(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 114330485 A
(43)申请公布日 2022.04.12

(21)申请号 202111358222.7

(22)申请日 2021.11.16

(71)申请人 国网冀北电力有限公司经济技术研究院
地址 100038 北京市海淀区羊坊店东路21号院1号楼7层701室
申请人 国家电网有限公司

(72)发明人 张晓曼 程序 李红建 耿鹏云 陈太平 安磊 齐霞 张妍 刘宣 路妍 董海鹏 曾凡梅 相静 张萌萌 谢品杰

(74)专利代理机构 上海科律专利代理事务所(特殊普通合伙) 31290
代理人 金碎平

(51)Int.Cl.
G06K 9/62(2022.01)
G06N 3/12(2006.01)
G06Q 10/04(2012.01)
G06Q 40/06(2012.01)
G06Q 50/06(2012.01)

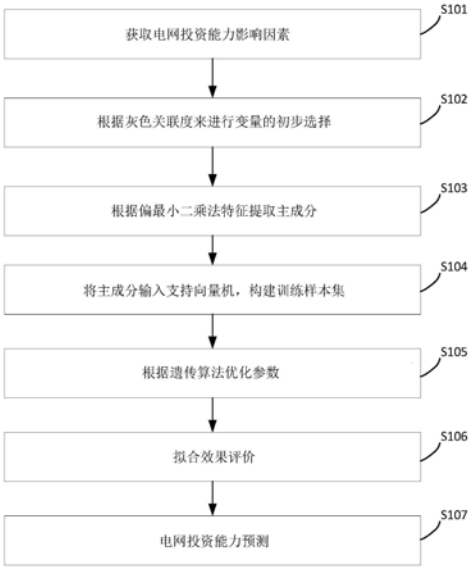
权利要求书2页 说明书10页 附图2页

(54)发明名称

基于PLS-SVM-GA算法的电网投资能力预测方法

(57)摘要

本发明公开了一种基于PLS-SVM-GA算法的电网投资能力预测方法,包括步骤:S101、确定电网企业投资能力的初始影响因素;S102、根据灰色关联度分析进行变量的初步选择;S103、利用偏最小二乘法中的主成分分析对初始影响因素进行提取;S104、将提取出的成分通过支持向量机模型构建训练样本集;S105、利用遗传算法对支持向量机的参数进行优化;S106、对电网投资能力的拟合效果进行评价;S107、利用优化后的支持向量机对电网投资能力进行预测并输出评价指标。本发明兼顾支持向量机和遗传算法的优点,能够更好地考虑非线性因素影响,使模型具有较好的鲁棒性和预测稳定性,从而大大提高预测结果的准确性。



CN 114330485 A

1.一种基于PLS-SVM-GA算法的电网投资能力预测方法,其特征在于,包括以下步骤:

步骤S101、确定电网企业投资能力的初始影响因素;

步骤S102、根据灰色关联度分析进行变量的初步选择;

步骤S103、利用偏最小二乘法中的主成分分析对初始影响因素进行提取;

步骤S104、将提取出的成分通过支持向量机模型构建训练样本集;

步骤S105、利用遗传算法对支持向量机的参数进行优化;

步骤S106、对电网投资能力的拟合效果进行评价;

步骤S107、利用优化后的支持向量机对电网投资能力进行预测并输出评价指标。

2.如权利要求1所述的基于PLS-SVM-GA算法的电网投资能力预测方法,其特征在于,所述步骤S101中初始影响因素包括:主营业务成本、资产负债率、净资产收益率、单位资产售电量、运行维护费、线损率、售电量、电力行业景气指数、全年高峰负荷、销售电价、GDP、固定资产投资额、城市化率、第二产业占比、能源消费强度、碳排放强度、经济发展目标和贷款利率。

3.如权利要求1所述的基于PLS-SVM-GA算法的电网投资能力预测方法,其特征在于,所述步骤S102对步骤S101中的初始影响因素分别进行灰色绝对关联度、灰色相对关联度分析,进而计算灰色综合关联度,并选择灰色综合关联度在0.5以上的影响因素进行下一步分析。

4.如权利要求3所述的基于PLS-SVM-GA算法的电网投资能力预测方法,其特征在于,所述步骤S102中灰色综合关联度ρ_{0i}的计算公式如下:

$$\rho_{0i}=\theta\varepsilon_{0i}+(1-\theta)r_{0i}$$

其中,$\theta\in[0,1]$,取$\theta=0.5$,表示对绝对量之间的关系和变化速率同等重视,r_{0i}为灰色相对关联度,ε_{0i}为灰色绝对关联度。

5.如权利要求1所述的基于PLS-SVM-GA算法的电网投资能力预测方法,其特征在于,所述步骤S103包括:

对数据进行标准化处理,生成标准化矩阵E_0和F_0,x_{ij}^*表示x_{ij}标准化后的数值;y_i^*表示y_i标准化后的数值;x_{ij}表示解释变量矩阵X中第j个变量x_j的第i个样本点;y_i表示因变量y的第i个样本值;

通过交叉有效性原则来确定PLS回归中成分的提取个数;

依次提取第一个主成分t_1、第二成分t_2、…和第h成分t_h,在确定h后停止迭代,其中h小于X的秩。

6.如权利要求5所述的基于PLS-SVM-GA算法的电网投资能力预测方法,其特征在于,所述PLS回归中成分的提取个数确定过程如下:

记y_i为原始数据,$t_1,t_2,\cdots,t_m$是在PLS回归过程中提取的成分,$\hat{y}_{hi}$为使用全部样本点并取成分$t_1,t_2,\cdots,t_h$回归建模所得的第i个样本点的拟合值,而$\hat{y}_{h(-i)}$是在回归时删去样本点i,再利用成分$t_1,t_2,\cdots,t_h$回归所得y_i的拟合值;

$$S_h=\sum_{i=1}^{n}(y_i-\hat{y}_{hi})^2$$

$$P_h=\sum_{i=1}^{n}(y_i-\hat{y}_{h(-i)})^2$$

则交叉有效性的定义为：

$$Q_h^2=1-\frac{P_h}{S_{h-1}}$$

当$Q_h^2<0.0975$时，停止增加新的成分t_h。

7.如权利要求1所述的基于PLS-SVM-GA算法的电网投资能力预测方法，其特征在于，所述步骤S104包括：

将步骤S103提取的主成分$t_1,t_2,\cdots,t_h$作为支持向量机的样本输入，$u_1=f(t)$作为支持向量机的输出；

构建训练集样本$\{(t_i,y_i),i=1,2,\cdots h\}$；其中$t_i(t_i\in R^d)$是第i个训练样本的输入列向量，$y_i\in R$为对应的输出值，建立如下回归函数：

$f(t)=w\Phi(t)+b$

其中$\Phi(t)$为将数据映射到高维特征空间的非线性映射函数；w为特征权向量；$b\in R$为阈值。

8.如权利要求1所述的基于PLS-SVM-GA算法的电网投资能力预测方法，其特征在于，所述步骤S105通过控制误差ε的取值对偏最小二乘支持向量回归模型中的参数集采用遗传算法进行近似寻优，将训练样本的均方误差MSE作为遗传算法的适应度函数，通过选择、交叉和变异操作来判断当前是否满足目标精度要求，若满足条件则通过解码输出SVM模型的最优参组合，否则重新用遗传算法进行计算。

9.如权利要求1所述的基于PLS-SVM-GA算法的电网投资能力预测方法，其特征在于，所述步骤S105包括：

步骤1：确定惩罚因子c和核参数σ的大致范围，对c、σ进行二进制编码，生成初始种群；

步骤2：构造适应度函数作为遗传算法与SVM的接口，将训练样本的均方误差MSE作为遗传算法的适应度函数，通过判断适应度函数的大小来决定是否终止参数寻优；

步骤3：设定种群规模、终止进化代数、交叉概率和变异概率；

步骤4：应用遗传算子选择、交叉、变异运算来产生下一代种群，然后转到步骤2来判断适应度值大小。

10.如权利要求1所述的基于PLS-SVM-GA算法的电网投资能力预测方法，其特征在于，所述步骤S106根据所得的预测模型，输入测试样本进行预测，并对输出数据反归一化处理；根据预测值和实际值的比较，采用评价预测模型的平均绝对误差MAE、平均相对误差MPE、均方根误差RMSE和Theil不等系数对所建预测模型进行评价。

基于PLS-SVM-GA算法的电网投资能力预测方法

技术领域

[0001] 本发明涉及一种电网投资能力预测方法,尤其涉及一种基于PLS-SVM-GA算法的电网投资能力预测方法。

背景技术

[0002] 在改革和规范电网企业运营模式的新电改背景下,对电网企业投资能力的研究越来越迫切。合理、客观地把握企业的投资能力是企业管理策略研究的核心内容之一。并且,随着电力市场化和电网企业体制改革的逐步深化,电网企业的经济效益因素在投资决策中所占的比重日益增大。因此,为满足电网企业对资金项目计划和预算管理的需要,有必要采取科学的方法对电网投资能力进行客观的预测,从而对电网公司的投资能力预测提供一定的决策支持。

[0003] 大多数学者采用主成分分析法提取影响因素,但是在影响因素选择上,电网投资涉及的影响因素众多,包括经营状况、管理水平、市场贡献、经济环境和政策环境等多方面指标的多层次、多维度的影响,已有文献有的仅从电网企业投资的内部因素或者外部因素进行分析,有的从内外部都分析了投资因素,均不够全面、客观。

[0004] 在预测方法选取上,应用较为广泛的方法包括计量回归方法以及综合评价方法。计量回归方法主要集中在多元回归以及协整分析,但是传统回归方法的重要缺点是没有考虑影响因素对电网投资规模的非线性影响,并且大多数文献在使用计量模型预测电网投资规模的时候忽略了计量模型的内生性问题。另外,许多学者利用综合评价方法建立电网投资影响因素体系,并在此基础上分析预测电网企业的投资能力。然而,综合评价方法在指标重要程度的判定上有很大的主观性,并且电网企业的投资的因素影响通常是非线性的,从而大大影响了预测的准确性。

发明内容

[0005] 本发明所要解决的技术问题是提供一种基于PLS-SVM-GA算法的电网投资能力预测方法,能够更好地考虑非线性因素影响,使模型具有较好的鲁棒性和预测稳定性,从而大大提高预测结果的准确性。

[0006] 本发明为解决上述技术问题而采用的技术方案是提供一种基于PLS-SVM-GA算法的电网投资能力预测方法,包括以下步骤:步骤S101、确定电网企业投资能力的初始影响因素;步骤S102、根据灰色关联度分析进行变量的初步选择;步骤S103、利用偏最小二乘法中的主成分分析对初始影响因素进行提取;步骤S104、将提取出的成分通过支持向量机模型构建训练样本集;步骤S105、利用遗传算法对支持向量机的参数进行优化;步骤S106、对电网投资能力的拟合效果进行评价;步骤S107、利用优化后的支持向量机对电网投资能力进行预测并输出评价指标。

[0007] 进一步地,所述步骤S101中初始影响因素包括:主营业务成本、资产负债率、净资产收益率、单位资产售电量、运行维护费、线损率、售电量、电力行业景气指数、全年高峰负

荷、销售电价、GDP、固定资产投资额、城市化率、第二产业占比、能源消费强度、碳排放强度、经济发展目标和贷款利率。

[0008] 进一步地，所述步骤S102对步骤S101中的初始影响因素分别进行灰色绝对关联度、灰色相对关联度分析，进而计算灰色综合关联度，并选择灰色综合关联度在0.5以上的影响因素进行下一步分析。

