文章编号:1001 - 9081(2010)01 - 0175 - 03

模糊核聚类支持向量机集成模型及应用

张　娜[1,2],张永平[1]
(1. 中国矿业大学 计算机学院,江苏 徐州 221116; 2. 宿迁高等师范学校 计算机系,江苏 宿迁 223800)
(zhangna8083031@126.com)

摘　要:为了进一步提高支持向量机在回归预测中的精度,提出一种基于模糊核聚类的最小二乘支持向量机集成方法。该方法采用模糊核聚类算法根据相互独立训练出的多个 LS-SVM 在验证集上的输出对其进行分类,并计算每一类中的所有个体在独立验证集上的泛化误差,然后取其中平均泛化误差最小的个体作为这一类的代表,最后经简单平均法得到集成的最终预测输出。在短期电力负荷预测中的实验结果表明,该方法具有更高的精确度。

关键词:最小二乘支持向量机;模糊核聚类;集成学习;短期负荷预测
中图分类号: TP273　　**文献标志码**:A

Support vector machine ensemble model based on KFCM and its application

ZHANG Na[1,2], ZHANG Yong-ping[1]
(1. *College of Computer, China University of Mining and Technology, Xuzhou Jiangsu* 221116, *China*;
2. *College of Computer, Suqian Higher Normal School, Suqian Jiangsu* 223800, *China*)

Abstract: To further enhance the regression prediction accuracy of support vector machine, a Least Squares Support Vector Machine (LS-SVM) ensemble model based on Kernel Fuzzy C-Means clustering (KFCM) was proposed. The KFCM algorithm was used to classify LS-SVMs trained independently by its output on validate samples, the generalization errors of LS-SVMs in each category to the validate set were calculated of the LS-SVM whose error was minimum would be selected as the representative of its category, and then the final prediction was obtained by simple average of the predictions of the component LS-SVM. The experiments in short-term load forecasting show the proposed approach has higher accuracy.

Key words: Least Squares Support Vector Machine (LS-SVM); Kernel Fuzzy C-Means clustering (KFCM); ensemble learning; short-term load forecasting

0　引言

支持向量机(Support Vector Machine, SVM)[1]以统计学习理论为基础,采用结构风险最小化准则和 VC 维理论,根据有限数据信息在模型的复杂度和学习能力之间寻找最佳折中,从而获得最好的推广能力,能够较好地解决小样本、非线性、高维数和局部极小点等问题。最小二乘支持向量机[2](Least Squares Support Vector Machine, LS-SVM)对传统支持向量机进行改进,提高了求解问题的速度和收敛精度。但在实际应用中,最小二乘支持向量机的训练和泛化性能受到正则化参数 γ 和核宽度 σ 取值的影响。集成学习可以通过简单地训练多个学习机(如神经网络、支持向量机等)并将其结果进行结合,从而有效地提高学习系统的泛化能力[3]。现有理论研究表明[4-5],对集成的个体进行必要的选择能够有效地降低神经网络集成的泛化误差。为了更好地解决 SVM 的模型选择问题,本文提出了一种模糊核聚类最小二乘支持向量机集成模型,并将其应用于短期电力负荷预测中。实验结果表明,该方法可以有效地提高短期电力负荷预测的精确度,具有更好的泛化性能。

1　最小二乘支持向量机的原理

LS-SVM[2,6]是对 SVM 的一种改进,它将传统 SVM 中的不等式约束改为等式约束,且将误差平方和损失函数作为训练集的经验损失,这样就把解二次规划问题转化为求解线性方程组问题,从而提高了求解问题的速度和收敛精度。

设给定训练样本集 $\{x_k, y_k\}(k = 1, \cdots, N)$, $x_k \subset \mathbf{R}^n$, $y_k \subset \mathbf{R}$,利用非线性映射 $\varphi(\cdot)$ 将输入空间映射为高维特征空间,再进行最优线性回归,对未知函数进行回归估计可表达为:

$$y(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\varphi(\boldsymbol{x}) + b \tag{1}$$

式中权向量 $\boldsymbol{w} \subset \mathbf{R}^n$,偏置量 $b \subset \mathbf{R}$。这样构造的函数 $y(x)$ 可使得对于样本集之外的输入 $\boldsymbol{x}$,也能精确地估计出相应的输出 y。LS-SVM 定义优化问题为:

$$\min_{\boldsymbol{w},b,\boldsymbol{e}} J(\boldsymbol{w},\boldsymbol{e}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}/2 + \gamma\sum_{k=1}^{N} e_k^2/2;\ \gamma > 0 \tag{2}$$

$$\text{s.t. } y_k = \boldsymbol{w}^{\mathrm{T}}\varphi(x_k) + b + e_k, k = 1, 2, \cdots, N$$

