文章编号: 1001-0920 (2014) 09-1661-06 DOI: 10.13195/j.kzyjc.2013.0774

基于贪心核特征提取方法的中期峰值负荷预测

李 军

(兰州交通大学 自动化与电气工程学院，兰州 730070)

摘 要: 针对中期电力负荷预测, 提出基于贪心核主元回归 (GKPCR)、贪心核岭回归 (GKRR) 的特征提取建模方法. 通过对核矩阵的稀疏逼近, GKPCR 和 GKRR 两种贪心核特征提取方法旨在寻找特征空间中数据的低维表示, 计算需求低, 适用于大数据集的在线学习. 将所提出的方法应用于不同地区的电力负荷中期峰值预测, 并与现有预测方法进行了比较. 实验结果表明, 在同等条件下, 所提出的方法能有效地改进预测精度, 而且性能更好, 显示了其有效性和应用潜力.

关键词: 贪心算法；特征提取；核主元回归；核岭回归；负荷预测

中图分类号: TN911 **文献标志码:** A

Greedy kernel feature extraction method for medium term electricity peak load forecasting

LI Jun

(School of Automation and Electrical Engineering，Lanzhou Jiaotong University，Lanzhou 730070，China．E-mail: lijun691201@mail.lzjtu.cn)

Abstract: For midterm electricity load forecasting, modeling methods of feature extraction based on greedy kernel principal component regression(GKPCR) as well as greedy kernel ridge regression(GKRR) are proposed. On the basis of sparse approximation of the kernel matrix, the proposed greedy kernel feature extraction methods aim to find a lower dimensional representation of data embedded in the feature space. Modeling methods of greedy kernel feature extraction have low computational requirements and allow on-line processing of large data sets. The proposed GKPCR and GKRR methods are then applied to electricity peak load forecasting instances in different areas. Compared to existing other kernel-based forecasting methods, experimental results show that, the employed methods may significantly improve the accuracy of peak load forecasting under the same condition, which have considerably better performance, and show the effectiveness and applicability.

Key words: greedy algorithm；feature extraction；kernel principal component regression；kernel ridge regression component；load forecasting

0 引 言

电力负荷的中长期预测对于电力部门有计划地制订电网规划、增容和改扩建方案至关重要. 构建适应性更强的电力负荷中长期预测模型, 对智能电网的发展规划和运行调度具有重要意义.

负荷预测通常依靠历史负荷数据和气候等因素构建预测模型. 近年来, Volterra 滤波器[1]、神经网络[2]、支持向量机 (SVM)[3-4]等预测方法已成功应用于电力负荷中短期预测领域, 而且 SVM 和高斯过程 (GP) 等基于核的学习方法被广泛关注. 另一方面, 将岭回归 (RR)、主元回归 (PCR) 和偏最小二乘 (PLS) 等方法引入高维特征空间, 可形成核岭回归 (KRR)[5]、核主元回归 (KPCR)[6]、核偏最小二乘 (KPLS)[7]等基于核学习的方法, 该类方法的求解避免了 SVM 中二次规划问题的求解, 提高了计算效率, 在建模方面同样具有很好的应用潜力. 然而, 与 SVM、GP 方法相似, 这些方法也存在随着数据集的增大核矩阵计算较为复杂的困难. 为了解决上述问题, 可以基于期望最大化算法 (EM)[8]迭代地计算提取主元特征, 也可以通过对核矩阵进行稀疏逼近, 使用基于贪心逼近算法的贪

收稿日期: 2013-06-10；**修回日期:** 2014-01-13.

基金项目: 甘肃省高等学校基本科研业务费专项资金项目(620026)；甘肃省教育厅硕导项目(1104-09).

作者简介: 李军(1969–), 男, 教授, 博士, 从事计算智能、机器学习、系统建模控制等研究.

心核主元分析 (GKPCA) 方法[9-10]. GKPCA 方法不仅适宜于大数据集的在线学习, 还能很好地挖掘数据的内在特征, 增强模型的推广性.

本文在贪心核逼近算法的基础上, 提出基于贪心核特征提取建模的 GKPCR 和 GKRR 方法, 并将其应用于不同地区的中期电力负荷峰值预测的实例中, 以评价预测的效果和性能. 在同等条件下, 与现有多种预测方法进行比较, 验证了所提出方法的有效性.

1 贪心核主元分析

1.1 贪心核逼近算法

与 PCA 方法提取特征不同, GKPCA 使用贪心核逼近算法提取非线性特征, 其基向量的选取直接取自训练数据而非训练数据的线性组合, 且满足数据重构误差最小, 因此可获得对训练数据集逼近的次优解.