[0009] 进一步地，所述步骤S102中灰色综合关联度ρ_{0i}的计算公式如下：

[0010] $\rho_{0i}=\theta\varepsilon_{0i}+(1-\theta)r_{0i}$

[0011] 其中，$\theta\in[0,1]$，取$\theta=0.5$，表示对绝对量之间的关系和变化速率同等重视，r_{0i}为灰色相对关联度，ε_{0i}为灰色绝对关联度。

[0012] 进一步地，所述步骤S103包括：对数据进行标准化处理，生成标准化矩阵E_0和F_0，x_{ij}^*表示x_{ij}标准化后的数值；y_i^*表示y_i标准化后的数值；x_{ij}表示解释变量矩阵X中第j个变量x_j的第i个样本点；y_i表示因变量y的第i个样本值；通过交叉有效性原则来确定PLS回归中成分的提取个数；依次提取第一个主成分t_1、第二成分t_2、…和第h成分t_h，在确定h后停止迭代，其中h小于X的秩。

[0013] 进一步地，所述PLS回归中成分的提取个数确定过程如下：记y_i为原始数据，t_1，t_2，…，t_m是在PLS回归过程中提取的成分，$\hat{y}_{hi}$为使用全部样本点并取成分t_1，t_2，…，t_h回归建模所得的第i个样本点的拟合值，而$\hat{y}_{h(-i)}$是在回归时删去样本点i，再利用成分t_1，t_2，…，t_h回归所得y_i的拟合值；

[0014] $$S_h=\sum_{i=1}^{n}(y_i-\hat{y}_{hi})^2$$

[0015] $$P_h=\sum_{i=1}^{n}(y_i-\hat{y}_{h(-i)})^2$$

[0016] 则交叉有效性的定义为：

[0017] $$Q_h^2=1-\frac{P_h}{S_{h-1}}$$

[0018] 当$Q_h^2<0.0975$时，停止增加新的成分t_h。

[0019] 进一步地，所述步骤S104包括：将步骤S103提取的主成分t_1，t_2，…，t_h作为支持向量机的样本输入，$u_1=f(t)$作为支持向量机的输出；构建训练集样本$\{(t_i,y_i),i=1,2,\cdots h\}$；其中$t_i$($t_i\in R^d$)是第i个训练样本的输入列向量，$y_i\in R$为对应的输出值，建立如下回归函数：

[0020] $f(t)=w\Phi(t)+b$

[0021] 其中$\Phi(t)$为将数据映射到高维特征空间的非线性映射函数；w为特征权向量；$b\in R$为阈值。

[0022] 进一步地，所述步骤S105通过控制误差ε的取值对偏最小二乘支持向量回归模型中的参数集采用遗传算法进行近似寻优，将训练样本的均方误差MSE作为遗传算法的适应度函数，通过选择、交叉和变异操作来判断当前是否满足目标精度要求，若满足条件则通过

解码输出SVM模型的最优参组合，否则重新用遗传算法进行计算。

[0023] 进一步地，所述步骤S105包括：步骤1：确定惩罚因子c和核参数σ的大致范围，对c、σ进行二进制编码，生成初始种群；步骤2：构造适应度函数作为遗传算法与SVM的接口，将训练样本的均方误差MSE作为遗传算法的适应度函数，通过判断适应度函数的大小来决定是否终止参数寻优；步骤3：设定种群规模、终止进化代数、交叉概率和变异概率；步骤4：应用遗传算子选择、交叉、变异运算来产生下一代种群，然后转到步骤2来判断适应度值大小。

[0024] 进一步地，所述步骤S106根据所得的预测模型，输入测试样本进行预测，并对输出数据反归一化处理；根据预测值和实际值的比较，采用评价预测模型的平均绝对误差MAE、平均相对误差MPE、均方根误差RMSE和Theil不等系数对所建预测模型进行评价。

[0025] 本发明提供的电网投资能力预测方法，可以从行业内外部角度进行分析，并利用GA-PLS-SVM模型对各指标及电网投资能力进行测算。本发明对比现有技术有如下的有益效果：1、本发明通过灰色关联度进行变量的初步选择，综合考虑内部指标以及外部指标共18个影响因素，提高预测结果的准确性。2、在成分提取上，大多数学者采用主成分分析法提取影响因素，本发明利用PLS(Partial least squares regression，偏最小二乘法回归))方法对初始影响因素进行提取，PLS方法是多元线性回归、典型相关分析和主成分分析的集成和发展。它与主成分分析法(PCA)的主要区别是它在特征提取过程中采用了信息综合和筛选技术。它所提取的成分既能很好地概括自变量系统中的信息，又能最好地解释因变量，同时又消除了系统中的噪声干扰。因而，PLS不仅能够完成类似PCA的降维工作，而且有效地解决了自变量间多重相关性情况下的回归建模问题。3、作为预测方法的一种，支持向量机(Support Vector Machines，SVM))可以更好地考虑了因素的非线性影响，本发明把利用PLS方法提取的主成分作为输入变量用于SVM回归建模，使模型具有较好的鲁棒性和预测稳定性。

附图说明

[0026] 图1为本发明电网投资能力预测流程示意图；

[0027] 图2为本发明电网投资能力GA和SVM相结合的流程示意图。

具体实施方式

[0028] 下面结合附图和实施例对本发明作进一步的描述。

[0029] 请参见图1，本发明提供的基于PLS-SVM-GA算法的电网投资能力预测方法，包括：获取电网投资能力影响因素；根据灰色关联度来进行变量的初步选择；根据偏最小二乘法特征提取主成分；将主成分输入支持向量机，构建训练样本集；根据遗传算法优化参数；拟合评价效果；电网投资能力预测。本发明可以从行业内外部角度进行分析，并利用GA-PLS-SVM模型对各指标及电网投资能力进行测算。