式中:优化目标函数 J 的第 1、2 项分别控制模型的复杂度和误差的范围;γ 为正则化参数(处罚因子);e_k 为不敏感损失函数的松弛因子。

引入 Lagrange 函数 L 求解式(2)的优化问题,即:

$$L(\boldsymbol{w},b,\boldsymbol{e},\boldsymbol{a}) = J(\boldsymbol{w},\boldsymbol{e}) - \sum_{k=1}^{N} a_k[\boldsymbol{w}^{\mathrm{T}}\varphi(\boldsymbol{x}_k) + b + e_k - y_k] \tag{3}$$

式中 a_k 为 Lagrange 乘子。根据 KKT 最优条件,可得到此优化问题的解析解为:

收稿日期:2009 - 07 - 07;**修回日期**:2009 - 09 - 11。

作者简介:张娜(1982 -),女,江苏宿迁人,讲师,硕士研究生,主要研究方向:信息处理、机器学习; 张永平(1958 -),男,辽宁丹东人,副教授,主要研究方向:计算机网络与信息安全、密码学。

$$\begin{bmatrix} b \\ \boldsymbol{a} \end{bmatrix} = \begin{bmatrix} 0 & \vec{1}^{\mathrm{T}} \\ \vec{1} & \boldsymbol{\Omega} + \gamma^{-1}\boldsymbol{I} \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ y \end{bmatrix} \tag{4}$$

其中,$\boldsymbol{y} = [y_1, y_2, \cdots, y_N]^{\mathrm{T}}$;$\vec{1} = [1, \cdots, 1]^{\mathrm{T}}$;$\boldsymbol{a} = [a_1, a_2, \cdots, a_N]^{\mathrm{T}}$;$\boldsymbol{I}$ 为 $N \times N$ 的单位矩阵;$\boldsymbol{\Omega}$ 为方阵,其第 k 列 l 行的元素为 $\Omega_{kl} = \varphi(\boldsymbol{x}_k)^{\mathrm{T}}\varphi(\boldsymbol{x}_l) = K(\boldsymbol{x}_k, \boldsymbol{x}_l)$,$K(\cdot, \cdot)$ 为核函数,它为满足 Mercer 条件的任意对称函数。这样,不需要知道非线性变换的具体形式,就可用核函数来实现算法的线性化。本文采用高斯径向基函数(Radical Basis Function,RBF)为核函数,其表达式为:

$$K(\boldsymbol{x}_k, \boldsymbol{x}_l) = \exp(-\| \boldsymbol{x}_k - \boldsymbol{x}_l \|^2/(2\sigma^2)) \tag{5}$$

式中 σ 为核宽度,反映了边界封闭包含的半径。

式(4)的线性系统可用最小二乘算法求解出 b 和 $\boldsymbol{a}$,再由式(3)进一步求出 $\boldsymbol{w}$,从而得到训练数据集的非线性逼近为:

$$y(\boldsymbol{x}) = \sum_{k=1}^{N} a_k K(\boldsymbol{x}, \boldsymbol{x}_k) + b \tag{6}$$

从式(6)可知,SVM 回归可用 3 层的网络结构来表示,其中输入层、隐层和输出层的节点数分别为 n、N 和 1,而输入层与隐层之间、隐层与输出层之间的连接权值分别为 1 和 a_k。

在实际应用中,最小二乘支持向量机的训练和泛化性能受到正则化参数 γ 和核宽度 σ 取值的影响,迫切需要一种切实可行的方法来进一步提高 LS-SVM 预测的精确度和稳定性。

2 基于 KFCM 的 LS-SVM 集成

2.1 模糊核聚类算法(KFCM)

KFCM 算法的基本思想是利用非线性映射 $\Phi(*)$ 把输入模式向量空间变换到一个高维特征空间,然后在该特征空间采用模糊 c-均值算法,对变换后的特征向量 $\Phi(\boldsymbol{x}_i)$ 进行模糊聚类分析[7-8]。它能够突出不同类别样本特征的差异,使得原来线性不可分的样本点在核空间中变得线性可分,从而实现更为准确的聚类。

