RS-LS-SVR 模型在煤与瓦斯突出预测的应用①

彭 泓，高 攀

(辽宁工程技术大学 电气与控制工程学院，葫芦岛 125105)

摘 要：在综合研究了各种算法的基础上，将粗集理论和最小二乘支持向量机算法结合，充分利用了粗集算法能够去除冗余信息，最小二乘支持向量机能够精确加快收敛速的优点。利用具体网络建立一个突出预测机制，并利用本预测机制对矿井瓦斯突出情况进行模拟预测。经过基于 MATLAB 工具箱的 BP 神经网络模型的实验对比表明，LS-SVR 能加快收敛速度。实验结果表明，基于 RS-LS-SVR 网络的预测模型可靠，收敛速度快，预测精度高，效果良好。

关键词：煤与瓦斯突出；粗糙集；最小二乘支持向量回归机；模拟预测

RS-LS-SVR Model in Predication of the Coal and Gas Outburst

PENG Hong, GAO Pan

(Faculty of Electrical and Engineering, Liaoning Technical University, Huludao 125105, China)

Abstract: Based on the comprehensive study of the various algorithms, with the rough set theory and the Least Squares Support Vector Regression, taking advantage of rough set method can remove redundant information, Least Squares Support Vector Regression can accurately accelerate the convergence speed advantages. Prominent use of a specific network prediction mechanism, and use this prediction of mine gas outburst mechanism to predict the situation. After based on the MATLAB neural network toolbox BP neural network method of experimental comparison shows that the LS-SVR can speed up the convergence rate. The experimental result reveals that Based on RS-LS-SVR neural network prediction model is reliable, fast convergence and high accuracy, good effect.

Key words: coal and gas outburst; rough set; LS-SVR; simulation and prediction

在我国煤炭生产中有 95%的煤矿都是地下作业，那么在井下作业就要面临着生产安全问题。煤与瓦斯突出事故不仅会造成采掘工作面和通风系统的破坏，同时大量煤与瓦斯以极快的速度喷出，还可能会充塞巷道，造成人员窒息和瓦斯爆炸、燃烧及煤（岩）埋人事故。有些学者通过采用人工神经网络来解决这种问题，但是神经网络算法收敛速度较慢，存在局部极小点而且需要无限样本，非线性映射的泛化能力较弱，同时存在过学习和欠学习的问题，有时较难达煤与瓦斯突出预测系统精度的要求。然而基于统计学习理论的支持向量机将输入变量通过非线性变换映射到高维特征空间，在高维特征空间中构造最优超平面，利用核函数巧妙避免了内积运算和“维数灾难”。支持向量机算法是求解凸二次优化问题，能够保证找到的极值解就是全局最优解，能较好地解决有限样本、非线性和高维数的问题，同时用粗糙集算法做前端处理，利用其定性分析能力出去原始数据冗余的属性，简化了学习样本，所以有必要采用基于 RS-LS-SVR 进行煤与瓦斯突出预测。

1 粗糙集与支持向量机原理

1.1 RS 理论[1]

粗糙集(Rough Sets,RS)理论的主要特点是具有很强的定性分析能力,或依据观察、度量到的某些不精确的结果而进行分类数据的能力,可以直接对不完备信息进行处理,而不必进行完备化。其基本原理为：

① 基金项目:国家自然科学基金(50874059)
收稿时间:2011-05-13;收到修改稿时间:2011-06-25

有非空有限集合 U,A，其中 U 为论域，A 为属性集，对于每一属性 $a \in A$，存在属性值的集合 $V_a = \{a(x) | \forall x \in U\}$ 称 $S = (U, A)$ 为信息系统。若 $A = C \cup D$，且 $C \cap D \neq \varnothing$，其中 C 为条件属性集，D 为决策属性集（一般 $D = \{d\}$），则称 $S = (U, C \cup D)$ 为决策系统。

设 $C \subseteq A, X \subseteq U$，分别定义 X 的 C 的上近似、C 下近似与边界为：

$$\begin{aligned} C^{-}(X) &= \bigcup \{Y \in U / C \mid Y \cap X \neq \varnothing\} \\ C_{-}(X) &= \bigcup \{Y \in U / C \mid Y \subseteq X\} \\ BN_C(X) &= C^{-}(X) - C_{-}(X) \end{aligned} \tag{1}$$