[0030] 为了得到电网投资能力预测值，选择灰色综合关联度在0.5以上的因素进行下一步分析。本发明根据偏最小二乘法特征提取主成分之后，兼顾支持向量机和遗传算法(Genetic Algorithm，GA)两种智能算法的优点，先将提取的主成分输入支持向量机，构建训练样本集，再通过控制误差ε的取值对偏最小二乘支持向量回归模型中的参数集采用遗传算法进行近似寻优，之后采用偏最小二乘支持向量回归对电网投资能力进行预测，从而

构建一套GA－SVM的预测模型体系进行回归预测。

[0031] 本发明建立的GA-SVM预测模型的预测步骤如图2所示：

[0032] (1)输入电网投资能力影响因素。

[0033] (2)采用灰色综合关联度来进行变量的初步筛选。

[0034] (3)根据偏最小二乘法特征提取主成分。

[0035] (4)设置SVM模型参数寻优区间以及遗传算法的初始化和参数设置。

[0036] (5)利用GA算法对SVM的模型参数进行寻优。将训练样本的均方误差MSE作为GA的适应度函数，通过选择、交叉和变异等遗传操作来判断当前是否满足目标精度要求，若满足条件则通过解码输出SVM模型的最优参数γ和σ^2组合，否则重新用遗传算法进行计算。

[0037] (6)建立参数优化后的SVM预测模型。根据步骤(3)得到的最佳参数γ和σ^2，利用训练样本，训练SVM预测模型，最后根据所得的预测模型，输入测试样本进行预测，并对输出数据反归一化处理。

[0038] (7)根据预测值和实际值的比较，采用评价预测模型的性能指标MAE、MPE、RMSE和Theil不等系数对所建预测模型进行评价。

[0039] (8)最后利用训练好的GA-SVM进行预测，预测得到未来某一时刻的电网投资能力指标数值，再进行投资能力的相关分析。

[0040] 下面对本发明的每个步骤一一展开说明。

[0041] 1、获取电网投资能力影响因素

[0042] 面对错综复杂的经济和社会环境，不仅需要从行业内部的角度来研究电网企业投资能力，还需要从宏观环境的角度审视经济发展对投资能力的影响，因此，需要分别从行业内部、外部来选取影响电网投资能力的指标，同时基于相关性、全面性、代表以及数据的可得性等原则，构建了如表1所示电网投资能力影响因素指标体系。

[0043] 表1电网投资能力影响因素

[0044]

内部指标	
经营状况	**管理水平**
主营业务成本	单位资产售电量
资产负债率	运行维护费
净资产收益率	线损率

外部指标		
市场环境	**经济环境**	**政策环境**
售电量	GDP	能源消费强度
电力行业景气指数	固定资产投资额	碳排放强度
全年高峰负荷	城市化率	经济发展目标
销售电价	第二产业占比	贷款利率

[0045] 2、根据灰色综合关联度来进行变量的筛选

[0046] 本发明对上文确定的18个指标与电网投资能力分别进行灰色绝对关联度、灰色相对关联度分析，进而计算灰色综合关联度，选择灰色综合关联度在0.5以上的因素进而下一步分析。

[0047] 灰色绝对关联度、灰色相对关联度和灰色综合关联度基本原理和计算方法如下：

[0048] (1)灰色绝对关联度

[0049] 设

[0050] $X_i=(x_i(1),x_i(2),\cdots,x_i(n))\ (i=1,2,\cdots,m)$

[0051] 记折线

[0052] $(x_i(1)-(x_i(1),x_i(2)-(x_i(1),\cdots,x_i(n)-(x_i(1))$ 为 X_i^0。令

[0053] $$s_i=\int_1^n X_i^0\,dt$$

[0054] $$|s_i-s_0|=\int_1^n (X_i^0-X_0^0)dt,\quad (i=0,1,2,\cdots,m)$$

[0055] 则，灰色绝对关联度为

[0056] $$\varepsilon_{0i}=\frac{1+|s_0|+|s_i|}{1+|s_0|+|s_i|+|s_i-s_0|}$$

[0057] 灰色绝对关联度ε_{0i}表征了折线X_0与X_i的绝对增量间的关系，用两条序列折线间所夹的面积大小来衡量两序列的关联性的大小，折线X_0与X_i的绝对增量越相似，ε_{0i}越大，反之就越小。

[0058] (2)灰色相对关联度

[0059] 设

[0060] $X_i=(x_i(1),x_i(2),\cdots,x_i(n))\ (i=1,2,\cdots,m)$

[0061] 则，灰色相关关联度为

[0062] $$r_{0i}=\frac{1+|s_0'|+|s_i'|}{1+|s_0'|+|s_i'|+|s_i'-s_0'|}$$

[0063] 其中，$s_i'=\int_1^n (X_i'-x_i(1)')dt$，$X'_i=X_i/x_i(1)$，$i=0,1,2,\cdots,m$

[0064] 灰色相对关联度r_{0i}表征了序列X_0与X_i相对于始点的变化速率之间的关系，X_0与X_i的变化速率越接近，r_{0i}越大，反之就越小。

[0065] (3)灰色综合关联度

[0066] 灰色综合关联度ρ_{0i}既体现了折线X_0与X_i的相似程度，又反映出X_0与X_i相对于始点的变化速率的接近程度，是较为全面地表征序列之间是否紧密的一个数量指标。其计算公式如下：

[0067] $\rho_{0i}=\theta\varepsilon_{0i}+(1-\theta)r_{0i}$

[0068] 其中，$\theta\in[0,1]$。为表示对绝对量之间的关系和变化速率同等重视，取$\theta=0.5$。

[0069] 3、根据偏最小二乘法(PLS,Partial Least-Squares Regression)提取主成分

[0070] 对于上述根据灰色综合关联度筛选出来电网投资能力影响因素，进一步利用PLS

回归分析相比于主成分分析，可以有监督地提取主成分，所以可以说PLS回归分析是主成分、典型相关分析及多元线性回归分析的有机结合，其具体步骤如下：

[0071] (1)数据标准化处理

[0072] 其目的在于使样本点的集合重心和原点重合，减少运算误差。

[0073]
$$\begin{cases} x_{ij}^{*}=\dfrac{x_{ij}-\bar{x}_j}{s_j} \\ y_i^{*}=\dfrac{y_i-\bar{y}}{s_y} \\ \mathbf{E}_0=(x_{ij}^{*})_{n\times k} \\ \mathbf{F}_0=(y_i^{*})_{n\times 1} \end{cases} \qquad (i=1,2,\cdots,n;\ j=1,2,\cdots k) \tag{3-1}$$