假设输入空间的样本 $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$,$\boldsymbol{X} \subseteq \mathbf{R}^p$,通过一个非线性映射 $\Phi: \chi \to \boldsymbol{F}$ 将输入空间 χ 变换至高维特征空间 $\boldsymbol{F}$,c 为预定的类别数目,$v_i (i = 1, 2, \cdots, c)$ 为第 i 个聚类的中心,$u_{ik} (i = 1, 2, \cdots, c;\ k = 1, 2, \cdots, n)$ 是第 k 个样本对第 i 类的隶属度函数,且 $0 \leqslant u_{ik} \leqslant 1$ 及 $0 < \sum_{k=1}^{n} u_{ik} < n$,则模糊核聚类算法的目标函数为:

$$J_m(\boldsymbol{U}, \boldsymbol{v}) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^m \| \Phi(\boldsymbol{x}_k) - \Phi(\boldsymbol{v}_i) \|^2 \tag{7}$$

式中,$\boldsymbol{U} = \{u_{ik}\}$,$\boldsymbol{v} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_c\}$,$m > 1$ 为加权指数,其约束为:

$$\sum_{i=1}^{c} u_{ik} = 1;\ \forall k = 1, 2, \cdots, n \tag{8}$$

定义核函数 $K(\boldsymbol{x}, \boldsymbol{y})$,满足 $K(\boldsymbol{x}, \boldsymbol{y}) = \Phi(\boldsymbol{x})^{\mathrm{T}}\Phi(\boldsymbol{y})$,KFCM 聚类的准则是求目标函数的极小值。根据 Lagrange 乘子寻优法,式(7)所示目标函数的最小值可由式(9)、(10)求得:

$$u_{ik} = \frac{(1/(K(\boldsymbol{x}_k, \boldsymbol{x}_k) + K(\boldsymbol{v}_i, \boldsymbol{v}_i) - 2K(\boldsymbol{x}_k, \boldsymbol{v}_i)))^{\frac{1}{m-1}}}{\sum_{j=1}^{c}(1/(K(\boldsymbol{x}_k, \boldsymbol{x}_k) + K(\boldsymbol{v}_j, \boldsymbol{v}_j) - 2K(\boldsymbol{x}_k, \boldsymbol{v}_j)))^{\frac{1}{m-1}}} \tag{9}$$

$$\boldsymbol{v}_i = \frac{\sum_{k=1}^{n} u_{ik}^m K(\boldsymbol{x}_k, \boldsymbol{v}_i)\boldsymbol{x}_k}{\sum_{k=1}^{n} u_{ik}^m K(\boldsymbol{x}_k, \boldsymbol{v}_i)} \tag{10}$$

综上所述,KFCM 算法的步骤如下:

步骤 1 设定聚类数目 c 和参数 m;

步骤 2 初始化各个聚类中心 $\boldsymbol{v}_i$;

步骤 3 重复下面的运算,直到各个样本的隶属度值稳定:1)用当前的聚类中心根据式(9)更新隶属度;2)用当前的聚类中心和隶属度根据式(10)更新各个聚类中心。本文中选择高斯径向基核函数。

2.2 LS-SVM 模糊核聚类集成

选择性集成的方法能够取得比用全部个体集成更好的泛化性能[4],本文采用模糊核聚类集成[9]的方法来解决集成中个体 LS-SVM 的选取问题。

首先用 KFCM 算法对相互独立训练出的 m 个 LS-SVM 个体进行分类,然后计算所有类别中每个个体 LS-SVM 在独立验证集上的泛化误差,最后分别选取每个类别中平均泛化误差最小的 LS-SVM 个体作为这一类的代表进行简单平均法集成。具体实现步骤如下:

步骤 1 将 m 个 LS-SVM 个体对每个独立验证样本的输出按照相同的顺序以列向量的形式存放到输出矩阵 $\boldsymbol{O}$ 中;

步骤 2 选择聚类数目 c 并设定加权指数 m,对 LS-SVM 的输出矩阵 $\boldsymbol{O}$ 进行模糊核聚类分析并可以得到隶属度矩阵 $\boldsymbol{U}$;

步骤 3 根据隶属度矩阵 $\boldsymbol{U}$,可得每个 LS-SVM 对所有 c 个类别的隶属度的最大值,然后将相应的 LS-SVM 归入具有隶属度最大值的类别;

步骤 4 计算每一类中的所有 LS-SVM 个体在验证集上的平均泛化误差,将每一类中平均泛化误差最小的 LS-SVM 个体选择出来;

