基于粗集属性约简的电力供应量SVM回归预测

汤义强[1],毛军军[1,2],李 侠[1],程白彬[1]
(1.安徽大学 数学科学院,安徽 合肥 230039;
2.安徽大学 计算智能与信号处理教育部重点实验室,安徽 合肥 230039)

摘 要:采用基于粗糙集属性约简的支持向量机回归预测模型对我国电力供应量进行预测。根据电力供应量及其影响因素的历史数据建立决策表,利用动态层次聚类法对决策表中的连续属性进行了离散化;运用属性约简算法进行约简,提取出主要因素,并将其作为样本的特征,应用支持向量机回归预测模型对电力供应量进行预测。五年预测结果表明:与SVR模型相比,结合了属性约简方法的RS&SVR模型充分利用了更少但是主要的预测因子的信息,预测精度有一定提高,应用效果较好。

关键词:支持向量机回归;粗糙集;属性约简;预测
中图分类号:TP18 **文献标识码**:A **文章编号**:1673-629X(2010)09-0048-04

China's Power Supply SVM Regression Forecast Based on Rough Set Attribute Reduction

TANG Yi-qiang[1], MAO Jun-jun[1,2], LI Xia[1], CHENG Bai-bin[1]
(1. School of Mathematical Sciences, Anhui University, Hefei 230039, China;
2. Ministry of Edu. Key Lab. of Intelligent Computing & Signal Processing, Anhui Univ., Hefei 230039, China)

Abstract: The application of RS&SVR method, which is support vector machine regression (SVR) based on attribute reduction algorithm of rough sets, on forecast of China's power supply is dealt with in this paper. According to historical data of power output and its influencing factors, a decision table is built up, and discretization of continuous attributes in the table is done by means of dynamic layer cluster. Using the attribute reduction algorithm to eliminate some redundant attributes from the table, the kernel factors are determined. Taking these kernel factors as the attributes of both training and testing samples, the power supply forecasting is conducted. Five-year forecasting results show that, compared with SVR which chooses attributes of input vectors in light of experience, the method of RS&SVR could make use of less but cardinal predictors' information, and the forecasting accuracy is improved.

Key words: support vector machine regression; rough sets; attribute reduction; forecast

0 引 言

电力供应在国民经济发展中承担重要角色,目前我国的电力供应还不能完全满足国民经济发展需求,供需矛盾在某些情况下还比较突出。因此,有必要对电力供求及早规划,做出合理决策,为国民经济安全、稳步发展提供保障。支持向量机(Support Vector Machines,SVM)是V. Vapnik[1]等人在统计学习理论基础上提出的小样本学习算法,并在广泛的分类和回归分析应用中表现出优越性能。目前支持向量机回归已被用于诸多领域,如股市预测、气象预测、电力负荷预测、粮食产量预测等[2~5]。影响电力供应的因素往往是不确定的、复杂的,并且电力产量一般以年或季度为统计时间标度,因而电力供应量的预测是一个典型的非线性、小样本问题。

支持向量机有很多优点,但是对于给定的数据样本,它不能对样本的特征属性给予评价,所以次要的属性往往对模型的训练和检验造成干扰,影响预测精度。因此,有的文献[4,6]尝试对输入样本的特征属性或输入样本权重做一些处理。粗糙集(Rough Sets,RS)理论在处理不确定知识、消除冗余信息、发现样本数据属性之间的本质关系上具有优势,它不依赖模型的先验知识,只从数据中寻找描述系统正常模型的最小预测规则集,具有客观性[7]。

收稿日期:2009-12-15;修回日期:2010-03-02
基金项目:国家自然科学基金(60675031);安徽省高等学校省级自然科学研究项目(KJ2008B093);安徽大学创新性实验项目(30007)
作者简介:汤义强(1984-),男,安徽巢湖人,硕士研究生,研究方向为粗糙集、运筹控制;毛军军,博士,副教授,研究方向是粒计算、智能计算理论及其应用。