在条件 C 下，下近似 $C_{-}(X)$ 中的元素肯定属于集合 X；上近似 $C^{-}(X)$ 中的元素可能属于集合 X，反映了集合中元素的模糊性；边界 $BN_C(X)$ 反映了集合的不确定性。在一个决策系统中，各个条件属性之间往往存在某些程度上的依赖或关联，约简可以理解为不丢失信息的前提下，可以最简单地表示决策系统的决策属性对条件属性集合的依赖性或关联度。C 的所有约简的集合记为 $RED_D(C)$.C 的所有约简的交集叫做核：

$$CORE_D(C) = \bigcap RED_D(C) \tag{2}$$

一个决策系统可能存在多个约简。由此粗糙集理论提供了分析多余属性的方法，通过对决策表中属性值的处理实现简约。

1.2 最小二乘支持向量回归机的基本原理[2]

支持向量机回归就是凸二次规划问题来求全局最优解，避免了神经网络计算过程中出现的局部极小值。将最优化问题转化成求解其对偶问题（公式 4），使高维特征空间避免了“维数灾难”。而最小二乘支持向量回归机（LS-SVR）是把凸二次规划问题转化成线性方程组求解。以下对 LS-SVR 算法进行详细的说明[2-4]:

假设给定了训练样本 $S = \{(x_i, y_i), x_i \in R^n, y_i \in R\}_{i=1}^{l}$，其中 x_i 是第 i 个输入学习样本向量,且为一维列向量 $x_i = [x_i^1, x_i^2, ..., x_i^d,]$，l 为样本数目。$y_i \in R$ 为对应的目标值。线性回归问题的目标就是找到回归函数:

$$f(x) = w^T \phi(x) + b \tag{3}$$

其中,w 为权重向量;b 称为偏置量,也称阈值。回归问题对应的优化问题为

$$\phi(w, \eta) = \frac{1}{2} \| w \|^2 + \frac{C}{2} \sum_{i=1}^{l} \eta_i^2 \tag{4}$$

其中，C 为惩罚参数，η_i 为误差变量。最小化上式为：

$$\begin{aligned} &\min_{w,d,\eta_i} \frac{1}{2} \| w \|^2 + \frac{C}{2} \sum_{i=1}^{l} \eta_i^2 \\ &s.t. y_i - (w\, \phi(x) + d) = \eta, i = 1, ..., n \end{aligned} \tag{5}$$

转化此函数的优化问题,可以构造 Lagrange 函数，同时引入 Lagrange 算子 α, α^*:

$$L(w, d, \eta, \alpha) = \frac{1}{2} \| w \|^2 + \frac{C}{2} \sum_{i=1}^{l} \eta_i^2 - \sum_{i=1}^{l} \alpha_i (w \bullet \phi(x) + d + \eta_i + y_i) \tag{6}$$

根据最优化理论,将 L 分别对各原变量求偏导得:

$$\begin{aligned} &\frac{\partial L}{\partial w} = 0, \rightarrow w = \sum_{i=1}^{l} \alpha_i \phi(x_i) \\ &\frac{\partial L}{\partial d} = 0, \rightarrow \sum_{i=1}^{l} \alpha_i = 0 \\ &\frac{\partial L}{\partial \eta} = 0, \rightarrow \alpha_i = C \eta_i \\ &\frac{\partial L}{\partial \alpha} = 0, \rightarrow y_i = w \phi(x) + d + \eta_i \end{aligned} \tag{7}$$

得到对偶最优化问题为

$$\begin{aligned} &\mathrm{W}(\alpha, \alpha^*) = \sum_{i=1}^{l} \alpha_i y_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j (< \phi(x_i) \phi(x_j) > + \frac{\lambda_{ij}}{C}) \\ &\lambda_{ij} = \begin{cases} 1, i = j \\ 0, i \neq j \end{cases} \end{aligned} \tag{8}$$

核函数为 $< \phi(x_i) \phi(x_j) >= K(x_i, x_j)$。优化问题转化为泛函最大值的求解问题：

$$\begin{aligned} &\max_{\alpha} \sum_{i=1}^{l} \alpha_i y_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j (< \phi(x_i) \phi(x_j) > + \frac{\lambda_{ij}}{C}) \\ &s.t. \sum_{i=1}^{l} \alpha_i = 0 \end{aligned} \tag{9}$$

