DOI:10.13334/j.0258-8013.pcsee.2006.18.002

第 26 卷 第 18 期 中 国 电 机 工 程 学 报 Vol.26 No.18 Sep. 2006
2006 年 9 月 Proceedings of the CSEE ©2006 Chin.Soc.for Elec.Eng.

文章编号：0258-8013 (2006) 18-0006-07 中图分类号：TM74 文献标识码：A 学科分类号：470·40

基于数据挖掘的 SVM 短期负荷预测方法研究

牛东晓，谷志红，邢 棉，王会青

(华北电力大学工商管理学院，河北省 保定市 071003)

Study on Forecasting Approach to Short-term Load of SVM Based on Data Mining

NIU Dong-xiao, GU Zhi-hong, XING Mian, WANG Hui-qing

(School of Business Administration, North China Electric Power University, Baoding 071003,Hebei Province,China)

ABSTRACT：The support vector machine (SVM) has been successfully applied to the load forecasting area, but it has some disadvantages of very large data amount and slow processing speed. Using advantages of the data mining technology in processing large data and eliminating redundant information, a SVM forecasting system based on data mining preprocess was proposed to search the historical daily load with the same meteorological category as the forecasting day and to compose data sequence with highly similar meteorological features. Taking the new data sequence as the training data of SVM, the data amount was decreased and the processing speed was improved. This approach has achieved greater forecasting accuracy comparing with the method of single SVM and BP neural network.

KEY WORDS： power system；data mining；meteorological factor；support vector machines；short-term load forecasting

摘要：支持向量机方法已成功地应用在负荷预测领域，但它在训练数据时存在数据处理量太大、处理速度慢等缺点。为此提出了一种基于数据挖掘预处理的支持向量机预测系统，引用在处理大数据量、消除冗余信息等方面具有独特优势的数据挖掘技术,寻找与预测日同等气象类型的多个历史短期负荷，由此组成具有高度相似气象特征的数据序列，将此数据序列作为支持向量机的训练数据，可减少数据量，从而提高预测的速度和精度，克服支持向量机的上述缺点。将该系统应用于短期负荷预测中,与单纯的 SVM 方法和 BP 神经网络法相比,得到了较高的预测精度。

关键词：电力系统；数据挖掘；气象因素；支持向量机；短期负荷预测

0 引言

在 96 点短期负荷预测工作中，影响预测精度提高的最大困难就是短期负荷受到多种随机干扰因素的影响和需要多种不同的预测模型拟合。其中，对短期负荷曲线变化影响最大的干扰因素是气象因素，如果不考虑气象因素，无论采用何种技术建立预测模型，均会导致预测失败，且误差较大，影响电网的稳定运行。目前，其他研究人员一般情况下采用人工神经网络模型、遗传算法模型和小波分析模型[1-4]进行短期负荷预测，其中人工神经网络模型应用最为广泛。但普通神经网络模型[5-6]大多不考虑气象因素影响，并且出现了一些新的问题。例如，BP网络学习算法实际上是利用梯度下降法调节权值使目标函数达到极小，而目标函数仅为各给定输入和相应输出差的平方和，导致了BP网络过分强调克服学习错误而泛化性能不强；隐单元的数目难以确定，网络的最终权值受初始值影响大等。

最近，由贝尔实验室的 Vapnik 等提出了一种被称为支持向量机(Support Vector Machines, SVM)的机器学习算法[7-8]，它与传统的神经网络学习方法不同，实现了结构风险最小化原理(SRM)，它同时最小化经验风险，对未来样本有较好的泛化性能。SVM 的另一优点是它的训练等价于解决一个线性约束的二次规划问题，存在唯一解。目前 SVM 已经扩展为解决非线性回归估计问题，而且与神经网络方法相比，有着显著的优越性。但是，经典的 SVM 算法在处理大数据量的模式分类和时间序列预测等方面存在速度慢、时间长的缺点。

基于以上分析，在此提出了一种提高负荷预测精度的新观点，认为提高负荷预测精度的关键在于历史数据的预处理方式和预测模型的改进，并提出了基于数据挖掘的支持向量机短期负荷预测新方

基金项目：国家自然科学基金项目(50077007)；河北省自然科学基金项目（G2005000584）。

Project Supported by National Natural Science Foundation of China (50077007).