文中将粗糙集属性约简方法结合在支持向量机回归样本的预处理阶段,建立对电力产量及其影响因素的历史数据样本的回归分析模型,并与传统支持向量机回归模型作对比分析。

1　理论概述

1.1　支持向量机回归原理

支持向量机是基于 VC 维理论和结构风险最小化(SRM)准则[1]的具体实现。支持向量回归(Support Vector Regression,SVR)的目的是利用少数支持向量代表整个样本集来寻找一个对训练数据集拟合得最优的回归超平面[8]。

考虑一给定样本数据 $\{(x_i,y_i)\mid i=1,2,\cdots,n\}$,其中 $x_i\in R^m$,$y_i\in R$,m 为样本特征维数,n 为样本个数。对于线性回归,要求回归函数形为

$$f(x)=<w,x>+b=\sum_{i=1}^{n}(a_i-a_i^*)<x,x_i>+b$$

其中 a_i,a_i^* 为 Lagrange 乘子,满足约束条件 $\sum_{i=1}^{n}(a_i-a_i^*)=0$ 且 $0\leqslant a_i,a_i^*\leqslant C(i=1,2,\cdots,n)$,少数 $a_i-a_i^*\neq 0$ 对应的样本点即为支持向量,寻找回归函数的详细过程可在文献[9]中找到。

对于非线性支持向量回归问题,通常引入一个非线性映射 Φ 将数据 x 映到更高维的 Hilbert 空间,在此空间进行线性回归,它对应原空间的非线性回归。核函数技术对低维空间输入实现了高维特征空间的点积,从而避免了在高维空间内的复杂点积运算,即 $K(x_i,x_j)=<\Phi(x_i),\Phi(x_j)>$,相似的,非线性回归的最优超平面解析式具有这样的形式:

$$f(x)=\sum_{i=1}^{n}(a_i-a_i^*)<\Phi(x),\Phi(x_i)>+b$$
$$=\sum_{i=1}^{n}(a_i-a_i^*)K(x_i,x_j)+b$$

常用核函数有:p 阶多项式核函数 $K(x_i,x_j)=(<(x_i,x_j>+c)^p(p\in N,c\geqslant 0)$;高斯径向基核函数 $K(x_i,x_j)=\exp\{-\|x_i-x_j\|^2/2\sigma^2\}$;Sigmoid 核函数 $K(x_i,x_j)=\tanh(\theta+\upsilon<x_i,x_j>)$。

1.2　粗集相关概念和属性集相对约简的描述

实际应用粗集理论分析数据时,一般是对决策表进行处理。

定义 1(决策表)[7]由四元组 $S=(U,A,V,f)$ 构成的信息表知识表达系统称为决策表,其中 U 为论域,即非空有限对象集;$A=C\cup D$ 为非空有限属性集,C,D 分别为条件属性集和决策属性集且 D 非空;$V=\bigcup V_a$ 为属性 $a\in A$ 的值域构成的集合;$f:U\times A\to V$ 为信息函数,指定论域 U 中每个对象各个属性的取值。当 $D=\{d\}$ 只有一个决策属性时,称此决策表为单一决策表。

定义 2(上、下近似)[7]对任意 $X\subseteq U$ 和 X 上的等价关系 B,在此关系下 X 的上、下近似集定义为:

$$B^-(X)=\{x\mid x\in U\wedge[x]_B\cap X\neq\varnothing\}$$
$$B_-(X)=\{x\mid x\in U\wedge[x]_B\subseteq X\}$$

当 $B^-(X)\neq B_-(X)$,称 X 为一个 B 粗集。

定义 3(正域)[7]设 P,Q 是 U 上的两个等价关系,则 Q 的 P 正域定义为:

$$POS_P(Q)=\bigcup_{X\in U\mid_Q}P_-(X)$$

它表示的是根据 $U\mid_P$ 的信息可以准确地划分到关系 Q 的等价类中开集的集合。

定义 4(约简)[7]设 P,Q 是论域 U 上的两个等价关系簇,若 $S\subset P$ 是 P 的 Q 独立子集(即 $\forall r\in S$,$POS_{S\setminus\{r\}}(Q)\neq POS_S(Q)$)且有 $POS_S(Q)=POS_P(Q)$,则称 S 是 P 的 Q 约简。