根据 KKT 条件可得:其中只有部分参数 α_i 不为 0,它们就是问题中的支持向量(SV),而且可以得到:

$$\overline{w} = \sum_{i=1}^{l} \alpha_i^* \phi(x_i), \overline{d} = y_i - \frac{\alpha_i^*}{C} - \sum_{j=1}^{l} \alpha_j^* K(x_j, x_i) \tag{10}$$

最后得到的非线性回归决策函数为

$$f(x) = \sum_{i=1}^{k} \alpha_i^* K(x_i, x) + d \tag{11}$$

2 煤与瓦斯突出预测模型建立

基于 RS—LS-SVR 的煤与瓦斯突出模型是以粗集作为模型的前端处理对信息进行预处理，LS-SVR 作为模型的后处理，用于样本集的训练和完成预测工作。模型流程如图 1 所示。其主要实现步骤为：

Step1：收集整理煤矿井下煤与瓦斯突出样本集；

Step2：对样本数据进行归一化处理；

Step3：数据预处理，对连属性数据离散化形成决策表；

Step4：根据最小约简删除决策表中的冗余属性，优化属性核值进行 RS 简约，实现对网络输入特征的精简;

Step5：结合归一化后的样本数据集构建 LS-SVR 模型，选择适当的核函数及其参数进行训练，直至满足预期的训练目标；

Step6：利用训练好非线性关系进行突出预测；

Step7：对得到的结果进行误差检验，评价结果分析。

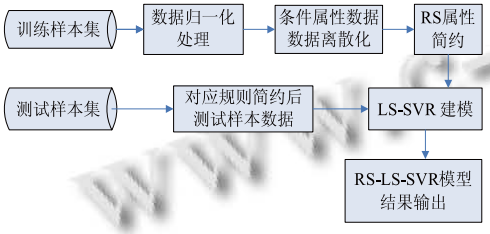

图 1 煤与瓦斯突出预测模型图

3 RS-LS-SVR模型应用

通过对煤矿现场的了解确定煤与瓦斯突出的主要影响因素，本文选取综合指标（D/K ）、地质构造（T）、煤层厚度（M）、最大钻屑量（S）、瓦斯涌出初速度（q）作为条件属性，对瓦斯突出进行预测，突出危险性中 0 代表安全，1 代表存在危险，2 代表瓦斯突出。预测的具体步骤如下：

1) 归一化处理

为了消除学习样本中各个因子由于量纲和单位不同带来的影响，需要对样本数据进行归一化处理，调整到[0，1]之间。具体处理办法如下所示：

① 使 $p = P^T$；

② 将矩阵 P 按照下述式子分别对每一行进行处理：

$$p = \frac{P(x) - \max P(x)}{\max P(x) - \min P(x)} \tag{12}$$

其中：p 为归一化后数据，P(m)为原始数据。

表 1 归一化时各输入变量的最大、最小值

变量	D/K	T	M	S	q
$\max P(x)$	0.22/11	5	2.3	6.8	14.5
$\min P(x)$	0.27/18	1	1.2	4.2	3.0

2) RS 简约

本文采用等距离法对上述一些数据进行离散化处理,把每一列的属性值分 4 个等级,得出各指标值转换成 Rough Set 的数据格式决策表。

表 2 决策表

序号	D/K	T	M	S	q	危险性
1	0	0	0	1	1	0
2	0	0	1	1	1	0
3	1	1	0	1	1	0
4	2	2	0	1	1	0
5	0	0	0	0	1	1
6	0	0	1	0	1	1
7	1	1	0	1	1	1
8	1	1	1	0	1	1
9	2	2	1	0	1	1
10	0	0	0	1	0	2
11	1	1	0	1	0	2
12	2	2	0	0	1	2

去掉冗余属性（T）没有出现不相容对策，地质构造（T）可以去掉，所以计算相容决策项的核值，去除冗余信息。通过（P,Q）相容算法计算每项决策的核值来去除冗余值。例如在第一项决策中 D/K0 M0 S1 q1→危险性 0,其中其中 S1 q1 为核值。因为规则 M0 S1 q1→危险性 0, D/K0 S1 q1→危险性 0 为真。而规则 D/K0 M0 q1→危险性 0, D/K0 M0 S1→危险性 0 为假。用类似方法可以消去条件属性的冗余值，得到核值表 3。