法：先通过天气预报了解预测日的整日气象特征，再利用数据挖掘技术寻找与预测日同等气象类型的多个历史短期负荷，组成具有高度相似气象特征的数据序列，从而减少 SVM 的训练数据；据此再构建支持向量机预测模型。将该系统应用于某地区短期负荷预测中，与 BP 神经网络及标准 SVM 方法相比,得到了较高的预测精度，计算速度也得以提高,从而表明了以数据挖掘技术作为信息预处理的 SVM 学习系统的优越性。

1　数据挖掘与 SVM 简介

1.1　数据挖掘

数据挖掘就是使用模式识别技术、统计和数学技术，在大量的数据中发现有意义的新关系、模式和趋势的过程，也就是从海量数据中挖掘出可能有潜在价值的信息技术[9-10]。它可实现以下功能：

（1）分类。按分析对象的属性、特征，建立不同的组类来描述事物，用于预测事件所属的类别。

（2）聚类。识别出数据分析内在的规则，按照这些规则把对象分成若干类。

（3）关联。发现有联系的事件或记录，由此推断事件间潜在的关联，识别可能重复发生的模式。

（4）预测。分析掌握对象的发展规律，对未来的趋势做出预见。

1.2　SVM 方法

SVM方法是统计学习理论(SLT)的一种成功实现，它建立在SLT的VC(vapnik chervonenkis)理论和结构风险最小化(structural risk minimization, SRM)原理基础上，根据有限样本信息在模型的复杂性(对特定训练样本的学习精度)和学习能力(无错误地识别任意样本的能力)之间寻求最佳折衷，以期获得更好的泛化能力[11-12]。它有如下特点：

（1）实现了SVM原则，它能最小化泛化误差的上界，而不是最小化训练误差，因此具有更好的泛化性能。

（2）与神经网络方法相比，SVM有更少的自由参数。在SVM算法中仅有3个自由参数，而神经网络却有大量的自由参数,需要凭经验主观选择。

（3）神经网络不一定能收敛到全局最优解,容易陷入局部最优解。而在SVM算法中,训练SVM就等价于解一个具有线性约束的二次凸规划问题，因而它的解是唯一的、全局的和最优的。

SVM的缺点是不能确定数据中哪些知识是冗余的，哪些是有用的,哪些作用大，哪些作用小。但由于SVM具有良好的泛化性能，目前已经成功地推广应用到了模式识别、函数逼近、信息融合、时间序列预测等领域。

2　基于数据挖掘的 SVM 预测方法

2.1　数据挖掘预处理

根据数据挖掘的基本概念和功能，本文将模糊分类器和灰色关联技术联合起来，设计了一种具有分类和关联功能的数据挖掘技术，步骤如下：

（1）短期负荷气象影响模糊分类器。

对负荷影响较大且天气预报可以给出的因素有日最高温度、日最低温度、日平均温度和日降雨量，如果条件具备，还可以增加其它因素。以本文实例所用电网为例，为了得到气象因素与负荷变化之间的关系，做散点图如图1。从图1可以看出：电力负荷随着最低温度的变化成非线性规律变化，即最低温度对电力负荷有明显的影响。同样的道理，可以作其他气象因素与历史负荷值散点图。从中可以得出结论：最高气温、最低气温、天气状况、湿度对每日负荷有较明显的影响。

现对这 4 类因素进行模糊赋值分类，对这 4 类气象因素的模糊性描述语言用数值向量表示出来，分类情况用向量（Z_1，Z_2，Z_3，Z_4）表示。Z_1、Z_2、Z_3 分别代表日最高温度、日最低温度和日平均温度，由于各地气温的低、中、高标准不同，并且日最高温度是低、中或高也有不同的标准，因此可按本地区的实际情况确定一定的模糊化标准，将三者分类为低、中、高，分别取值 1、2、3； Z_4代表日降雨量，雨量的小、中、大标准各地也不同，也可按地区特点设定，或采用模糊聚类分析方法确定，可将其分类为无雨、小雨、中雨、大雨，分别取值为 0、1、2、3，则可将每日的历史负荷分类如下：

$$(Z_1,Z_2,Z_3,Z_4)=\begin{cases}Z_1=1,2,3\\ Z_2=1,2,3\\ Z_3=1,2,3\\ Z_4=0,1,2,3\end{cases}$$

每日将历史短期负荷曲线 96 点输入数据库，同时输入每日的气象影响因素的模糊分类标记，例如：昨日最高温度中等，平均温度中等，最低温度高等，有中雨，则昨日的气象模糊分类类别是（2，2，3，2），这样，每日都有一个气象模糊分类标记。

如果根据气象预报知道预测日气象类别是（2，2，3，2），则从历史负荷库中反向抽取具有（2，2，3，2）类别特征的历史负荷日，将具有这一气象特

征的所有负荷日抽出，形成一个气象模糊分类库，这个库中的所有负荷日都具有（2，2，3，2）这一气象特征。

（2）根据灰色关联分析方法选定预测所需负荷日数据。

第（1）步所建的同一类别气象模糊分类库中的负荷日只是挖掘出了粗糙相似的气象特征，为了提高预测精度，再运用灰色关联分析理论，在分类库中进一步抽取与预测日具有高度关联的若干历史负荷日，以这些负荷日作为下一步预测建模的历史数据。这些历史负荷日通过联合数据挖掘技术中的分类和关联，达到了在气象特征上与预测日气象特征的高度一致，用此来建模，无疑将较大地提高负荷预测的精度。