[0074] 其中，x_{ij}表示解释变量矩阵X中第j个变量x_j的第i个样本点；$\bar{x}_j$表示解释变量矩阵X中第j个变量x_j的均值；s_j表示x_j的标准差；y_i表示因变量y的第i个样本值；$\bar{y}$表示y的均值；s_y表示y的标准差；x_{ij}^{*}表示x_{ij}标准化后的数值；y_i^{*}表示y_i标准化后的数值。

[0075] (2)第一成分t_1的提取

[0076] 对于标准化矩阵E_0和F_0，从E_0中提取第一个主成分$t_1=E_0w_1$，其中w_1为E_0的第一主轴，即$||w_1||=1$。

[0077] 则有

[0078]
$$w_1=\frac{E_0^{T}F_0}{\left\|E_0^{T}F_0\right\|}=\frac{1}{\sqrt{\sum_{j=1}^{k}r^2(x_j,y)}}\begin{bmatrix} r(x_1,y) \\ \cdots \\ r(x_k,y) \end{bmatrix} \tag{3-2}$$

[0079]
$$t_1=E_0w_1=\frac{1}{\sqrt{\sum_{j=1}^{k}r^2(x_j,y)}}[r(x_1,y)E_{01}+r(x_2,y)E_{02}+\cdots+r(x_k,y)E_{0k}] \tag{3-3}$$

[0080] 其中，E_{0i}(i=1,2,…,k)表示E_0的第i列；r(x_j,y)(i=1,2,…,k)表示x_j和y的相关系数。求得轴w_1后，可得成分t_1。接下来，分别求E_0和F_0对t_1的回归方程

[0081] $E_0=t_1P_1^{T}+E_1$，$F_0=t_1r_1+F_1$ (3-4)

[0082] 其中，$P_1=E_0^{T}t_1/\left\|t_1\right\|^2$为$E_0$对$t_1$的回归系数；$r_1=F_0^{T}t_1/\left\|t_1\right\|^2$为$F_0$对$t_1$的回归系数。并由此可以求得回归方程(4-32)的残差矩阵为：

[0083] $E_1=E_0-t_1P_1^{T}$，$F_1=F_0-t_1r_1$ (3-5)

[0084] (3)第二成分t_2的提取

[0085] 以E_1取代E_0，F_1取代F_0，重复建模步骤(2)，可以求得第一主轴w_1和第二成分t_2，此时，注意到E_1不再是标准化矩阵，故有：

[0086] $$w_2=\frac{E_1^TF_0}{\left\|E_1^TF_0\right\|}=\frac{1}{\sqrt{\sum_{j=1}^{k}Cov^2(E_{1i},F_1)}}\begin{bmatrix}Cov(E_{11},F_1)\\ \cdots\\ Cov(E_{1k},F_1)\end{bmatrix}\quad(3\text{-}6)$$

[0087] $t_2=E_1w_2$ (3-7)

[0088] 其中，Cov(E_{1j},y)表示E_{1j}与y的协方差。然后再施行E_1、F_1对t_2的回归，有

[0089] $$E_1=t_2P_2^T+E_2,\qquad F_1=t_2r_{21}+F_2\quad(3\text{-}8)$$

[0090] 其中，$P_2=E_2^Tt_2/\left\|t_2\right\|^2$为$E_1$对$t_2$的回归系数；$r_2=F_1{}^Tt_2/||t_2||^2$为$F_1$对$t_2$的回归系数。

[0091] (4)第h成分t_h的提取

[0092] 重复上述步骤(2)(3)，可以求得第h成分t_h。PLS回归中成分的提取个数h可以使用交叉有效性来确定，在确定h后停止迭代，其中h小于X的秩。

[0093] (5)交叉有效性原则

[0094] 根据上述PLS回归建模步骤，可以知道，PLS回归方程并不需要选用全部的成分进行回归建模，对此，可以通过考察增加一个新的成分后，能否对模型的预测功能有明显的改进来考虑，即可以通过交叉有效性原则来确定PLS回归中成分的提取个数。

[0095] 记y_i为原始数据，$t_1,t_2,\cdots,t_m$是在PLS回归过程中提取的成分，$\hat{y}_{hi}$为使用全部样本点并取成分$t_1,t_2,\cdots,t_h$回归建模所得的第i个样本点的拟合值，而$\hat{y}_{h(-i)}$是在回归时删去样本点i，再利用成分$t_1,t_2,\cdots,t_h$回归所得y_i的拟合值。记：

[0096] $$S_h=\sum_{i=1}^{n}(y_i-\hat{y}_{hi})^2\quad(3\text{-}9)$$

[0097] $$P_h=\sum_{i=1}^{n}(y_i-\hat{y}_{h(-i)})^2\quad(3\text{-}10)$$

[0098] 则，交叉有效性的定义为：

[0099] $$Q_h^2=1-\frac{P_h}{S_{h-1}}\quad(3\text{-}11)$$

[0100] 所谓交叉有效性原则是指，当$Q_h^2<0.0975$时，停止增加新的成分t_h。

[0101] 4、将主成分输入支持向量机，构建训练样本集

[0102] 在利用PLS法从自变量X和因变量Y中分别提取主成分$t_1,t_2,\cdots,t_h$和u_1后(它们包含自变量与因变量的绝大部分信息)，将以上提取的主成分$t_1,t_2,\cdots,t_h$作为支持向量机的样本输入，$u_1=f(t)$作为支持向量机的输出。携带自变量X绝大部分信息的前h个主成分被提取出来，构成了支持向量机的输入空间，从而实现了输入空间由$R_n\rightarrow R_h$的变换，达到了特征提取和变量降维的目的，从而提高了模型运行的效率和预测的精度。

