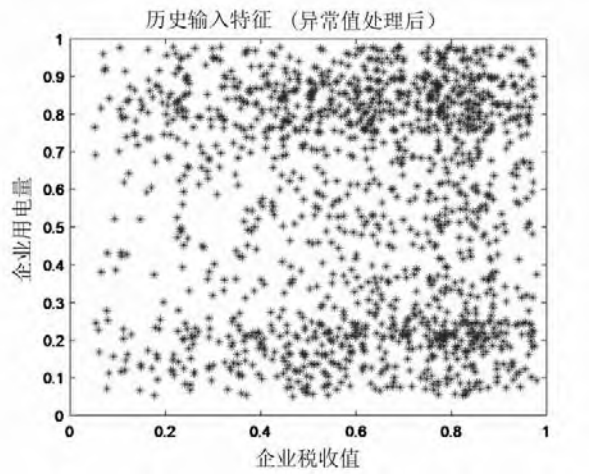

图 4　异常与缺失特征回归处理

2.4　核函数非线性映射

使用非线性映射 φ 把训练数据映射到一个高维特征空间,然后在高维特征空间里进行线性回归,映射回原空间后就成了输入空间的非线性分类。用核函数 $K(x,y)$ 就可以实现非线性回归,如图 5 所示。

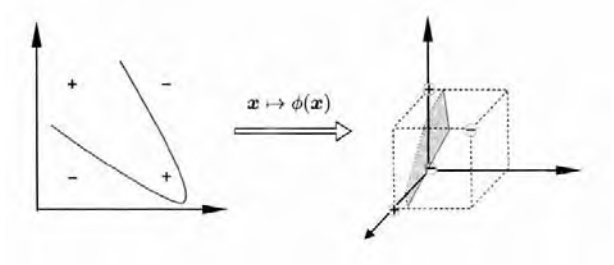

图 5　核函数的非线性映射

2.5　网格搜索与验证

采用网格搜索法来选择了 441 组超参数(即不同的 C,γ,ε 组合),一组超参数可以确定一个 SVM 模型,选择网格搜索多个超参数以提高模型的边化能力,C 为惩罚因子,γ 为核函数的参数,ε 为一个大于 0 的常数。在使用相同数据集的情况下,训练了 441 个不同的 SVM 模型。用一部分数据来做验证,避免过拟合。通过网格搜索得到的 441 个模型,在验证集上取误差最小的前 25 个模型,测试数据的最后结果取 25 个回归模型的平均值。

网格搜索法是一种直接的方法,它将不同组合的 γ、C 和 ε 值逐个进行测试,查看情况,网格搜索中,令:

$$C = (2^{(-5)}, 2^{(-3)}, 2^{(-1)}, 2^{(1)}, 2^{(3)}, 2^{(5)}, 2^{(7)}, 2^{(9)}, 2^{(11)})$$
$$\gamma = (2^{(-15)}, 2^{(-13)}, 2^{(-11)}, 2^{(-9)}, 2^{(-7)}, 2^{(-5)}, 2^{(-3)})$$
$$\varepsilon = (2^{(-15)}, 2^{(-13)}, 2^{(-11)}, 2^{(-9)}, 2^{(-7)}, 2^{(-5)}, 2^{(-3)}) \quad (7)$$

不同的参数组合依次求解,得到最优的超参数。

基于组合支持向量回归的排污企业生产识别流程如图 6 所示。

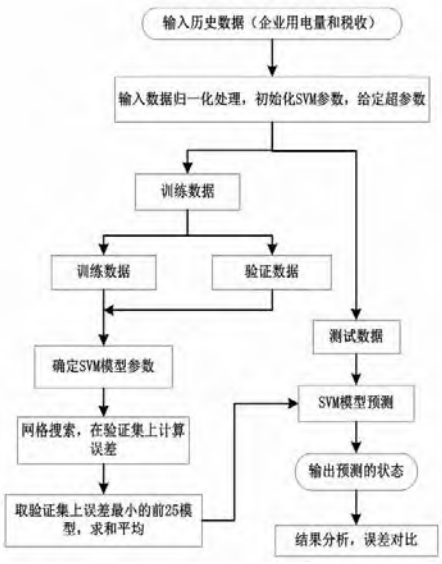

图 6　基于组合支持向量回归的排污企业生产识别流程

3　测试结果分析

对企业 1500 个实际生产数据进行测试识别。输入数据为这 1500 个实际生产下的纳税值和用电量,其中,输入数据存在部分缺失和异常。组合支持向量机模型给出对于 1500 个输入特征下对应的生产判断。