1）灰色关联分析理论。

关联分析是灰色系统理论提出的一种分析系统中各因素关联程度的方法，其基本思想是根据曲线间相似程度来判断关联程度。计算步骤如下：

①构造序列矩阵。在通过模糊分类器对历史负荷进行了初步分类挖掘并形成了与预测日具有相似气象特征的历史数据分类库后，进一步进行关联排序分析。参考序列(又称预测日气象特征)用 $\boldsymbol{T}_0$ 表示，若预测日气象预报为最高气温 40℃，平均气温 30℃，最低气温 20℃，雨量 15mm，则 $\boldsymbol{T}_0$=[$T_0(1)$，$T_0(2)$，$T_0(3)$，$T_0(4)$]=(40，30，20，15)。同理用已得分类库中的每日气象数据组成比较序列，以 $\boldsymbol{T}_1$，$\boldsymbol{T}_2$，…，$\boldsymbol{T}_n$ 表示，这 n+1 个序列构成序列矩阵如下：

$$(\boldsymbol{T}_0,\boldsymbol{T}_1,\boldsymbol{T}_2,\ldots,\boldsymbol{T}_n)=\begin{bmatrix} T_0(1) & T_1(1) & \ldots & T_n(1) \\ | & | & | & | \\ T_0(m) & T_1(m) & \ldots & T_n(m) \end{bmatrix} \tag{1}$$

②无量纲化。为消除量纲，用初值化方法进行数据处理。采用式(2)得无量纲矩阵如式(3)所示：

$$T_i'(k)=T_i(k)/T_i(1), i=0,1,2,\ldots,n;\ k=1,2,\ldots,m \tag{2}$$

$$(\boldsymbol{T}_0',\boldsymbol{T}_1',\boldsymbol{T}_2',\ldots,\boldsymbol{T}_n')=\begin{bmatrix} T_0'(1) & T_1'(1) & \ldots & T_n'(1) \\ | & | & | & | \\ T_0'(m) & T_1'(m) & \ldots & T_n'(m) \end{bmatrix} \tag{3}$$

③计算关联系数。

$$x_{0i}(k)=\frac{\min\limits_i\min\limits_k|x_0(k)-x_i(k)|+r\max\limits_i\max\limits_k|x_0(k)-x_i(k)|}{|x_0(k)-x_i(k)|+r\max\limits_i\max\limits_k|x_0(k)-x_i(k)|} \tag{4}$$

式中，$i=0,1,2,\ldots,n$；　$k=1,2,\ldots,m$，$r\hat{I}[0,1]$，r 为分辨系数，通常取r=0.5，得关联系数矩阵为

$$\begin{bmatrix} x_{01}(1) & \ldots & x_{0n}(1) \\ | & | & | \\ x_{01}(m) & \ldots & x_{0n}(m) \end{bmatrix} \tag{5}$$

④计算关联度。

$$r_{0i}==\frac{1}{m}\sum_{k=1}^{m}x_i(k),\quad i=1,2,\ldots,n \tag{6}$$

2）确定预测建模所需历史负荷序列。

本文以预测日的气象因素指标向量为参考序列 $\boldsymbol{T}_0$，通过模糊分类器所得气象分类库中提供的历史数据中每一日的气象因素指标向量为比较序列 $\boldsymbol{T}_t$，计算 $\boldsymbol{T}_0$ 与 $\boldsymbol{T}_t$ 之间的关联度 r_t。设定一阈值a，取关联度 $r_t\geqslant a$的负荷日，或给定一 n 值，一般 n 取 5~7 即可够建模使用，按关联度从大到小的顺序取前 n 个负荷日，然后将这些负荷日按时间先后排序作为新的历史数据序列，这样就完成了挖掘提取工作。

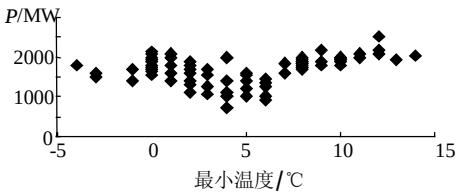

图 1 最小温度与高峰负荷关联散点图

Fig. 1 Diagram of relation scatters points between the minutest temperature and peak-load

2.2 SVM 预测原理

用 SVM 算法估计回归函数时，其基本思想是通过一个非线性映射 f，把输入空间的数据 x 映射到一个高维特征空间中去，然后在这一高维空间中作线性回归[13-14]。