令 $\boldsymbol{\Gamma}=\{\phi(\boldsymbol{x}_1),\phi(\boldsymbol{x}_2),\cdots,\phi(\boldsymbol{x}_l)\}$ 为训练数据集映射至高维特征空间的集合, $\boldsymbol{I}=\{1,2,\cdots,l\}$ 为训练样本集 $\boldsymbol{\Gamma}$ 的索引集, $\boldsymbol{J}=\{j_1,j_2,\cdots,j_m\}$ 为从训练样本集 $\boldsymbol{\Gamma}$ 中选择的样本子集 $\boldsymbol{S}$ 的索引集, 即 $\boldsymbol{S}=\{\phi(\boldsymbol{x}_i):j\in J\}$. 首先利用选择的训练子集作为非线性特征逼近原训练样本数据集, 即

$$\widetilde{\phi}(\boldsymbol{x}_i)=\sum_{j\in J}[\boldsymbol{\xi}_i]_j\phi(\boldsymbol{x}_j),\ \forall i\in I. \tag{1}$$

其中: $\boldsymbol{J}\in\boldsymbol{I}$ 包含 m 个所选择的样本子集 $\boldsymbol{S}$, 系数向量 $\boldsymbol{\xi}_i\in\mathcal{R}^m$ 为线性组合的系数. 依赖于所选择的子集, 逼近目标是由均方误差所定义的重构误差, 即

$$\varepsilon_{\mathrm{MS}}(I|J)=\frac{1}{l}\sum_{i\in I}\left\|\phi(\boldsymbol{x}_i)-\sum_{j\in J}[\boldsymbol{\xi}_i]_j\phi(\boldsymbol{x}_j)\right\|^2. \tag{2}$$

若给定选择的子集, 则借助核函数, 最优的系数可通过最小化式 (2) 求得, 即

$$\boldsymbol{\xi}_i=\mathop{\arg\min}_{\xi\in R^m}\left\|\phi(\boldsymbol{x}_i)-\sum_{j\in J}[\boldsymbol{\xi}_i]_j\phi(\boldsymbol{x}_j)\right\|^2=$$

$$(\boldsymbol{K}_s)^{-1}\boldsymbol{k}_s(\boldsymbol{x}_i),$$

其中: 核矩阵 $\boldsymbol{K}_s\in\mathcal{R}^{m\times m}$ 由所选择样本子集构成, 且

$$[\boldsymbol{K}_s]_{i,j}=\langle\phi(\boldsymbol{x}_{j_i}),\phi(\boldsymbol{x}_{j_j})\rangle=k(\boldsymbol{x}_{j_i},\boldsymbol{x}_{j_j}),$$

$\langle\bullet\rangle$ 为内积符号; 样本 $\boldsymbol{x}_i$ 与选择的子集 S 中的样本形成核函数向量

$$\boldsymbol{k}_s(\boldsymbol{x}_i)=$$

$$[k(\boldsymbol{x}_{j_1},\boldsymbol{x}_i),k(\boldsymbol{x}_{j_2},\boldsymbol{x}_i),\cdots,k(\boldsymbol{x}_{j_m},\boldsymbol{x}_i)]^{\mathrm{T}}\in\mathcal{R}^m.$$

对训练子集的选择是满足重构误差最小化的优化问题, 即

$$\boldsymbol{J}^*=\mathop{\arg\min}_{\boldsymbol{J}\in\boldsymbol{I}}\varepsilon_{\mathrm{MS}}(\boldsymbol{\Gamma}|\boldsymbol{J}), \tag{3}$$

且子集集合的基数满足 Card $(\boldsymbol{J})=m$.

为加快计算, 逐次选择增加到子集中的样本时, 需要满足在第 t 次迭代时, 均方重构误差的最大上界最小化, 即

$$\varepsilon_{\mathrm{MS}}{}^{(t)}(\boldsymbol{\Gamma}|\boldsymbol{J})=$$

$$\frac{1}{l}\sum_{i\in I}\|\phi(\boldsymbol{x}_i)-\tilde{\phi}^{(t)}(\boldsymbol{x}_i)\|^2\leqslant$$

$$\frac{1}{l}(l-t)\max_{j\in I\backslash J^{(t)}}\|\phi(\boldsymbol{x}_j)-\tilde{\phi}^{(t)}(\boldsymbol{x}_j)\|^2. \tag{4}$$

其中: t 为所选择子集 $\boldsymbol{J}^{(t)}$ 的样本数目, $j\in\boldsymbol{I}\backslash\boldsymbol{J}$ 为 j 属于集合 $\boldsymbol{I}$ 且不包含于集合 $\boldsymbol{J}$.