[0103] 将上面提取的主成分构建训练集样本{(t_i,y_i)，i=1,2,…h}(其中t_i($t_i\in R^d$)是第i个训练样本的输入列向量，$y_i\in R$为对应的输出值)，建立如下回归函数：

[0104] $f(t)=w\Phi(t)+b$ (4-1)

[0105] 其中Φ(t)为将数据映射到高维特征空间的非线性映射函数;w为特征权向量;b∈R为阈值。

[0106] 定义ε的线性不敏感损失函数:

[0107] $$L(f(t),y,\varepsilon)=\begin{cases}0,|y-f(t)|\le\varepsilon\\ |y-f(t)|-\varepsilon,|y-f(t)|>\varepsilon\end{cases} \quad (4\text{-}2)$$

[0108] 其中,f(t)为回归拟合函数的预测值;y对应的实际值,即表示若f(t)与y之间的差别小于等于ε,则损失等于0。

[0109] 引入松弛变量ξ_i,ξ_i^*,建立如下约束条件:

[0110] $$min\frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}(\xi_i+\xi_i^*) \quad (4\text{-}3)$$

[0111] s.t.y_i-wΦ(t_i)-b≤ε+ξ_i

[0112] $$-y_i+w\Phi(t_i)+b\le\varepsilon+\xi_i^*$$

[0113] $$\xi_i\ge0,\xi_i^*\ge0$$

[0114] i=1,2,…,l

[0115] 其中,C为惩罚因子。

[0116] 引入Largrange函数,并将其转化为对偶形式:

[0117] $$max[-\frac{1}{2}\sum_{i=1}^{h}\sum_{j=1}^{h}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)K(t_i,t_j)-\sum_{i=1}^{l}(\alpha_i+\alpha_i^*)\varepsilon+\sum_{j=1}^{h}(\alpha_i-\alpha_i^*)y_i] \quad (4\text{-}4)$$

[0118] $$s.t.\sum_{i=1}^{h}(\alpha_i-\alpha_i^*)$$

[0119] 0≤α_i≤C

[0120] $$0\le\alpha_i^*\le C$$

[0121] 其中,C为惩罚因子,K(t_i,t_j)=Φ(t_i)Φ(t_j)为核函数。

[0122] 计算此规划问题求得最优解α和α^*。

[0123] 利用KKT(Karush-Kuhn-Tucker)条件,即

[0124] $$b^*=y_i-\sum_{i=1}^{h}(\alpha_i-\alpha_i^*)K(t_i,t)-\varepsilon,\alpha_i\in(0,C)$$

[0125] 或者

[0126] $$b^*=y_i-\sum_{i=1}^{h}(\alpha_i-\alpha_i^*)K(t_i,t)+\varepsilon,\alpha_i\in(0,C)$$

[0127] 计算得到偏置量b^*。

[0128] 得到SVM回归预测的拟合函数为:

[0129] $$f(t)=w^*\Phi(t)+b^*=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)K(t_i,t)+b^* \quad (4\text{-}5)$$

[0130] 5、根据遗传算法优化参数

[0131] 遗传算法在非线性优化问题中表现良好，它对模型的连续性、是否线性、可微性没有严格的要求，也不受待优化参数个数的限制，通过自适应学习能够很快获得最优解。目前GA算法已经在神经网络、结构设计、机器学习、函数优化以及图像处理领域有广泛的应用。

[0132] 遗传算法不同于搜索算法、启发式、枚举等传统算法，它通常具有以下的特点：

[0133] a.将问题参数间接的抽象为参数编码集。

[0134] b.可以处理复杂的非结构化问题，具有智能性、灵活的组织性和适应性。不需要事先描述整个问题的特点。

[0135] c.具有较强的并行化，思想简单，实现步骤规范，易于将实际问题具体化。

[0136] 综上分析遗传算法具有很强的全局搜索能力，用遗传算法寻找有效的最小二乘支持向量机的参数是一种可行的方式。

[0137] 由SVM的算法过程知道，不敏感损失函数中的ε、惩罚因子C和径向基函数中的σ^2(也称径向核)这3个参数取值不同将得到不同的支持向量机模型，因此，本发明将通过控制误差ε的取值对参数集(C,σ^2)采用遗传算法进行近似寻优，从而构建GA－PLS－SVM模型进行回归预测。

[0138] (1)建立位串空间

[0139] 位串空间：$S^L=\{a_1,a_2,\ldots a_k\}$，$a_k=(a_{k1},a_{k2},\ldots,a_{kL})$，$a_{k1}\in\{0,1\}$将位串个体从位串空间转化成问题参数空间的译码函数$\Omega:\{0,1\}^L\rightarrow[u,v]$的公式定义如下：

[0140] $$x_k=\Omega(a_{k1},\ a_{k2},\ldots,\ a_{kL})=u+\frac{v-u}{2^L-1}\left(\sum_{j=1}^{L}a_{kj}2^{L-j}\right) \quad (5\text{-}1)$$

[0141] 可以利用二进制对p、q编码。本发明算法采用5位二进制码对p、q进行编码，编码长度L＝10，设前5位表示p，后5位表示q，由此构成候选解空间S，其大小为2^L。

[0142] (2)自适应交叉算子

[0143] 本发明引进一种新的自适应遗传算子，建立交叉算子和适应度函数f(x)的关系，从而使交叉概率Pc随着适应度的波动而灵活改变，使Pc满足时变性，提高算法的灵活度。交叉算子和适应度函数f(x)的关系如下：