给定一数据点集如下：

$$G=\{(\boldsymbol{x}_i,d_i)\}_{i=1}^{n}$$

式中：$\boldsymbol{x}_i$ 为输入向量；d_i 为望值；n 为数据点的总数。

SVM 采用下式来估计函数：

$$y=f(x)=wf(x)+b \tag{7}$$

式中：$f(x)$为从输入空间到高维特征空间的非线性映射；系数w和 b 通过最小化下式来估计。

$$R_{\mathrm{SVM}}(c)=c\frac{1}{n}\sum_{i=1}^{n}L_e\left[d_i,wf(\boldsymbol{x}_i)+b\right]+\frac{1}{2}\|w\|^2 \tag{8}$$

式(8)中，采用 Vapnik 的e不敏感损失函数为

$$L_e(d,y)=\begin{cases}0, & |d-y|\leq e\\ |d-y|-e, & \text{其它}\end{cases}$$

其目的是用稀疏数据点来表现由式(7)给出的决策函数。在式(8)给出的正则化风险泛函中，第 1 部分

$c\frac{1}{n}\sum_{i=1}^{n}L_e[d_i,y_i]$是经验风险,它们由$e$不敏感损失函数来度量。第 2 部分$\|w\|^2/2$是正则化部分，$c$是正常数,它决定着经验风险与正则化部分之间的平衡。

为了寻找系数w和 b，需要引入松驰变量x_i和x_i^*,使下式成立：

$$\begin{cases}\text{minimize}\quad R_{SVM}(w,x^{(*)})=\frac{1}{2}\|w\|^2+c\sum_{i=1}^{n}(x_i+x_i^*)\\ \text{s.t.}\quad wf(x_i)+b_i-d_i\leqslant e+x_i^*\quad x_i^*\geqslant 0\\ \qquad d_i-wf(x_i)-b_i\leqslant e+x_i\quad x_i\geqslant 0\end{cases}\tag{9}$$

最后，依靠引入拉格朗日乘子a_i和a_i^*，由式(7)给出的决策函数就变成下面的精确形式：

$$y=f(x,a_i,a_i^*)=\sum_{i=1}^{n}(a_i-a_i^*)k(x,x_i)+b\tag{10}$$

对任何 $i=1,\cdots,n$ 都有等式$a_i\times a_i^*=0$，$a_i\geqslant 0$，$a_i^*\geqslant 0$成立。要使式(9)成立，在引入拉格朗日乘子后,就可以把这一凸优化问题简化为对一个二次优化问题寻找向量w的问题，在这种情况下，要找到所求的向量：

$$w=\sum_{i=1}^{n}(a_i-a_i^*)x_i\tag{11}$$

此时必须找到参数a_i和a_i^*，$i=1,\ldots,n$，使下式成立：maximize

$$R(a_i,a_i^*)=-e\sum_{i=1}^{n}(a_i+a_i^*)+\sum_{i=1}^{n}d_i(a_i^*-a_i)-\frac{1}{2}\sum_{i,j=1}^{n}(a_i^*-a_i)(a_j^*-a_j)K(x_i,x_j)\tag{12}$$

$$\text{s.t.}\quad\sum_{i=1}^{n}a_i=\sum_{i=1}^{n}a_i^*,\quad 0\leqslant a_i\leqslant c,\quad 0\leqslant a_i^*\leqslant c$$

通过在二次优化方法中控制 c 和e 参数，就可以控制(即使在高维空间中)SVM 的泛化能力。根据二次规划中的库恩——塔克条件，在式(12)中系数$(a_i^*-a_i)$只有一部分数目是非零值，它们所对应的数据点就是支撑向量。这些数据点位于决策函数的e边界上或在边界外。在方程(12)中，由于其它数据点的系数$(a_i^*-a_i)$都等于零，从而证实了在所有的数据点中只有支撑向量能够决定决策函数。

一般说来，e 值越大，支撑向量数目就越少，因而解的表达就越稀疏。然而，大的e 值也能降低数据点的逼近精度，从这一意义上讲，e 也是解的表达的稀疏程度与数据点的密度之间的平衡因子。在式(12)中，$K(x_i,x_j)$称为核函数,核函数的值等于 2 个向量 x_i 和 x_j 在其特征空间中的像 $f(x_i)$ 和 $f(x_j)$ 的内积，即：

$$K(x_i,x_j)=f(x_i)\times f(x_j)\tag{13}$$

任何函数只要满足 Mercer 条件都可用作核函数，采用不同的函数作为核函数，可以构造实现输入空间中不同类型的非线性决策面的学习机器。

2.3 基于数据挖掘的 SVM 预测模型

2.3.1 基于数据挖掘的SVM结构

从前面分析的数据挖掘技术和SVM方法各功能、特点中可发现它们存在2个互补性的差别：

（1）SVM处理信息一般不能将输入信息空间维数简化，所以当输入信息空间维数较大时，就会导致SVM训练时间较长，而数据挖掘技术却能够通过发现数据间的关系，既可以去掉数据中的冗余信息，又可以简化输入信息的数据空间维数。