决策表信息系统中的一个属性对应着一个等价关系,条件属性集和决策属性集对论域的划分形成了对论域样本上的分类知识。属性约简的目的就是从条件属性集中发现部分必要属性,相对于决策属性,这部分必要属性具有和所有条件属性相同的分类能力。设 $S=(U,A,V,f)$ 是一个决策表,$U=\{x_1,x_2,\cdots,x_n\}$ 是论域,$A=C\cup\{d\}$,$a\in A$ 是一个连续属性,属性 a 的值域为 $[\min_a,\max_a]$,其中 $\min_a,\max_a$ 分别为属性 a 的最小值和最大值。Pawlak 提出的粗集理论只能处理离散属性,因此含有连续属性的决策表需要进行某种离散化处理。文中采用聚类分析中凝聚层次聚类法实现决策表的离散化[10]。依据定义 4,可以找到条件属性集合 C 的一个约简。条件属性集合 C 的相对约简 C^{red} 就是条件属性集合 C 相对于决策属性 $\{d\}$ 的最大独立子集[7]。

2　数据准备与预处理

影响电力供应量的因素很多,可能有经济总体发展水平、主要工农业产品的生产、人民的生活水平等方面。文中参考文献[11],选取了在这些方面具有代表性的因素:电力产量(亿千瓦小时)d;国内生产总值(亿元)a_1;第一产业(亿元)a_2;第二产业(亿元)a_3;第三产业(亿元)a_4;人均国内生产总值(元/人)a_5;原油产量(万吨)a_6;生铁产量(万吨)a_7;原煤产量(亿吨)a_8;铁路货运量(万吨)a_9。1978～2008 年电力产量及其影响因素年度统计数据如表 1 所示。

表 1 1978 ~ 2008 年电力产量及其影响因素年度统计数据

year	d	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9
1978	2566.00	3645.2	1027.5	1745.2	872.5	381	10405.00	3479.00	6.18	110119
1979	2820.00	4062.6	1270.2	1913.5	878.9	419	10615.00	3673.00	6.35	111893
1980	3006.00	4545.6	1371.6	2192.0	982.0	463	10595.00	3802.00	6.20	111279
1981	3093.00	4889.5	1559.5	2255.5	1076.6	492	10122.00	3417.00	6.22	107673
1982	3277.00	5330.5	1777.4	2383.0	1163.0	528	10212.00	3551.00	6.66	113495
1983	3514.00	5985.6	1978.4	2646.2	1338.1	583	10607.00	3738.00	7.15	118784
1984	3770.00	7243.8	2316.1	3105.7	1786.3	695	11461.00	4001.00	7.89	124074
1985	4107.00	9040.7	2564.4	3866.6	2585.0	858	12490.00	4384.00	8.72	130709
1986	4495.00	10274.4	2788.7	4492.7	2993.8	963	13069.00	5064.00	8.94	135635
1987	4973.00	12050.6	3233.0	5251.6	3574.0	1112	13414.00	5503.00	9.28	140653
1988	5452.00	15036.8	3865.4	6587.2	4590.3	1366	13705.00	5704.00	9.80	144948
1989	5848.00	17000.9	4265.9	7278.0	5448.4	1519	13764.00	5820.00	10.54	151489
1990	6212.00	18718.3	5062.0	7717.4	5888.4	1644	13831.00	6238.00	10.80	150681
1991	6775.00	21826.2	5342.2	9102.2	7337.1	1893	14099.00	6765.00	10.87	152893
1992	7539.00	26937.3	5866.6	11699.5	9357.4	2311	14210.00	7589.00	11.16	157627
1993	8395.00	35260.0	6963.8	16454.4	11915.7	2998	14524.00	8739.00	11.50	162794
1994	9281.00	48108.5	9572.7	22445.4	16179.8	4044	14608.00	9741.00	12.40	163216
1995	10070.00	59810.5	12135.8	28679.5	19978.5	5046	15004.95	10529.27	13.61	165982
1996	10813.00	70142.5	14015.4	33835.0	23326.2	5846	15733.39	10722.50	13.97	171024
1997	11356.00	78060.8	14441.9	37543.0	26988.1	6420	16074.14	11511.41	13.73	172149
1998	11670.00	83024.3	14817.6	39004.2	30580.5	6796	16100.00	11863.67	12.50	164309
1999	12393.00	88479.2	14770.0	41033.6	33873.4	7159	16000.00	12539.24	12.80	167554
2000	13556.00	98000.5	14944.7	45555.9	38714.0	7858	16300.00	13101.48	12.99	178581
2001	14808.00	108068.2	15781.3	49512.3	44361.6	8622	16395.87	15554.25	13.81	193189
2002	16540.00	119095.7	16537.0	53896.8	49898.9	9398	16700.00	17084.60	14.55	204956
2003	19105.75	135174.0	17381.7	62436.3	56004.7	10542	16959.98	21366.68	17.22	224248
2004	22033.09	159586.7	21412.7	73904.3	64561.3	12336	17587.33	26830.99	19.92	249017
2005	25002.60	184088.6	22420.0	87364.6	73432.9	14053	18135.29	34375.19	22.05	269296
2006	28657.26	213131.7	24040.0	103162.0	84721.4	16165	18476.57	41245.19	23.73	288224
2007	32815.53	259258.9	28627.0	124799.0	103879.6	19524	18631.82	47651.63	25.26	314237
2008	34668.82	302853.4	34000.0	146183.4	120486.6	22698	19001.24	47067.41	27.88	330354