输入的数据中,纳税数据小于 0 时为异常值,纳税数据为 0 时为缺失值。异常值和缺失值都是因为实际管理等原因造成的真实数据不能查询。

输入数据的总数为 1500 个,异常即小于 0 的数据为 313 个,占总输入数据的 20.87%;部分缺失数据数量为 29 个,占总输入数据的 1.93%,如表 1 所示。

表 1　输入数据分析

数据类型	数量/个	比例/%
总输入数据	1500	100.00
异常数据	313	20.87
部分缺失数据	29	1.93

通过回归模型对异常数据进行处理,如图 7 所示,异常和部分缺失输入数据进行回归处理后,其规则化后的范围为(0,1)。

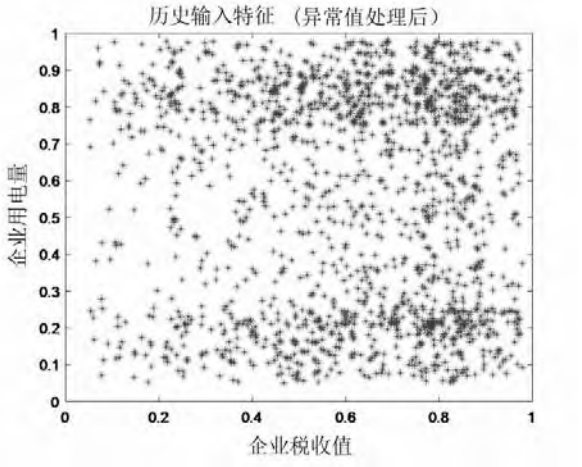

图 7　异常和部分缺失数据处理后的输入特征数据

通过组合支持向量回归模型对异常和部分缺失值处理后的数据进行生产状态识别,结果如图 8 所示。

组合支持向量回归模型识别企业生产状态结果如表 2 所示。其中,正确识别的生产状态 1484 个,识别正确率为 98.93%,多个支持向量机组合模型具有分类精度高的特点。由于异常和部分缺失数据

(下转第 77 页)

参考文献

[1] 杜至刚. 中国特高压电网发展战略规划研究[D]. 济南: 山东大学, 2008.
[2] 李光范, 王晓宁, 李鹏, 等. 1000 kV 特高压电力变压器绝缘水平及试验研究[J]. 电网技术, 2008, 32(3): 1 – 6.
[3] 邢运民, 罗建, 周建平, 等. 变压器铁心剩磁估量[J]. 电网技术, 2011, 35(2): 169 – 172.
[4] 仇明. 大型变压器铁心剩磁的危害及消除方法[J]. 变压器, 2018(2): 74 – 75.
[5] 余克光. 电力变压器铁心剩磁检测及消除[J]. 自动化应用, 2012(3): 73 – 74.
[6] 胡海将. 变压器剩磁对启机的影响[C]//中国电机工程学会. 中国电机工程学会年会论文集, 2013.
[7] 陈文臣, 雷晓燕, 王磊. 电力变压器铁心剩磁检测方法研究[J]. 陕西电力, 2009, 37(10): 45 – 48.
[8] 刘勇, 陈凌, 李英锋, 等. 一种大型电力变压器剩磁检测方法: CN201510213548.9[D]. 2015 – 08 – 12.
[9] 张建军, 刘宏亮, 陈志勇, 等. 一种基于最小二乘法的变压器铁心剩磁检测方法: CN201610382451.5[P]. 2016 – 11 – 09.
[10] 戈文祺. 电力变压器铁芯剩磁的仿真、测量与削弱[D]. 天津: 河北工业大学, 2014.
[11] 周建平, 罗建, ZHOUJian – ping, 等. 变压器铁心剩磁的一种估算方法[J]. 热力发电, 2010, 39(3): 61 – 64.
[12] 戈文祺, 汪友华, 陈学广, 等. 电力变压器铁心剩磁的测量与削弱方法[J]. 电工技术学报, 2015, 30(16): 10 – 16.
[13] 李钜, 乌云高娃, 刘涤尘, 等. Preisach 模型剩磁计算与抑制励磁涌流合闸角控制规律[J]. 电力系统自动化, 2006, 30(19): 37 – 41.
[14] 李景丽, 贺鹏威, 邱再森. 电力变压器铁心剩磁测量方法研究综述[J]. 高压电器, 2018(7): 98 – 105.
[15] 梁汉城, 陈向胜, 苏全. 降低大型变压器铁心剩磁的方法和相关测试[J]. 电世界, 2013, 54(6): 38 – 39.