（2）数据挖掘技术在实际应用过程中对噪声比较敏感，因而用无噪声的训练样本学习推理的结果在有噪声的环境中应用效果就不太好，也就是说数据挖掘技术的泛化性能较差，而SVM方法有较好的抑制噪声干扰的能力，且具有良好的泛化能力。

因此，根据它们的互补性把两者结合起来，用数据挖掘技术先对数据进行预处理，使得历史数据序列的变化规律有较大的简化、随机性弱化且含预测日气象特征，以便建立支持向量机预测模型时可以不再重复考虑气象特征的输入，就可使预测结果自然含有预测日的气象特征，使模型训练速度明显加快。再根据数据挖掘预处理后的信息结构来构成SVM的信息预测系统。这种系统的结构如图2所示。

在图2中，假设训练样本为$G=\{(x_i,d_i)\}_{i=1}^{n}$，其中 $x_i\in R^d$，是输入向量；$y_i\in R$ 是期望值；n是数据点的总数。在所有输入向量中,假设支持向量有N个,分别为 SV_1,SV_2,…, SV_N,相应的满足Mercer条件的 $f(x_i)$ 和 $f(x_j)$ 为2个向量 x_i 和 x_j 在其特征空间中的像， $K(x_i,x_j)$ 为核函数。

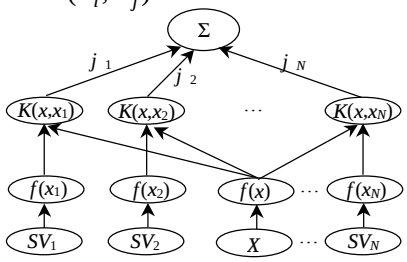

图 2 基于数据挖掘的支持向量机结构图

Fig. 2 Structure of support vector machines based on data mining

2.3.2 SVM算法的具体实现

在此选取Shevade等人提出的改进SMO算法[15]

来确定拉格朗日乘子 a_i 和 a_i^* 以及阈值 b,它是目前实现SVM算法中效率最高的一种,其具体实现如图3所示。其步骤为:

(1) 输入历史数据并进行预处理。数据预处理方法采用本文提出的联合数据挖掘技术,形成具有高度相似性气象特征的训练和测试样本集。

(2) 对SVM的模型参数进行初始化。将拉格朗日乘子 a_i 和 a_i^* 以及阈值b赋以随机的初始值。

(3) 利用训练样本建立形如式(9)的目标函数,然后采用改进的SMO算法求解目标函数式(9),得到 a_i 和 a_i^* 以及b的值。

(4) 将得到的参数值代入式(10),用测试样本计算未来某一时刻的预测值。

(5) 计算误差函数,当误差的绝对值小于预先设定的某个正数时,则结束学习过程(或设置迭代次数控制学习过程),否则返回步骤(3)继续学习。

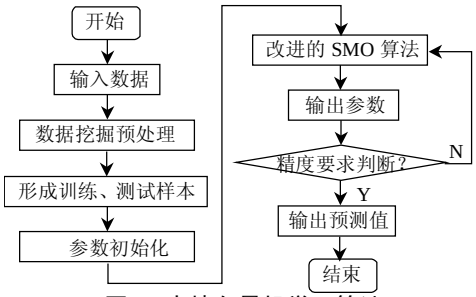

图 3 支持向量机学习算法

Fig. 3 Algorithm of support vector machines

3 实证分析

3.1 系统简介

运用本文所提新方法,对山西某地区电网短期负荷预测进行研究。该地区电网的电力负荷变化规律受气象影响非常显著,季节气温变化明显,年温差较大,雨量相对集中,农业负荷比例很大,因此按照通常经典预测方法和普通人工神经网络方法预测有较大困难,这些方法没有考虑气象诸因素的影响,历史负荷受到的干扰严重,明显导致预测结果不准确。下面采用本文提出的新方法,对该地区电网的负荷预测工作进行实证分析,并与单纯的SVM模型、普通人工神经网络模型进行预测精度分析比较。

3.2 样本选择

选取该电网自 2002 年 9 月—2004 年 8 月期间的负荷数据库作为训练样本,2004 年 9 月—2004 年 12 月期间的负荷数据作为测试样本。对于训练样本,通过本文提出的数据挖掘技术找出和预测点在气象特性和预测时段都相同的数据作为 SVM 中的 y 值,相应的 x 值(即样本输入量)分为以下几类:

(1) $A=\{a_1,a_2,\cdots,a_n\}$。预测日之前 n 日内的在预测时段的负荷数据。

(2) $B=\{b_1,b_2,\cdots,b_m\}$。预测日前一日预测时段之前 m 的负荷数据。

(3) $C=\{c_1,c_2,\cdots,c_m\}$。预测日的气象预报,共 s 个数据,包含平均气温、最高气温、最低气温、日降雨量。

(4) $D=\{d_1,d_2,\cdots,d_m\}$。预测日之前 n 日内的每日气象数据,其中任何一个元素 d_i 包含 s 个如上所述的气象数据。

对于输入量,还可以根据实际情况考虑负荷的工作日与非工作日、星期属性等其他因素。

3.3 参数分析

本文采用以下高斯函数作为核函数:

$$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp[-1/d^2(\boldsymbol{x}_i-\boldsymbol{x}_j)^2] \qquad (14)$$

式中d^2为高斯核的宽度参数。

参数的选择分别为d^2=20,c=100 和e=0.001。在本文研究中发现,核参数d^2和c对 SVM算法的表现起着非常重要的作用,当分别把c和e 固定在 10 和 0.001 时,训练集的标准均方差随着d^2的增大而增大。另一方面,测试集的标准均方差随着d^2的增大而起初减小,随后增大。这表明d^2的值太小(0.1~1),会对训练集造成过学习现象,d^2的值太大(100~100000),会对训练集造成欠学习现象。d^2的适合值应在 1~100 之间。由此可见:d^2对SVM的泛化性能起着关键作用,当分别把d^2和e固定在10和0.001时,训练集的标准均方差随着c的增大而单调减小;同时,当c的值从0.1增大到10时,测试集的标准均方差逐渐减小;当c的值从10增大到100时,测试集的标准均方差几乎保持为一常量;当c的值超过100时,测试集的标准均方差开始增大,其原因在于小的c值会对训练数据造成欠学习现象,c值太大容易对训练数据造成过学习现象而导致泛化性能恶化,因此,c的适合值应在10到100之间;当d^2和c固定不变时,e的变化对训练集和测试集的标准均方差的影响不大,这表明SVM的性能对e不敏感;当分别把d^2和c都固定在10时,训练集和测试集的标准均方差非常稳定,因而不受e 值变化的影响;一般情况下,支持向量的数目随着e 的增大而减小,然而大的e 值也能降低数据点的逼近精度,因此e 不能太大,根据训练过程,当e=0.001时,支持向量的数目较少,而数据点的逼近精度也较高。

普通神经网络模型采用标准的3层BP网络,分

别为输入层、隐含层和输出层。输入层的节点数为输入向量的分量数，取为9，其中5个为相似日同一时刻的负荷数值，由于普通神经网络模型并没有对负荷序列进行数据挖掘预处理，因此负荷序列不具有特定的气象特征，为了考虑气象因素影响，还需要再加上4个气象影响因素分量，即：日最高温度、日最低温度、日平均温度和日降雨量；输出层节点数为1，即短期负荷预测值；隐含层的节点数由经验取为5。网络初始化时，给定任意初始结构，将各连接权值$W_{ij}^{(l)}$随机置到小的随机数，给定学习训练次数(初设为100次)，学习率取0.2，初始衰减率取为5×10^{-4}，并设定学习速度h=0.01。

表1显示了本文方法(DMSVM)与单纯的SVM方法和BP神经网络方法训练精度的比较。由表1可知, DMSVM算法的e值(见下节)小于SVM算法和BP算法的e值。当训练5000次时所用训练时间DMSVM算法为5s,而SVM算法为30s，BP算法则为547s。由此可见,在收敛性和训练速度上, DMSVM算法明显优于SVM算法和BP算法。

表 1 DMSVM 与 SVM 模型和 BP 模型的 *e* 值
Tab.1 Comparison of *e* between DMSVM, SVM and BP

训练次数	*e*/%		
	DMSVM	SVM	BP
100	2.530	2.745	3.105
200	1.891	2.099	2.502
500	1.134	1.325	2.121
1000	1.027	1.208	1.848
2000	0.983	1.103	1.447
5000	0.331	0.566	1.282

3.4　误差测量与结果分析

本文选取平均相对误差e作为各种方法预测效果判断的根据：

$$e=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{A(i)-F(i)}{A(i)}\right|\times100\% \qquad (15)$$

式中　$A(i)$和$F(i)$分别为实际负荷值和预测负荷值。本文以 3%为评判标准，若某点的$e>3\%$ ，则判定该点预测结果为不合格。定义r为该工作日负荷预测结果的不合格率：

$$r=[num(e>3\%)/n]\times100\% \qquad (16)$$

式中　$num(e>3\%)$为某工作日预测结果平均相对误差超过3%的个数，n为负荷点总数。

为突出本文新方法的先进性，用本文模型（简记为DMSVM）和单纯的SVM方法、普通人工神经网络模型（ANN）对2004年9月12日—25日连续14日某点的负荷和8月7日的24h负荷进行预测，并比较三者的精度，其结果见表2和表3。从表2和表3可看出本文方法比其他2种方法的预测误差明显要小。由表2知：本文方法的相对误差总平均值为2.432%，比SVM法的3.091%和ANN法的3.584%小，本文方法的不合格率平均值为4.494%，也均比SVM法的5.476%和ANN法的6.625%小。