表 3 核值表

序号	D/K	M	S	q	危险性
1	X	X	1	1	0
2	0	X	1	1	0
3	X	0	1	1	0
4	X	0	1	1	0
5	X	X	0	1	1
6	X	X	0	1	1
7	1	X	1	1	1
8	X	X	0	1	1
9	X	1	0	1	1
10	X	X	X	0	2
11	X	X	X	0	2
12	2	0	0	X	2

去掉核值表中重复的记录 3 条得到优化核值，例

如第 3 与第 4 行事相同的，发现去掉所有冗余的对象后不会改变决策规则。优化后的核值表位 9 组数据[3]。

3) LS-SVR 训练

经粗糙集优化后样本数据大大减少，降低了模型训练的复杂度。把优化后的核值表带入采集样本的原值进行最小二乘支持向量机训练。支持向量机的核函数的选择有许多种，这里选择高斯径向基核函数。

$$K(x,x')=\exp\left\{\frac{-\|x-x'\|}{2\sigma^2}\right\},\sigma>0 \qquad (13)$$

其参数主要是中心向量和核宽度参数σ，中心向量即为所求支持向量，由算法自动选定。

4　仿真及预测结果分析

利用 steveGulm 开发的 SVM 工具箱，调用工具箱中的相关函数，在仿真软件 MATLAB 下编程实现支持向量机回归算法，从而得到相应的非线性关系。若训练不收敛或达不到预期目标时，不断调整上述设置的系数，直到训练收敛且精度满足要求。

核宽度参数σ的数值需要提前设定，核宽度参数过小，有可能发生过拟合，如果核宽度参数过大则有可能发生欠拟合，核宽度系数会直接影响模型的回归性能，如图 2 所示。惩罚因子 C 越大对数据的拟合程度越高，需要的训练时间越长，为了降低模型的复杂度，通常希望 C 小一点，但这又会增大模型的误差，所以 C 不能取值太小，这里 C 取值为 100。图 3 为σ=0.5 时 C 对预测误差的影响[5]。

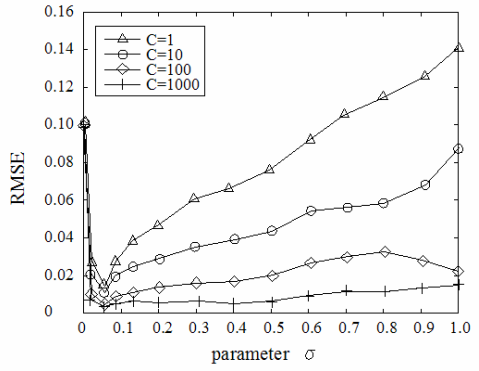

图 2　预测误差随变化情况

模型的输入向量为 4 维，选取第 3 6 10 组数据为测试样本，其余 6 组数据位训练样本。同时我们训练 BP 神经网络进行对比学习[6]。神经网络方法：影响 BP 神经网络性能的两个重要因素是隐含层神经元的个数和训练次数。采用 3 层网络 BP 神经网络训练，输入层神经元数设为 4 个；中间隐含层的神经元个数为（2*4+1）个，而 LS-SVR 值需要σ和 C 两个参数。为研究二者对模型性能的影响，给出了训练次数为 100 时，预测误差的影响，并与 LS-SVR 的相应结果进行对比。

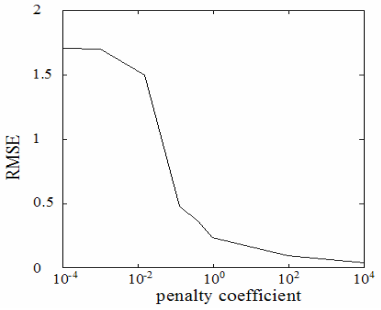

图 3　惩罚因子对预测误差的影响

表 4　预测结果比较

	BP	LS-SVM
均方根误差（RMSE）	0.6471	0.0219
训练时间（T/s）	1.2976	0.0358

可以看出运用 BP 神经网络也能对煤与瓦斯突出很好的预测，但观察表 4 中的数据发现，LS-SVR 训练时间远小于 BP 神经网络，误差也很小。所以选用 BP 神经网络的试验结果不如 LS-SVR 的好。因为神经网络是基于经验风险最小化和样本趋于无穷时的渐进理论在训练中存在不收敛和过拟合的问题，它需要有足够多的学习样本和先验知识，否则会导致模型泛化能力降低，性能下降；而 LS-SVR 是基于统计学习理论的算法，对具有少样本(支持向量)的煤与瓦斯突出具有很好的适应性、预测更加准确。因此在本实验中 LS-SVR 优于 BP 神经网络模型。