步骤 5 给定阈值 λ,在每个类别中的最佳个体 LS-SVM 中选择平均泛化误差小于 λ 的个体构成最终的集成个体;

步骤 6 将最终选择出的个体 LS-SVM 对测试样本的输出经简单平均法得到集成的最终输出。

该方法既能够保证最终集成中的个体 LS-SVM 具有较高的精确度,而且也同时保证了个体之间具有较大的差异度,相关的理论研究表明[4,9],这种集成方法可以进一步地提高集成学习的泛化能力。

3 短期负荷预测实例分析

短期负荷预测对制定发电调度计划、确定燃料供应计划及合理安排机组检修计划等均有重要指导作用,其预测的精确性极大地影响着供电部门的经济效益。本文采用我国南方某电网 2006 年 7 月 5 日到 8 月 10 日的整点有功负荷值,在负荷预测日的前一天中,每隔 2 小时对电力负荷进行一次测量,这样一来,一天共测得 12 组负荷数据。由于负荷值曲线相邻的点之间不会发生突变,因此后一时刻的值必然和前一时刻的值有关,除非出现重大事故等特殊情况。所以这里将前一天的实时负荷数据作为 LS-SVM 的样本数据。此外,由于电力负荷还与环境因素有关,因此,还需要通过天气预报等手段获得预测日的最高气温、最低气温和天气特征值,其中 0 表示晴天,0.5 表示阴天和 1 表示雨天。这里将电力负荷预测日当天的气象特征数据也作为 LS-SVM 的输入变量。因此,输入变量就是一个 15 维的向量。目标向量就是预测日当天的 12 个负荷值,这样输出变量就是一个 12 维的向量。获

得输入和输出变量后,要对其进行归一化处理,将数据处理为区间为[0, 1]的数据,归一化方法有很多种形式,本文采取如下方法:

$$\hat{x} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{11}$$

LS-SVM 中的正则化参数 γ 分别取 100、300、500、700 和 900,而核宽度 σ 分别取 0.25、0.5、0.75 和 1,这样组合起来就有 20 个参数不同的 LS-SVM 进行学习。然后用 7 月 5 日到 8 月 3 日共 30 天的负荷及天气数据对这 20 个 LS-SVM 进行训练;再用训练好的 LS-SVM 对 8 月 4 日到 8 月 9 日这 6 天的负荷数据进行验证,并根据输出的预测结果对 LS-SVM 进行聚类分析,在 KFCM 算法中,聚类数目 c 取 5,参数 m 取为 2,算法停止的条件为相邻迭代步数间的隶属度值的差的绝对值小于 0.001,初始聚类中心取为 0 到 1 间的随机数,其中核函数选用高斯核函数,并取 σ = 32;最后,将最佳的 5 个 LS-SVM 对 8 月 10 日的预测输出经简单平均法得到集成的最终输出。运行结果如图 1 所示。可以看出,本文方法取得了很好的预测效果,最终的预测均方误差为 7.651 7e - 004,而 KFCM 算法中所选择的 5 个最佳 LS-SVM 的预测均方误差分别为 7.690 6e - 004、7.717 3e - 004、7.689 5e - 004、7.705 1e - 004 和7.662 2e - 004。图 2 所示为预测误差曲线。

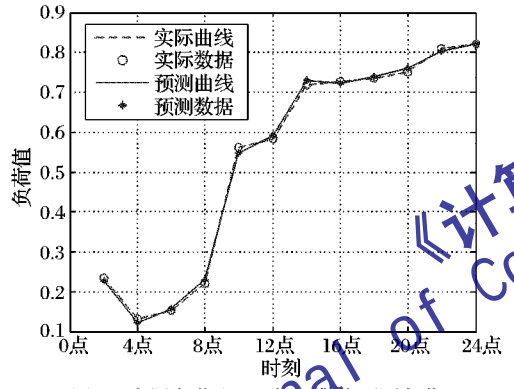

图 1　实际负荷和 LS-SVM 集成预测负荷

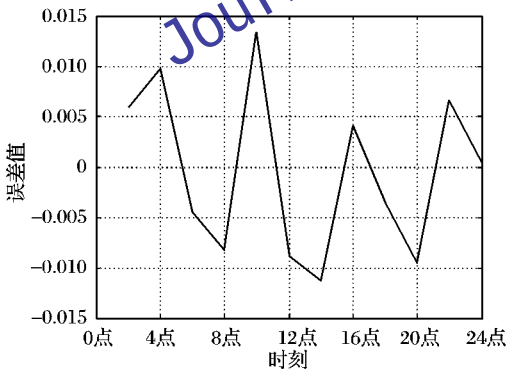

图 2　LS-SVM 集成预测误差曲线

将 LS-SVM 的参数采用随机选取的方法,即正则化参数 γ 的取值范围为[0, 1 000],核宽度 σ 的取值范围为[0, 1],在上述范围内随机选取 20 个 LS-SVM 进行训练,仍然采用本文方法进行预测,多组实验结果表明,其最终的预测均方误差也不超过 7.85e - 004。表 1 所示为几种方法进行比较的结果。