每次迭代时, 仅有最大逼近误差的样本被增加到样本子集 $\boldsymbol{J}^{(t)}$ 中, 使得式 (4) 不断下降, 可最小化逼近均方重构误差. 进一步利用施密特正交化方法, 渐进选择子集 $\boldsymbol{S}$ 中的样本, 即基向量, 从而形成正交基向量构成的子集. 因此, 贪心核逼近算法的具体实现步骤如下.

Step 1: 令 $t=0$, 置 $\boldsymbol{J}^{(0)}=\{0\}$, $\varepsilon_i{}^{(0)}=\boldsymbol{k}(x_i,x_i)$, $i\in\boldsymbol{I}$;

Step 2: 令 $t=t+1$, 选择 $j_t\in\mathop{\arg\max}_{j\in\boldsymbol{I}\backslash\boldsymbol{J}^{(t-1)}}\varepsilon_j^{(t-1)}$;

Step 3: 计算

$$[\boldsymbol{z}_i]_t=\frac{1}{\sqrt{\varepsilon_{j_t}^{(t-1)}}}\Big(\boldsymbol{k}(\boldsymbol{x}_i,\boldsymbol{x}_{j_t})-\sum_{h=1}^{t-1}[\boldsymbol{z}_i]_h[\boldsymbol{z}_{j_t}]_h\Big);$$

Step 4: 计算

$$\boldsymbol{q}_t=\frac{1}{\sqrt{\varepsilon_{j_t}^{(t-1)}}}\Big(\boldsymbol{\delta}(t)-\sum_{h=1}^{t-1}[\boldsymbol{z}_{j_t}]_h\boldsymbol{q}_h\Big),$$

$$\varepsilon_i{}^{(t)}=\varepsilon_i{}^{(t-1)}-([\boldsymbol{z}_i]_t)^2;$$

Step 5: 置 $\boldsymbol{J}^{(t)}=\boldsymbol{J}^{(t-1)}\bigcup\{j_t\}$;

Step 6: 迭代计算 Step 2 ~ Step 6, 至满足 Card $(\boldsymbol{J})=m$ 或达到期望的重构误差为止.

算法实现步骤中的 $\boldsymbol{\delta}(t)\in\mathcal{R}^m$ 是单位向量, 其第 t 个元素为 1, 其余为零. 每次迭代时, 具有最大化均方误差的一个样本 $\boldsymbol{x}_{j_t}$, 其索引号被包括到子集 $\boldsymbol{J}^{(t)}$ 中, 然后重新计算重构误差. 算法的停止条件为达到期望的均方重构误差或迭代至 $t=m$ 为止. 算法通过选择最优子集 $\boldsymbol{J}^{(*)}$ 可得到基集合 $\boldsymbol{S}$, 参数矩阵 $\boldsymbol{Q}=[\boldsymbol{q}_1,\boldsymbol{q}_2,\cdots,\boldsymbol{q}_m]$ 的维数为 $m\times m$, 由正交基向量 $\boldsymbol{\alpha}_t$ 形成, $\boldsymbol{Q}$ 矩阵定义的正交基集合 $\boldsymbol{W}$ 与基集合 $\boldsymbol{S}$ 等价, 具有同样的线性张成.

1.2 贪心核主元分析

与普通 KPCA 方法相似, GKPCA 方法计算所有样本在所选定的特征空间数据子集上的投影, 以满足数据重构误差最小, 即

$$\boldsymbol{z}=\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{k}_s(\boldsymbol{x})+\boldsymbol{b}. \tag{5}$$

其中: $\boldsymbol{z}\in\mathcal{R}^m$ 为数据映射到核子空间上的投影; $\boldsymbol{z}$ 为 $\phi(\boldsymbol{x})$ 在贪心核逼近算法所选择的正交基向量形成的子集上的投影, 即训练样本在正交基向量集合 $\boldsymbol{W}$ 上

的投影.

考虑所有数据在正交基向量形成的子集上的投影逼近, 则有 $\boldsymbol{Z}=[\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_l]$, 其维数为 $m\times l$. 贪心核逼近算法依据方阵 $\boldsymbol{Z}$ 逼近原始核数据矩阵, 近似核矩阵满足定义 $\widetilde{K}\approx\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}$.