[0144] $$Pc=\begin{cases}1-\dfrac{\sum_{i=1}^{N}(f_i-\overline{f})^2}{N\max(f_i-\overline{f})^2}, & T\le t\\ k_1, & T>t\end{cases} \quad (5\text{-}2)$$

[0145] (3)自适应变异算子

[0146] 本发明采用基本位变异对个体编码串以变异概率Pm随机指定某一位或某几位基因作变异运算。建立Pm与适应度函数f(x)的关系如下：

[0147] $$Pm=\begin{cases}k_2\left(1-\dfrac{\sum_{i=1}^{N}(f_i-\overline{f})^2}{N\max(f_i-\overline{f})^2}\right), & T\le t\\ k_3, & T>t\end{cases} \quad (5\text{-}3)$$

[0148] (4)GA优化PLS-SVM模型参数

[0149] 主要的实现步骤如下：

[0150] 步骤1：确定惩罚因子c和核参数σ的大致范围，对c、σ进行二进制编码，生成初始种群。

[0151] 步骤2：构造适应度函数，这是遗传算法与SVM的接口，通过判断适应度函数的大小来决定是否终止参数寻优。

[0152] 步骤3：对本算法的实际问题设定种群规模(如M＝20)、终止进化代数(T＝60)、交叉概率(Pc＝0.85)、变异概率(Pm＝0.001)。遗传算子中的选择运算是根据每个个体的适应度大小来确定的，本算法试验中适应度值小的个体将有大的概率被选择到下一代。

[0153] 步骤4：应用遗传算子选择、交叉、变异运算来产生下一代种群，然后转到步骤2来判断适应度值大小。

[0154] 6、预测效果评价

[0155] 利用以下指标评价模型的预测能力：

[0156] 平均绝对误差：$$\mathrm{MAE}=\frac{1}{N}\sum_{h=1}^{N}\left|y_h-\hat{y}_h\right| \quad (6\text{-}1)$$

[0157] 平均相对误差：$$\mathrm{MPE}=\frac{1}{N}\sum_{h=1}^{N}\left|\frac{y_h-\hat{y}_h}{y_k}\right| \quad (6\text{-}2)$$

[0158] 均方根误差：$$RMSE=\sqrt{\frac{1}{N}\sum_{h=1}^{N}(\hat{y}_t-y_t)^2} \quad (6\text{-}3)$$

[0159] Theil不等系数：$$U=\sqrt{\frac{1}{N}\sum_{h=1}^{N}(\hat{y}_t-y_t)^2}\Big/\sqrt{\frac{1}{N}\sum_{h=1}^{N}\hat{y}_t^{\,2}}+\sqrt{\frac{1}{N}\sum_{h=1}^{N}y_t^{\,2}} \quad (6\text{-}4)$$

[0160] 其中，y_k为实际值，$\hat{y}_k$则为预测值，N为时间序列的长度。其中，MAE和RMSE这两个统计量受因变量量纲影响，可以用来比较不同模型的预测效果，其值越小，表明相应模型的预测效果越好。其余两个统计量与因变量量纲无关的相对指标，MPE反映了相对误差的大小，其值亦是越小越好；Theil不等系数U通常介于0～1之间，其中当其值为0时，表示拟合程度达至100％。

[0161] 虽然本发明已以较佳实施例揭示如上，然其并非用以限定本发明，任何本领域技术人员，在不脱离本发明的精神和范围内，当可作些许的修改和完善，因此本发明的保护范围当以权利要求书所界定的为准。