作者简介:

谭志红(1966), 工学博士, 主要从事输变电设备运检管理、SF_6 气体局部放电下分解组分研究等工作。

(收稿日期: 2020 – 03 – 04)

(上接第 32 页)

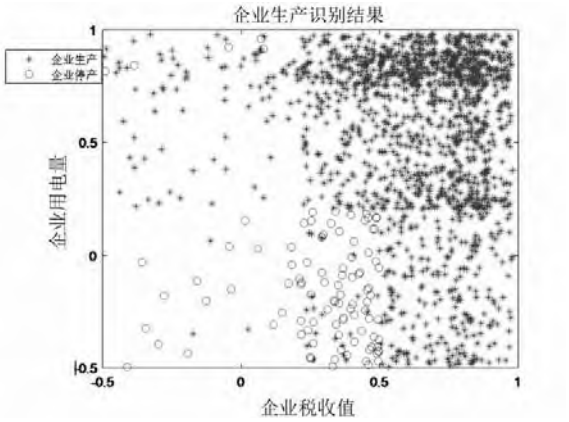

图 8 企业生产状态识别结果

有 342 个, 未进行异常和缺失值处理时, 正确识别 1142 个, 识别正确率为 76.13%, 异常和缺失值回归处理提高企业生产状态识别率 22.8%。

表 2 组合支持向量回归模型识别企业生产状态结果

输入数据	正确识别个数	正确识别率/%
回归处理后	1484	98.83
未回归处理	1142	76.13

4 结 语

提出了一种综合考虑企业用电和纳税信息的污染企业生产状况识别方法。该识别方法考虑了实际应用过程中的数据部分病态的问题, 采用回归分析, 解决输入特征异常和部分缺失的问题, 提高了分析结果鲁棒性; 同时, 通过网格搜索选择多个支持向量机组合模型识别污染企业生产状况, 增加了模型的泛化能力, 提高了分类精度。

参考文献

[1] 牛哲文, 郭采珊, 唐文虎, 等. “互联网 + 智慧能源”的技术特征与发展路径[J]. 电力大数据, 2019, 22(5), 6 – 10.
[2] 杜若, 谢川, 吴群艳. 电力环保大数据平台开发及智能运用[J]. 电力大数据, 2017, 20(8): 64 – 67.
[3] 罗勇智. “电力大数据 + 环保监管”助力蓝天保卫战[J]. 大众用电, 2019, 34(3): 15 – 16.
[4] Nekkaa M, Boughaci D. A Memetic Algorithm with Support Vector Machine for Feature Selection and Classification[J]. Memetic Computing, 2015, 7(1): 59 – 73.
[5] 周志华. 机器学习[J]. 航空港, 2018(2): 94 – 94.
[6] Sebastian Raschka. Python Machine Learning[M]. Packt Publishing, 2014.

作者简介:

靳 旦(1995), 助理工程师, 主要从事大数据开发, 数据分析与挖掘工作;

唐 伟(1990), 工程师, 主要从事大数据计算模式, 存储和管理工作。

(收稿日期: 2020 – 03 – 25)