在贪心核逼近算法的基础上, GKPCA 进一步应用常规 PCA 方法, 首先针对维数为 $m\times m$ 的核矩阵 $\boldsymbol{K}_1=\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}$, 提取相应的特征值和特征向量 λ_i、$\boldsymbol{V}^i$ $(i=1,2,\cdots,m)$. $\boldsymbol{V}$ 为特征向量矩阵, $\boldsymbol{V}^i$ 为 λ_i 对应的特征向量, 位于矩阵 $\boldsymbol{V}$ 的第 i 列; 然后, 按特征值大小排列, 选取非线性主元的数目 $p\,(p\leqslant m)$; 最后, 计算投影在 $\boldsymbol{V}^k$ 上的第 k 个非线性主元如下:

$$\beta_k(\boldsymbol{x})=(\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{V})^k\boldsymbol{k}_s(\boldsymbol{x})+b=$$
$$(\boldsymbol{A})^k\boldsymbol{k}_s(\boldsymbol{x})+b=\sum_{i=1}^{m}\alpha_i^k\boldsymbol{k}_s(\boldsymbol{x})+b. \tag{6}$$

其中 $\boldsymbol{\alpha}^k\in\mathcal{R}^m$ 为 $\boldsymbol{A}$ 的第 k 个列向量, 它形成参数矩阵 $\boldsymbol{A}=[\boldsymbol{\alpha}^1,\boldsymbol{\alpha}^2,\cdots,\boldsymbol{\alpha}^p]$, 其维数为 $m\times p$. 因此, 训练数据集矩阵 $\boldsymbol{X}\in\mathcal{R}^{l\times n}$ 将投影在 $\boldsymbol{V}$ 的前 p 个特征向量上, 投影矩阵 $\boldsymbol{P}$ 的维数为 $l\times p$.

KPCA 方法需要针对整个数据集形成的核矩阵提取非线性主元, 其计算复杂度为 $O(l^3)$. 若贪心逼近算法选取的基向量数目为 m, 则 k 的取值至多达到 m, 所以 GKPCA 的计算复杂度为 $O(lm^2)$. 可见, 对于处理大数据集而言, 当 m 选取较小时, GKPCA 的运算效率显著提高.

2　贪心核特征提取建模

将 GKPCA 与标准的 ε-SVM 相结合, 形成多层 SVM, 它对回归建模问题具有稀疏解. 然而, 在高斯噪声下, 二次平方损失函数的最小二乘方法将对回归建模问题具有更好的逼近. 因此, 在 GKPCA 的基础上, 本文提出基于贪心核特征提取的 GKPCR 方法和 GKRR 建模方法.

2.1　GKPCR 方法

考虑在特征空间上的标准回归模型, 有

$$\boldsymbol{y}=\boldsymbol{\Phi}\boldsymbol{w}+\boldsymbol{\varepsilon}. \tag{7}$$

其中: $\boldsymbol{y}$ 为 l 个目标输出所构成的向量; $\boldsymbol{\Phi}$ 为输入数据集映射至特征空间上且中心化后的由 $\phi(\boldsymbol{x}_i)_{i=1}^{l}$ 构成的 $l\times M$ 矩阵, M 为特征空间的维数; $\boldsymbol{w}$ 为回归系数向量; $\boldsymbol{\varepsilon}$ 为系数向量.

为了应用 GKPCA 方法, 考虑将 $\boldsymbol{\Phi}(\boldsymbol{x}_i)$ 首先投影到样本子集 $\boldsymbol{S}$ 上, 再投影到 $\boldsymbol{V}^k$ 上, 形成第 k 个非线性主元 $\boldsymbol{\beta}_k(\boldsymbol{x})$, 由式 (6) 求取.

将所有数据 $\boldsymbol{\Phi}$ 投影到非线性主元上, 式 (7) 改写为

$$\boldsymbol{y}=\boldsymbol{B}\boldsymbol{\eta}+\boldsymbol{\varepsilon}. \tag{8}$$

其中: $\boldsymbol{B}=\boldsymbol{\Phi}\boldsymbol{V}$ 为变换后的一个 $l\times M$ 的数据矩阵, 其各列是正交的; $\boldsymbol{V}$ 为由特征向量 $\boldsymbol{V}^k$ 作为列向量构成的矩阵, 维数为 $M\times M$.

系数向量 $\boldsymbol{w}$ 的最小二乘估计为

$$\hat{\boldsymbol{\eta}}=(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B})^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{y}=\boldsymbol{\Lambda}^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{y}, \tag{9}$$

其中 $\boldsymbol{\Lambda}$ 为对角阵, 由核矩阵 $\boldsymbol{K}_1$ 的特征值 $\lambda_1,\lambda_2,\cdots,\lambda_M$ 构成其元素. 因此, 在特征空间中, 原始回归模型 (7) 中的 $\boldsymbol{w}$ 估计值计算为

$$\hat{\boldsymbol{w}}=\boldsymbol{V}(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B})^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{y}=\sum_{i=1}^{M}\lambda_i^{-1}\boldsymbol{V}^i(\boldsymbol{V}^i)^{\mathrm{T}}. \tag{10}$$