图4为3种模型对连续14日负荷的拟合曲线与实际负荷曲线的对比分析图。从图4可看出本文方法的拟合精度最好，可见本文所提方法预测效果好、精度高，是一种新的实用型方法，按此方法处理气

表 2　2004 年 9 月 12 日—25 日预测误差的比较
Tab. 2　Comparison of forecasting error from September 12, 2004 to September 25, 2004

日期	DMSVM		SVM		ANN	
	e/%	*r*/%	*e*/%	*r*/%	*e*/%	*r*/%
2004-9-12	2.37	4.197	2.70	5.486	3.07	6.375
2004-9-13	2.62	4.445	2.75	5.149	3.14	6.038
2004-9-14	2.64	5.833	6.01	7.222	7.98	8.234
2004-9-15	1.20	3.750	1.34	4.197	1.45	5.086
2004-9-16	1.61	4.208	1.73	6.181	2.95	7.071
2004-9-17	3.33	3.861	4.91	6.875	4.32	7.159
2004-9-18	1.65	5.250	1.37	3.403	1.70	6.734
2004-9-19	3.84	4.556	3.18	5.375	4.03	6.264
2004-9-20	1.95	4.319	2.01	5.250	2.27	6.361
2004-9-21	3.29	3.972	6.81	6.128	7.23	7.017
2004-9-22	1.60	5.361	1.78	6.086	2.27	7.197
2004-9-23	3.31	4.667	3.58	5.070	3.81	6.969
2004-9-24	2.56	4.420	2.76	5.764	3.19	6.875
2004-9-25	2.08	4.083	2.34	4.484	2.76	5.373

表 3　2004年8月7日24h预测误差比较
Tab. 3　Comparison of the 24-hour forecasting error of August 7, 2004

预测时刻	*e*/%		
	DMSVM	SVM	ANN
00:00:00	0.38	1.56	2.44
01:00:00	1.73	2.96	3.47
02:00:00	0.87	2.01	3.46
03:00:00	1.54	1.88	3.11
04:00:00	1.16	2.74	2.54
05:00:00	2.78	3.02	3.08
06:00:00	3.11	3.22	3.36
07:00:00	2.41	4.17	4.47
08:00:00	2.63	3.32	4.05
09:00:00	2.98	4.16	5.23
10:00:00	1.50	3.33	6.59
11:00:00	2.33	4.45	5.56
12:00:00	2.68	4.78	5.38
13:00:00	3.21	4.90	5.99
14:00:00	3.16	4.89	6.08
15:00:00	2.61	4.22	5.32
16:00:00	2.86	3.64	4.00
17:00:00	1.72	3.31	4.44
18:00:00	0.57	2.58	2.76
19:00:00	1.48	3.23	3.87
20:00:00	2.12	3.77	4.02
21:00:00	1.93	2.99	3.33
22:00:00	2.55	3.06	4.41
23:00:00	1.08	3.71	3.99

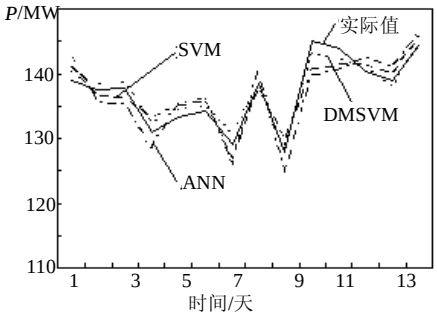

图 4 三种模型的拟合曲线图

Fig. 4 Fitting curve of three models

象影响问题，能够有效地提高短期负荷预测的精度，能相对简化负荷预测模型，免去SVM建模时的气象特征输入量，同时，SVM方法能有效地克服普通网络的缺点，使预测误差进一步减小，预测精度有了较大的提高，达到了实用的要求。

4 结论

（1）本文充分考虑了气象中的主要影响因素对负荷预测的影响，提出了基于模糊分类器和灰色关联分析的数据挖掘技术，对训练样本进行了有效的约简，提高了训练速度，并使预测结果自然地含有气象因素，不需要人工干预。

（2）经过数据挖掘预处理后，用具有更好泛化能力和全局寻优能力的支持向量机建立预测模型，较好地解决了有限样本学习问题，并有唯一的全局最优解。

（3）经过与单纯的SVM模型和ANN模型的实际预测比较，证明本文所建模型较大地提高了短期负荷预测的精度和系统的实用性，易于软件实现，能够有效地应用于电力负荷预测的管理工作。

参考文献

[1] 谢宏，程浩忠，张国立，等．基于粗糙集理论建立短期电力负荷神经网络预测模型[J]．中国电机工程学报，2003，23(11)：1-4.
Xie Hong，Cheng Haozhong，Zhang Guoli，et al．Applying rough set theory to establish artificial neural networks for short term load forecasting[J]．Proceedings of the CSEE，2003，23(11)：1-4(in Chinese).