利用上述方式我们可以对需要预测的数据进行 RS-LS-SVR 预测出结果为：

表 5　预测结果输出

3	6	10
0.0020	1.0000	2.0017
无突出危险	危险	突出

5　结语

为了提高煤与瓦斯突出预测系统检测精度，本文利用粗糙集理论和 LS-SVR 算法构建对瓦斯突出危害

（下转第 35 页）

Research Board,2006,(1944):35–40.
5 吴文祥,黄海军.平行路径网络中信息对交通行为的影响研究.管理科学学报,2003,6(2):12–16.
6 干宏程,叶昕.行程时间波动性对路径选择影响的离散选择分析.交通运输系统工程与信息,2010,10(1):140–144.
7 曾松,史春华,杨晓光.基本实验分析的驾驶员路线选择模式研究.公路交通科技,2002,19(4):85–88.
8 熊轶,黄海军,李志纯.交通信息系统作用下的随机用户均衡模型与演进.交通运输系统工程与信息,2003,3(3):44–48.
9 Richards AC, Mcdonald M. Questionnaire surveys to evaluate user response to variable message signs in an urban network. Intelligent Transport Systems, 2007,1(3):177–185.
10 Erke A, Sagberg F, Hagman R. Effects of route guidance variable message signs (VMS) on driver behaviour. Transp–ortation Research Part F, 2007,10(3):447–457.
11 杨晓光,伍速锋,云美萍.日常出行中的交通信息有效性仿真研究.计算机工程与应用,2007,43(4):12–15.
12 秦进,黎新华.交通信息的有效性研究.公路交通科技,2005,22(2):104–107.
13 黄海军.城市交通网络平衡分析理论与实践.北京:人民交通出版社,1994.180–190.
14 石小法,王炜,李文权.交通信息对交通网络的影响研究.系统工程学报,2001,16(3):167–171.
15 魏赟,范炳全,韩印,干宏程.交通诱导信息对路网中车辆行为的影响.交通运输工程学报,2009,9(6):114–126.
16 杨珍珍,干宏程.面向大型社会活动的快速路网控制策略仿真评价方法.计算机应用研究,2010,27(12):4473–4475.
17 干宏程,汪晴,范炳全.基于宏观交通流模型的行程时间预测.上海理工大学学报,2008,30(5):409–413.
18 Messmer A, Papageorgiou M. METANET: A macroscopic simulation program for motorway networks. Traffic Eng–ineering and Control, 1990,31(8/9):466–470,549.
19 Kotsialos A, Papageorgiou M, Mangeas M, Haj-Salem H. Coordinated and integrated control of motorway networks via non-linear optimal control. Transportation Research Part C, 2002,10(1):65–84.

(上接第 68 页)

的预报体系。降低了特征选择方法的计算复杂性，提出一种基于粗糙集约简的最小二乘支持向量机算法。借助于粗糙集对数据集的简约能力，在保证样本集分辨能力的情况下简化了样本空间,从而加快了 LS-SVR 的训练速度和改善其泛化能力,通过 LS-SVR 训练与 BP 神经网络 MATLAB 仿真结果训练研究,得出在处理小样本非线性系统建模问题时，LS-SVR 表现出更高的精度和更快的速度,使煤与瓦斯预测结果更能够满足现实的需求。

参考文献

1 曾黄麟.粗集理论及其应用.重庆:重庆大学出版社,1996.
2 师旭超,韩阳.煤与瓦斯突出预测的支持向量机(SVM)模型.中国安全科学学报,2009,19(7):26–30.
3 高隽.智能信息处理方法导论.北京:机械工业出版社,2004.
4 Wu XH, Liu J, Liang YC, et al. Application of Support Vector Machine in Transformer Fault Diagnosis. Journal of Xi'an Jiaotong University, 2007,41(6):722–726.
5 邓乃扬,田英杰.数据挖掘中的最优化方法—支持向量机.北京:科学出版社,2004.
6 张德丰.MATLAB 神经网络应用设计.北京:机械工业出版社,2009.