利用 GKPCA 方法提取前 p 个非线性主元, 基于正交的回归变量在特征空间中构建线性回归模型, 形成 GKPCR 模型如下:

$$f(\boldsymbol{x})=\sum_{k=1}^{p}\eta_k\beta_k(\boldsymbol{x})+b=$$
$$\sum_{k=1}^{p}\eta_k\sum_{i=1}^{m}\alpha_i^k\boldsymbol{k}(\boldsymbol{x}_{j_i},\boldsymbol{x})+b=$$
$$\sum_{i=1}^{m}c_i\boldsymbol{k}(\boldsymbol{x}_{j_i},\boldsymbol{x})+b, \tag{11}$$

其中 $c_i=\sum_{k=1}^{p}\eta_k\boldsymbol{\alpha}_i^k,\ i=1,2,\cdots,m$. 实际应用时, 舍弃具有小方差的主元, 能够消除由多重共线性引起的回归模型输出估计误差. 另外, 由式 (11) 可见, 与 KPCR 针对整个数据核矩阵提取非线性主元相比, GKPCR 主元的提取是针对特征空间的子集核矩阵 $\boldsymbol{K}_1=\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}$ 进行的, 因此能够处理大数据集, 满足核矩阵的计算需求.

2.2　GKRR 方法

KRR 是另一种能消除多重共线性的回归建模方法. 考虑式 (7) 的回归模型, 其解需要满足正则化风险最小化, 即

$$(\boldsymbol{w}^*,b^*)=$$
$$\underset{w\in H,b\in R}{\arg\min}\left[\frac{1}{l}\sum_{i=1}^{l}(y_i-f(\boldsymbol{x}_i))^2+\vartheta\|\boldsymbol{w}\|^2\right]. \tag{12}$$

其中: $f(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x})+b$, ϑ 为正则化常数, H 为特征空间. 系数向量 $\boldsymbol{w}$ 的最小二乘估计为

$$\widehat{\boldsymbol{w}}=(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}+\vartheta\boldsymbol{E}_l)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{y}. \tag{13}$$

其中: $\boldsymbol{E}_l\in\mathcal{R}^l$ 为单位阵, $\boldsymbol{y}\in\mathcal{R}^l$ 为目标输出向量.

对核矩阵中心化预处理可消除偏置项 b. 利用表示定理[5], 由线性 RR 方法的对偶形式可得到 KRR 模型如下:

$$f(\boldsymbol{x})=\sum_{i=1}^{l}\alpha_i\boldsymbol{K}(\boldsymbol{x}_i,\boldsymbol{x})=\langle\boldsymbol{\alpha},\boldsymbol{k}(\boldsymbol{x})\rangle, \tag{14}$$

其中$\boldsymbol{\alpha}=[\alpha_1,\alpha_2,\cdots,\alpha_l]^{\mathrm{T}}$. 将式(14)代入(12), 则$\boldsymbol{\alpha}$的最优解为

$$\boldsymbol{\alpha}=(\boldsymbol{K}+l\vartheta\boldsymbol{E}_l)^{-1}\boldsymbol{y}. \tag{15}$$

当l较大时, 求取式(15)中核矩阵的逆会带来计算和内存需求的困难.

为克服KRR的缺点, GKRR方法利用贪心逼近算法, 首先选择一个数目为m的正交基向量构成的数据子集, 在特征空间中, 将所有训练数据投影到所选择的数据子集上完成核投影逼近, 则有

$$\boldsymbol{Z}=[z_1,z_2,\cdots,z_l]\in\mathcal{R}^{m\times l},$$

与式(12)相似. 在正则化风险最小化的条件下, 式(14)的KRR模型变换为GKRR模型, 即

$$f(\boldsymbol{x})=\langle\boldsymbol{\nu},\boldsymbol{z}\rangle=\langle\boldsymbol{\nu},\boldsymbol{A}^{\mathrm{T}}\boldsymbol{k}_s(\boldsymbol{x})\rangle. \tag{16}$$

其中: 核函数$\boldsymbol{k}_s(\boldsymbol{x})=[k(\boldsymbol{x}_{j_1},\boldsymbol{x}),\cdots,k(\boldsymbol{x}_{j_m},\boldsymbol{x})]^{\mathrm{T}}$, 系数向量$\boldsymbol{\nu}\in\mathcal{R}^m$. 与式(15)的求解相同, $\boldsymbol{\nu}$的最优解为

$$\boldsymbol{\nu}^*=\left(\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}+m\vartheta\boldsymbol{E}_m\right)^{-1}\boldsymbol{Z}\boldsymbol{y}, \tag{17}$$

其中$\boldsymbol{E}_m$为单位阵.

由式(17)可见, GKRR方法仅针对维数为$m\times m$的矩阵进行求逆运算, 由于$m\ll l$, 可适用于大数据集的求解, 且满足复杂运算需求.