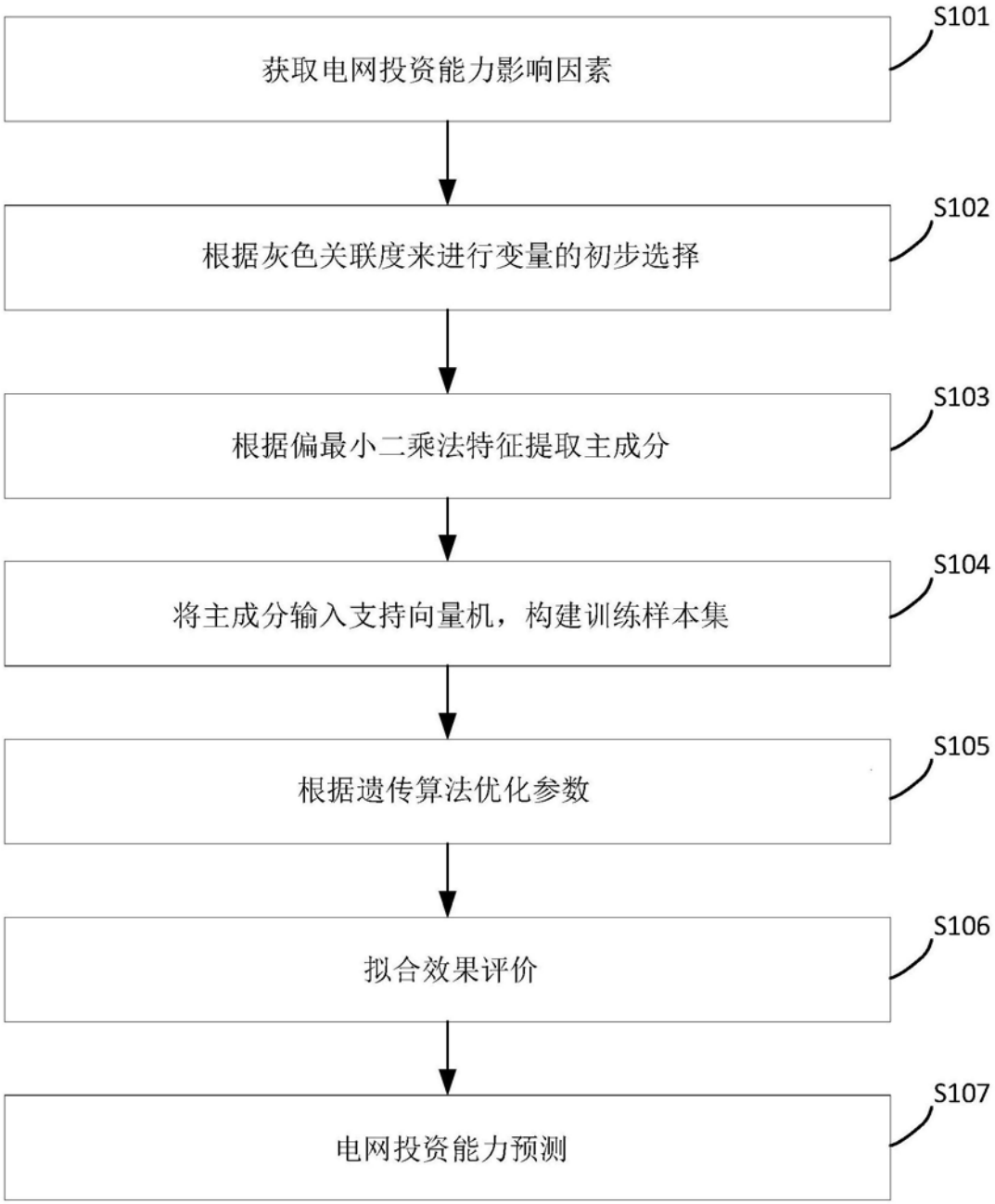

S101
获取电网投资能力影响因素
S102
根据灰色关联度来进行变量的初步选择
S103
根据偏最小二乘法特征提取主成分
S104
将主成分输入支持向量机，构建训练样本集
S105
根据遗传算法优化参数
S106
拟合效果评价
S107
电网投资能力预测

图1

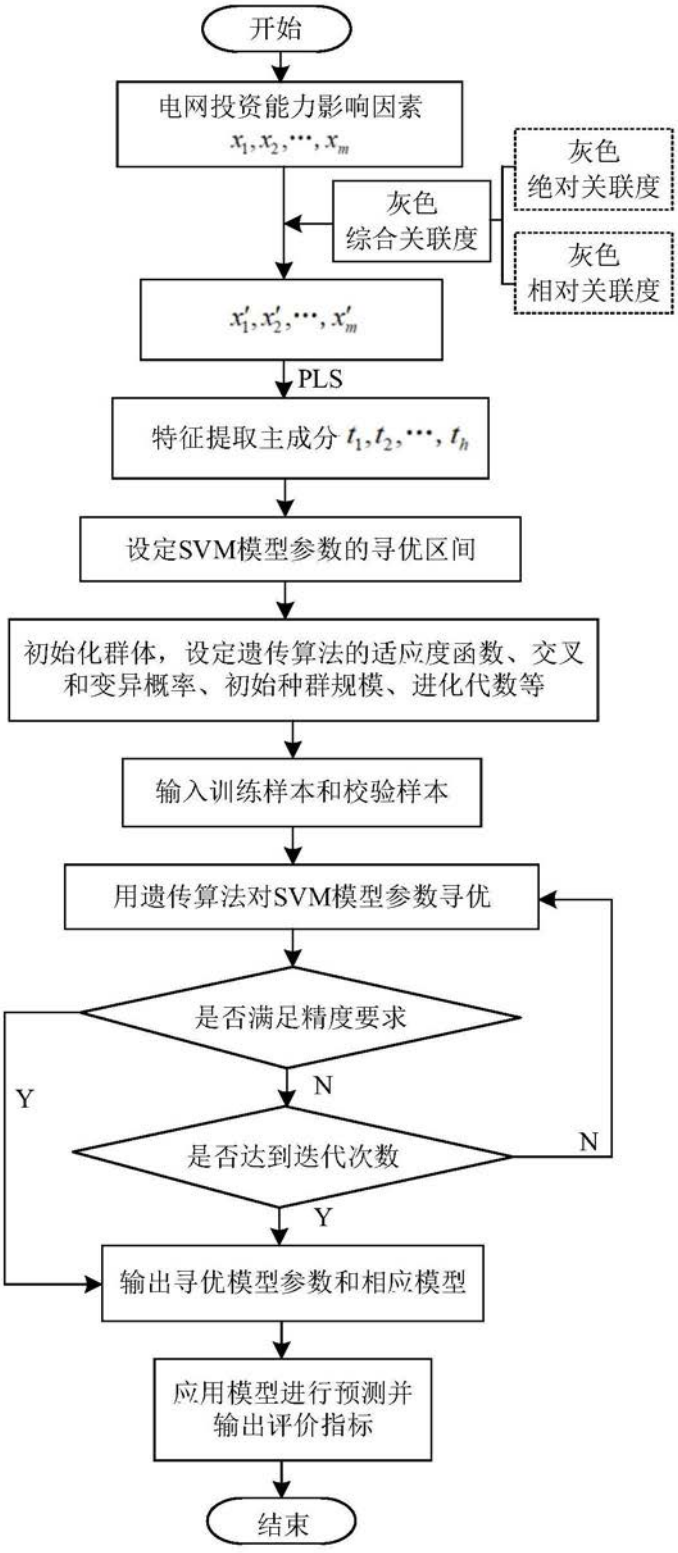

开始
电网投资能力影响因素
$x_1, x_2, \cdots, x_m$
灰色
绝对关联度
灰色
综合关联度
灰色
相对关联度
$x'_1, x'_2, \cdots, x'_m$
PLS
特征提取主成分 $t_1, t_2, \cdots, t_h$
设定SVM模型参数的寻优区间
初始化群体，设定遗传算法的适应度函数、交叉和变异概率、初始种群规模、进化代数等
输入训练样本和校验样本
用遗传算法对SVM模型参数寻优
是否满足精度要求
Y
N
是否达到迭代次数
N
Y
输出寻优模型参数和相应模型
应用模型进行预测并输出评价指标
结束

图2