资料来源:《中国统计年鉴 2009》

根据历史数据作如下处理:

Step1 将原数据表中的电力年产量作为电力供应量并视为决策属性 d,其余因素集合$\{a_1,a_2,\cdots,a_9\}$ 视为条件属性集 C,建立一个单一决策表;

Step2 利用层次聚类法对每个属性分别聚类,实现连续属性的离散化,划分得太粗或太细,会导致信息的缺失或冗余,本例类别数限定在 3 ~ 8 类;

Step3 计算 C 的相对约简 C^{red},根据相对约简,剔除原数据表中冗余属性,得到属性约简后的样本;

Step4 采用最小 - 最大值规范化方法,将属性约简后样本的每个属性分别规范化到[- 1,1] 区间,计算公式为 $v' = 2\dfrac{v-\min_a}{\max_a-\min_a}-1, v\in[\min_a,\max_a]$;

Step5 将得到的规范化样本数据进行拆分,用 1978 年~2003 年的数据作训练样本,2004 年~2008 年数据做测试样本。训练用的 26 个样本分成 10 份,以作交叉验证来优选 SVR 模型的控制参数。

通过以上步骤,对实验样本作了简化降维,降低了样本数据的复杂性,同时提取出了影响电力产量的一些主要因素。

3 支持向量机回归结果与分析

采用来自 http://www.isis.ecs.soton.ac.uk/resources/svminfo 的通用 MATLAB SVM Toolbox 实现[12]。核函数选择了高斯径向基函数 $K(x_i,x_j)=\exp\{-\|x_i-x_j\|^2/2\sigma^2\}$,损失函数选用线性 ε - insensitive 函数 $L(f(x),y)=\max(0,|f(x)-y|-\varepsilon)$。

参数 C,σ,ε 的选择采用交叉验证。为作比较,将原样本也按上面 2 节中步骤 4~5 处理。两种回归的相关参数和结果见表 2~4。

表 2　参数值和支持向量个数

模型	C	σ	ϵ	nSV
SVR	400	12	0.01	16(61.5%)
RS & SVR	40	11	0.01	18(69.2%)

表 3　电力产量预测值与实际值

年份	SVR	RS & SVR	实际值
2004	22230.19	21958.09	22033.09
2005	25577.35	25293.83	25002.60
2006	28815.96	28560.00	28657.26
2007	32684.36	32924.35	32815.53
2008	34448.61	35163.74	34668.82

单位:亿千瓦小时(100 million kWh)

表 4　模型预测精度比较

模型	RMSE	THEIL_IC	MAPE	MAX_APE
SVR	303.3361	0.0023	0.9564	2.2987
RS & SVR	267.0918	0.0021	0.7208	1.4276

其中:

(1) $\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-y_i)^2}$

(2) $\mathrm{THEIL\_IC}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-y_i)^2}/\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i^2}+\sqrt{\frac{1}{n}\sum_{i=1}^{n}y_i^2}\right)$,希尔不等系数通常介于 0 ~ 1,且值越小表明预测精度越高;