3 中期电力负荷峰值预测实验

将贪心核特征提取建模方法应用于不同地区的电力负荷预测的实例中. 应用贪心核特征提取方法时, 贪心核逼近算法的停止条件为满足基向量m的规定数目或$\varepsilon_{\max}=\max\varepsilon_j(x)=0.05$最大重构误差. 另外, 高斯径向基核函数与多项式核等其他核函数相比, 具有良好的性能. 因此, 相同条件下, 用于对比实验的核函数仍选取高斯核函数, 即

$$k(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp(-0.5\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2/\sigma^2).$$

由于负荷预测的复杂性, 不仅考虑电力负荷历史观测值, 还需要考虑日历信息、节假日信息、气象信息等因素. 按照时间序列建模的方式, 预测输入包含历史负荷值$(y_{t-1},y_{t-2},\cdots,y_{t-\Delta})$, Δ表示嵌入维数, 还包括日历、节假日信息. 为了避免建模过程中出现的计算饱和现象, 电力负荷值的实际序列数据归一化为零均值、单位方差的数值. 按照多步迭代预测的方法进行中期峰值负荷预测, 即训练时采用GKPCR或GKRR方法构建单步预测模型. 测试时, 递推地将输出作为下一次预测输入, 滚动进行直至预测结束.

用平均绝对百分比误差(MAPE)、最大误差(ME)和相对误差(RE)评价预测性能, 即

$$\mathrm{MAPE}=100\sum_{i-1}^{n}\left|\frac{y_i-\widehat{y_i}}{y_i}\right|\Big/n,$$

$$\mathrm{ME}=\max|y_i-\widehat{y_i}|,\ \mathrm{RE}=(\widehat{y_i}-y_i)/y_i.$$

其中: y_i为预测日的实际负荷值, $\widehat{y_i}$为模型预测值, n为预测的天数.

3.1 EUNITE竞赛的电力峰值负荷预报实验

实验选取欧洲EUNITE网络组织的中期负荷预测竞赛提供的电力负荷实际数据集[11], 目标是预测未来31天内, 即1999年1月的每日最大电力负荷值, 称为每日峰值负荷. 为了弥补气象因素的影响, 以数据分割的方式选取1997年1月～3月、1997年10月～1998年3月、1998年10月～12月冬季时间段的历史负荷数据作为训练数据集. 电力负荷值的实际时间序列数据的嵌入维取为7, 结合日历信息和节假日信息, 日历信息用七位二进制编码表示, 节假日信息用一位二进制编码表示, 模型的输入为15维, 这与文献[4]一致. 文献[4]的结果是其作者用自己开发的LIBSVM软件获得, 且赢得当年EUNITE竞赛第一名.

实验中, 通过交叉验证的方法选取高斯核函数的超参数$\sigma=2$, m固定为全部训练数据集数目的20% ($m=0.2l$). 在GKPCR方法中, 非线性主元为25, GKRR方法中, 正则化参数$\vartheta=1\mathrm{e}-2$时, 可获取较优的实验结果. 为了衡量本文方法的预测效果, 在相同条件下, 将GKPCR、GKRR、GKPCA结合具有线性核函数的SVM与其他方法构建的预测模型进行比较. 不同方法的预测结果评价由表1列出.

表1 GKPCR、GKRR与其他方法的预测结果对比

预测方法	MAPE	ME/MW
SVM ($C=4\,096$, $\varepsilon=0.5$)	1.937 9	48.231 4
PCA+ε-SVM ($C=100$, $\varepsilon=0.5$)	2.028 6	39.670 8
KPCA+ε-SVM ($C=4\,096$, $\varepsilon=0.5$)	1.776 5	49.710 7
GKPCA+ε-SVM ($C=4\,096$, $\varepsilon=0.5$)	1.658 7	41.638 7
KPCR	1.698 6	41.969 2
EM-KPCR	1.698 2	41.974 8
KPLS	1.843 7	48.629 8
GKPCR	1.597 7	38.711 7
正则化的RBF网络	1.951 9	47.640 1
KRR	1.738 8	46.578 4
GKRR	1.570 5	39.919 0

在PCA结合具有高斯核函数的标准SVM方法中, 主元选为10. 在KPCA结合具有线性核函数的SVM、KPCR、EM-KPCR方法中, 非线性主元均取为25. KRR的正则化参数$\vartheta=1\mathrm{e}\text{-}2$. 在KPLS方法中, 提取主元是对输入和输出同时进行的, 其主元数目取为5. PCA、KPCA、GKPCA结合SVM构建的多层SVM中, 标准ε-SVM算法的实现使用LIBSVM软件完成. 正则化RBF网络方法中, RBF的聚类中心取为9, 正则化参数为1e-4, 使用共轭梯度算法优化20次作为算法的停止条件, 它是文献[12]的延伸. 由表1

可见, GKPCR 与 GKRR 方法的 MAPE 值最低, 与其他方法相比, 贪心核特征提取建模方法的预测精度显著提高.