[2] 郎能灵，侯志俭．小波模糊神经网络在电力系统中短期负荷预测中的应用[J]．中国电机工程学报，2004，24(1)：24-29.
Tai Nengling，Hou Zhijian．New short-term load forecasting principle with the wavelet transform fuzzy neural network for the power system [J]．Proceedings of the CSEE，2004，24(1)：24-29(in Chinese).

[3] 姚李孝，姚金雄，李宝庆，等．基于竞争分类的神经网络短期电力负荷预测[J]．电网技术，2004，28(10)：45-48.
Yao Lixiao，Yao Jinxiong，Li Baoqing，et al．Short-time load forecasting using neural network based on competitive learning classification[J]．Power System Technology，2004，28(10)：45-48(in Chinese).

[4] 谢开贵，李春燕，俞集辉．基于遗传算法的短期负荷组合预测模型[J]．电网技术，2001，25(8)：20-23.
Xie Kaigui，Li Chunyan，Yu Jihui．Genetic algorithm based combination forecasting model for short term load [J]．Power System Technology，2001，25(8)：20-23(in Chinese).

[5] Senjyu T．One-hour-ahead load forecasting using neural network [J]．IEEE Trans．on Power Syst．，2002，17(1)：113-118.

[6] Youshen Xia，Jun Wang．A general projection neural network for solving monotone variational inequalities and related optimization problems[J]． IEEE Transactions on Neural Networks，2004，15(2)：318-328.

[7] Shevade S K，Keerthi S S，Bhattacharyy C，et al． Improvements to SMO algorithm for SVM regression[J]．IEEE Trans．on Neural Network，2000，11(5)：1188-1193.

[8] Suykens J A K，Lukas L，Vandewalle J．Approximation using least squares support vector machine[A]．IEEE Int Symposium 071 Circuits and Systems[C]．Geneva，2000，(1)：757-760.

[9] Chen L D，Toru S．Data mining methods，applications，and tools[J]．Information system management，2000，17(1)： 65-70.

[10] 于之虹，郭志忠．数据挖掘与电力系统[J]． 电网技术，2001，25(8)：58-62.
Yu Zhihong，Guo Zhizhong．Data mining and power system [J]．Power System Technology，2001，25(8)：58-62(in Chinese).

[11] 赵登福，王蒙，张讲社，等．基于支持向量机方法的短期负荷预测[J]．中国电机工程学报，2002，22(4)：26-30.
Zhao Dengfu，Wang Meng，Zhang Jiangshe，et al．A supprot vector machines approach for short-term load forecasting [J]．Proceedings of the CSEE，2002，22(4)：26-30(in Chinese).

[12] 李元诚，方廷健，于尔铿．短期负荷预测的支持向量机方法研究[J]．中国电机工程学报，2003，23(6)：55-59.
Li Yuancheng，Fang Tingjian，Yu Erkeng．Study of support vector machines for short-term load forecasting[J]．Proceedings of the CSEE， 2003，23(6)：55-59(in Chinese).

[13] 张林，刘先珊，阴和俊．基于时间序列的支持向量机在负荷预测中的应用[J]．电网技术，2004，28(10)：38-41.
Zhang Lin，Liu Xianshan，Yin Hejun．Application of support vector machines based on time sequence in power system load forecasting [J]．Power System Technology，2004，28(10)：38-41(in Chinese).

[14] Chapelle O，Vapnik V．Choosing multiple parameters for support vector machines[R]．New York：AT&T Research Labs，2001.

[15] ShevadeS K，Keerthi S S，Bhattacharyy C，et al． Improvements to SMO algorithm for SVM regression[J]．IEEE Transaction Neural Networks，2000，11(5)：356-362.

收稿日期：2006-04-17。

作者简介：

牛东晓(1962—)，男，汉，安徽宿县人，教授，博士生导师，主要研究方向为电力负荷预测技术、电力市场技术、综合评价理论方法应用、技术经济评价理论方法；

谷志红(1980—)，男，汉，河北石家庄人，华北电力大学工商管理学院硕士研究生，主要研究电力市场理论、电力负荷预测 laobing618@163.com;laobing618@126.com；

王会青(1980—)，女，汉，山西临汾人，硕士研究生，主要研究电力市场营销、综合评价理论方法应用。

（责任编辑 喻银凤）