(3) $\mathrm{MAPE}=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i-y_i}{y_i}\times 100\right|$,一般认为 MAPE 的值小于 10,则预测精度较高;

(4) $\mathrm{MAX\_APE}=\max\left(\left|\frac{\hat{y}_i-y_i}{y_i}\times 100\right|\right)$

以上式中 $\hat{y}_i$,y_i 分别为预测值和实际值。

从支持向量个数来看,RS&SVR 模型与 SVR 模型大体相当,前者仅比后者多了两个支持向量;从预测效果来看,RS&SVR 模型具有更小的均方根误差和希尔不等系数,说明预测值更接近实际值,拟合效果更好。因此,RS&SVR 模型用于回归预测,方法是有效的。

4　结束语

支持向量机回归应用于电力供应量分析是一种较理想的模型,而粗糙集理论在处理不确定知识、消除冗余信息方面具有优势,将属性约简方法应用于数据的预处理,简化了样本的特征维。两者的结合比单一模型产生了更好的结果。相信粗糙集理论方法与支持向量机的结合在应用领域具有广阔的前景。

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] 施燕杰.基于支持向量机(SVM)的股市预测方法[J].统计与决策,2005(4):123-125.

[3] 滕卫平,俞善贤,胡　波,等.SVM 回归法在汛期旱涝预测中的应用研究[J].浙江大学学报:理学版,2008,35(3):343-347.

[4] 王晓红,吴德会.基于 WLS-SVM 回归模型的电力负荷预测[J].微计算机信息,2008,24(4):312-314.

[5] 程　伟,张燕平,赵　姝.支持向量机在粮食产量预测中的应用[J].安徽农业科学,2009,37(8):3347-3348.

[6] 姜德民,王　磊,徐义田,等.基于粗糙集理论与支持向量回归的预测模型[J].统计与决策,2008(10):32-33.

[7] 王国胤.Rough 集理论与知识获取[M].西安:西安交通大学出版社,2001.

[8] 王定成,方廷健,高理富,等.支持向量机回归在线建模及应用[J].控制与决策,2003,18(1):89-91.

[9] Cristianini N,Shawe-Taylor J.支持向量机导论[M].李国正,王　猛,曾华军,译.北京:电子工业出版社,2004:98-104.

[10] 苗夺谦.Rough Set 理论中连续属性的离散化方法[J].自动化学报,2001,27(3):296-302.

[11] 袁　卫,庞　皓,曾五一,等.统计学习题与案例[M].北京:高等教育出版社,2006:236-241.

[12] Gunn S R. Support vector machines for classification and regression[R]. Faculty of Engineering, Science and Mathematics, School of Electronics and Computer Science, University of Southampton,1998.

(上接第 47 页)

[3] 侯　勇.关键字驱动的自动化测试系统的研究[D].西安:西安电子科技大学,2006.

[4] 徐志发.从 eTOM/SID 看电信商业智能系统的发展思路[J].电信科学,2008(1):17-21.

[5] 徐振良,樊滨温,王志鹏.关键字驱动技术在 SAFS 中的研究[J].软件时空,2006,22:270-272.

[6] 冯玉才,唐　艳,周　淳.关键字驱动自动化测试的原理和实现[J].计算机应用,2004,24(8):140-142.

[7] 凌永发,张云生,郭秀萍.软件测试自动化中的脚本技术[J].云南民族学院学报,2002(11):544-548.

[8] 芦彩林,丁刚毅..Net 框架下测试脚本自动生成技术研究[J].微计算机应用,2008,29(5):101-104.

[9] 赵斌飞,刘　磊.测试脚本自动生成器的设计与实现[J].计算机科学,2008(6):276-279.

[10] Fewster M,Graham D.软件测试自动化技术与实例详解[M].舒智勇,包晓露,焦　跃,等译.北京:电子工业出版社,1999:450-560.

[11] Mosley D J,Posey B A.软件测试自动化[M].邓　波,黄丽娟,译.北京:机械工业出版社,2003:440-442.

[12] Yuan Xun,Atif M.Generating event sequence-based test cases using GUI runtime state feedback[J]. IEEE Transactions on Software Engineering,2010,36(1):81-95.