图 1 给出了待预测日的实际电力负荷峰值和 $m = 0.2l$ 时, GKPCR、GKRR、GKPCA 结合 SVM 与其他方法(包括文献[4]的 SVM 预测输出)峰值预测结果的比较. 由图 1 可见, 所提出的贪心核特征提取建模方法具有明显的预测效果.

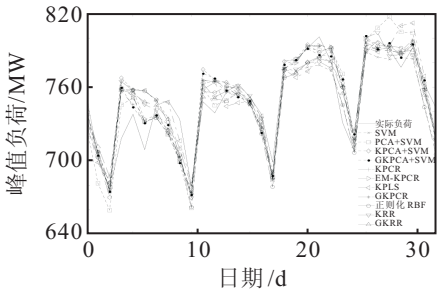

图 1　日峰值负荷预测结果的比较

不同预测方法对未来 31 天的滚动预测相对误差(RE) 如图 2 所示. 由图 2 可见, 基于贪心核特征提取建模的电力负荷中期预测方法相对误差波动较小, 能够取得很好的预测效果.

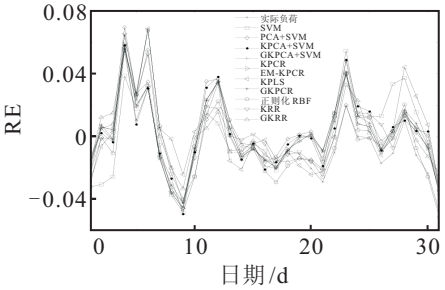

图 2　日峰值负荷预测相对误差的结果比较

3.2　新英格地区中期电力峰值负荷实验

实验数据取自美国新英格兰地区 2004～2005 年期间每隔 1 小时的负荷数据集. 目标是预测未来 21 天内, 即 2006 年 5 月 1～21 日期间的每日电力负荷峰值. 采用数据分割方法选取 2004 年 4～5 月、2005 年 4～5 月和 2006 年 4 月数据为训练数据. 预测模型的输入包括过去一周的历史负荷峰值和节假日、日历信息, 仍为 15 维, 以多步迭代预测方式进行. 实验中, 由交叉验证法取核参数 $\sigma = 5$, 基向量 m 固定为训练数据的 20% ($m = 0.2l$), GKPCR 中非线性主元为 18. GKRR 中, 正则化参数 $\vartheta = 1\text{e-}3$ 时, 可获取较优的实验结果.

为了评价预测效果, 在相同条件下, 将 GKPCR、GKRR、GKPCA 结合具有线性核函数的 SVM 与其他预测模型进行比较. 不同方法的预测结果评价由表 2 给出. 在 PCA 结合 SVM 方法中, 主元选为 5. 在 KPCA 结合 SVM、KPCR、EM-KPCR 方法中, 非线性主元均取为 18. KRR 的正则化参数 $\vartheta = 1\text{e-}3$. KPLS 方法中的主元数目取为 5, 可获得较好的预测效果. SVM 算法的实现使用 LIBSVM 软件完成, 正则化 RBF 网络中的聚类中心为 9, 正则化参数为 1e-2, 使用共轭梯度算法优化 20 次作为算法的停止条件. 由表 2 可见, GKPCR 与 GKRR 方法的 MAPE 值、ME 值均较低, 与其他方法相比, 贪心核特征提取建模方法的预测精度显著提高.

表 2　GKPCR、GKRR 与其他方法的预测结果对比

预测方法	MAPE	ME/MW
SVM ($C = 100$, $\varepsilon = 0.5$)	1.774 5	778.758 6
PCA +ε-SVM ($C = 100$, $\varepsilon = 0.5$)	1.354 2	618.357 7
KPCA +ε-SVM ($C = 100$, $\varepsilon = 0.5$)	1.318 5	502.961 6
GKPCA +ε-SVM ($C = 100$, $\varepsilon = 0.5$)	1.348 4	518.341 5
KPCR	1.295 3	454.988 8
EM-KPCR	1.314 3	414.036 4
KPLS	1.715 4	592.980 1
GKPCR	1.242 9	411.524 6
正则化的 RBF 网络	1.371 2	485.000 1
KRR	1.281 7	507.962 1
GKRR	1.263 1	422.785 7

图 3 给出了待预测日的实际电力负荷峰值和 $m = 0.2l$ 时, GKPCR、GKRR、GKPCA 结合 SVM 与其他方法(包括文献[4]方法的 SVM 预测输出)峰值预测结果比较. 由图 3 可见, 所提出的贪心核特征提取建模方法具有明显的预测效果.