表 1　三种方法平均预测误差的比较

预测方法	平均预测误差
单个 LS-SVM	8.0513 e - 004
LS-SVM 简单平均集成	7.9472 e - 004
LS-SVM 模糊核聚类集成	7.6885 e - 004

其中,LS-SVM 模糊核聚类集成的预测误差并不一定就比单个最佳的 LS-SVM 预测误差要小,但是,事先是无法知道哪个 LS-SVM 的预测误差是最小的,即使训练误差较小的 LS-SVM 也不一定在测试中就能取得很好的效果,而多组实验结果表明,本文方法可以进一步提高负荷预测的精确度。

4　结语

本文根据选择性集成学习的思想,利用模糊核聚类的方法对多个独立训练出的 LS-SVM 进行聚类分析,并从每个类中选择最佳的个体 LS-SVM 进行集成。对短期负荷预测的实验结果表明,本文方法能够有效地提高负荷预测的精确度。

参考文献:

[1] VAPNIK V N. 统计学习理论的本质[M]. 张学工, 译. 北京: 清华大学出版社, 2000.

[2] SUYKENS J A K, VANDEWALLE J. Recurrent least squares support vector machines[J]. IEEE Transactions on Circuits and Systems, 2000, 47(7): 1109 - 1114.

[3] HANSEN L K, SALAMON P. Neural network ensembles[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1990, 12(10): 993 - 1001.

[4] ZHOU ZHIHUA, WU JIANXIN, TANG WEI. Ensembling neural networks: Many could be better than all[J]. Artificial Intelligence, 2002, 137(1/2): 239 - 263.

[5] 蔡俊伟, 胡寿松, 陶洪峰. 基于选择性支持向量机集成的混沌时间序列预测[J]. 物理学报, 2007, 56(12): 6820 - 6828.

[6] 程启明, 杜许峰, 郭瑞青, 等. 基于最小二乘支持向量机的多变量逆系统控制方法及应用[J]. 中国电机工程学报, 2008, 28(35): 96 - 101.

[7] 张道强. 基于核的联想记忆及聚类算法的研究与应用[D]. 南京: 南京航空航天大学, 2004.

[8] 普运伟, 金炜东, 朱明, 等. 核模糊 C 均值算法的聚类有效性研究[J]. 计算机科学, 2007, 34(2): 207 - 210.

[9] 李凯, 黄厚宽. 一种基于聚类技术的选择性神经网络集成方法[J]. 计算机研究与发展, 2005, 42(4): 594 - 598.

(上接第 139 页)

[4] RFC4988. Mobile IPv4 Fast Handovers [S], 2007.

[5] RFC4068. Fast Handovers for Mobile IPv6 [S], 2005.

[6] 唐宏, 陶京涛, 王柏丁, 等. 基于移动通信切换特性的 L2-Trigger 方法[J]. 计算机应用, 2006, 26(12): 2796 - 2799.

[7] 唐宏, 陈前斌, 吴中福. 移动 IP 技术中 L2-TRIGGER 方法研究[J]. 重庆邮电学院学报, 2003, 15(4): 88 - 91.

[8] AKYILDIZ I, WANG W. The predictive user mobilty profile framework for wireless multimedia networks[J]. IEEE/ACM Transactions on Networking, 2004, 12(6): 1021 - 1035.

[9] CHOI Y H, PARK J, CHUNG Y U, *et al*. Cross-layer handover optimization using linear regression model[C]// IEEE International Conference on Information Networking. New York: IEEE, 2008: 23 - 25.

[10] 邓聚龙. 灰色系统基本方法 [M]. 2 版. 武汉: 华中科技大学出版社, 2005.

[11] Mobile WIMAX-Part 1: A technical overview and performance evaluation [EB/OL]. [2009 - 03 - 15]. http://www.wimaxforum.org/technology/downloads/Mobile_WiMAX_Part1_Overview_and_Performance.pdf.