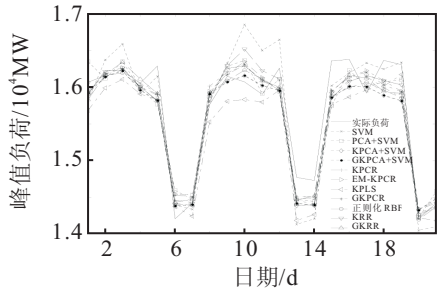

图 3　日峰值负荷预测结果的比较

不同预测方法对未来 21 天的滚动预测相对误差(RE) 如图 4 所示. 由图 4 可见, 基于贪心核特征提取建模的电力负荷中期预测方法相对误差波动较小, 能够取得很好的预测效果.

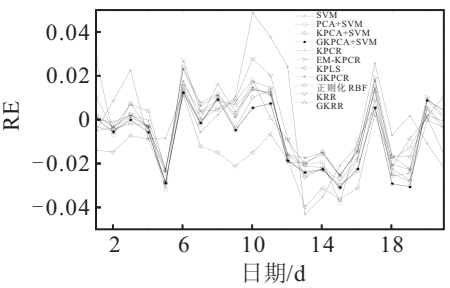

图 4　日峰值负荷预测相对误差结果比较

在实验中, 基向量的数目m固定为全部训练数据集的50% ($m = 0.5l$) 以下时, 贪心核特征提取建模方法均可取得较好的预测效果, 这表明本文方法的鲁棒性也较好. m的取值通常为数据集的25%左右, 且随着m的增大, 算法的训练时间也会加长.

4 结 论

针对实际的电力负荷预测, 提出了基于贪心核特征提取建模的GKPCR和GKRR方法. GKPCR和GKRR方法通过贪心核逼近算法获取新的低维正交基向量构成的数据子集, 以尽可能小的最小均方重构误差逼近训练数据集, 而且使得投影函数的复杂性最小化. 贪心核特征提取建模方法通过控制非线性主元的数目或正则化参数, 抓住了数据的本质特征, 提高了模型的推广性. GKPCR和GKRR方法是KPCR和KRR方法的进一步延伸, 能克服由于数据集增大而引起的核矩阵计算困难和内存不足的缺点, 适应于大数据集的在线学习. 最后通过不同地区的中期电力负荷峰值预测实例, 验证了所提出方法的有效性.

参考文献(References)

[1] 杜杰, 徐立中, 曹一家, 等. 短期负荷预测Volterra滤波器模型[J]. 控制与决策, 2009, 24(12): 1903-1908.
(Du J, Xu L Z, Cao Y J, et al. Short-term load forecasting model based on Volterra filter[J]. Control and Decision, 2009, 24(12): 1903-1908.)

[2] Ghiassia M, Zimbrab D K, Saidane H. Medium term system load forecasting with a dynamic artificial neural network model[J]. Electric Power Systems Research, 2006, 76(5): 302-316.

[3] Elattar E E, Goulermas J Y, Wu Q H. Electric load fore-casting based on locally weighted support vector regression[J]. IEEE Trans on SMC, 2010, 40(4): 438-447.

[4] Chen B J, Chang M W, Lin C J. Load forcasting using support vector machines: A study on eunite competition 2001[J]. IEEE Trans on Power Load Systems, 2004, 19(4): 1821-1830.

[5] Saunders C, Gammerman A, Volk V. Ridge regression algorithm in dual variables[C]. Proc of the 15th Int Conf on Machine Learning. Madison-Wisconsin: Morgan Kaufmann Publishers, 1998: 515-521.

[6] Scholkopf B, Smola A, Muller K. Nonlinear component analysis as a kernel eigenvalue problem[J]. Neural Computation, 1998, 10(5): 1299-1319.

[7] Rosipal R. Kernel partial least squares for nnlinear regression and discrimination[J]. Neural Network World, 2003, 13(3): 291-300.

[8] Rosipal R, Girolami M. An expectation-maximization approach to nonlinear component analysis[J]. Neural Computation, 2001, 13(3): 505-510.

[9] Franc V, Hlavac V. Greedy algorithm for a training set reduction in the kernel methods[C]. Proc of Computer Analysis of Images and Patterns. Berlin: Springer, 2003: 426-433.

[10] Franc V. Optimization algorithms for kernel methods[D]. Prague: Department of Cybernetics, Czech Technical University, 2005.

[11] Sincak P. World-wide competition within the EUNITE network[EB/OL]. (2001-08-05)[2012-11-06]. http://neuron.tuke.sk/competition/index.php.

[12] Moody J, Darken C. Fasting learning in networks of locally-tuned processing units[J]. Neural Computation, 1989, 1(2): 281-294.

(责任编辑: 郑晓蕾)