基于非线性特征提取和加权 K 最邻近元回归的预测模型*

唐　黎[1]，潘和平[2]，姚一永[1]
（1.西南财经大学天府学院 智能金融学院，四川 成都 610052）
（2.成都大学 商学院，四川 成都 610106）

摘　要　本文提出了一种智能的金融时间序列预测模型。该模型采用前向滚动经验模态分解（forward rolling empirical mode decomposition，FEMD）对金融时间序列进行信号分解，采用主成分分析（principal component analysis，PCA）对分解后产生的高维向量组进行降维，整个过程是一个复杂的非线性特征提取过程。再将提取的特征输入一种新的利用 PCA 输出的加权 K 最邻近法（K-nearest neighbor，KNN）进行回归预测。该模型在特征提取过程的构造和整体结构上都是具有创新性的，并提出了比简单的 KNN 预测更有效的改进算法。实证结果证实了该模型对中国股票指数的预测效果。

关键词　经验模态分解，PCA，K 最邻近法，特征提取，预测

中图分类号　F832.5

1　引言

金融市场是一个庞大的、具有复杂运动模式的系统，受到来自各方面多重因素的影响。时间序列作为金融市场中最为主要的数据，是金融市场复杂内在的综合表现形式。通过对金融时间序列的分析及预测，我们可以发现市场潜在的规律及信息特征，为金融活动及决策提供重要依据，具有非常重要的现实意义。因此，对金融时间序列的分析及预测成了最重要的、最具挑战性的方向之一，而现有的文献已经取得了很多有价值的研究成果。

Abu-Mostafa 和 Atiya[1]证实了金融时间序列具有高噪声、非线性、非平稳和混沌等特征。高噪声表现为，我们不可能完全捕获的相关信息，在模型中未被考虑而形成的噪声。非平稳表现为，金融时间序列的分布随时间而改变。混沌则表现为，金融时间序列的趋势变化从短期来看是随机的，但从长期来看却是有确定性发展趋势的，这也为金融时间序列的预测分析提供了理论基础。

现有的对金融时间序列进行预测的模型包括线性和非线性两大类，其中比较有代表性的模型包括：建立于有效市场和随机漫步理论基础上的差分自回归移动平均模型[2]、自回归条件异方差模型[3]、广义自回归条件异方差模型[4]，认为在金融市场中具有可预测性的是价格波动率，而并非价格本身。而一些跨学科的研究从不同的角度，提出了基于混沌理论的模型[5]、神经网络（neural network，NN）[6]及支持向量机（support vector machine，SVM）[7]，这些模型可以通过对历史数据的有效处理来分析和挖掘金融市场的潜在规律并对市场进行概率性的预测。对金融市场的分析和预测主要在于对金融时间序列的分

* 基金项目：国家社会科学基金项目（17BGL231）。
通信作者：潘和平，成都大学高等研究院、商学院教授，重庆大学管理科学与房地产学院教授，博士生导师，E-mail：panhp@swingtum.com。

析与预测。1998 年 Huang 等[8]提出了 Hilbert-Huang 变换，其核心在于经验模态分解（empirical mode decomposition，EMD）和 Hilbert 变换，是一种数据驱动式自适应算法。其中 EMD 具有局部特征表现能力，能够有效反映系统本身的物理特性，是一种更适用于处理非线性、非平稳金融时间序列的方法。丁志宏和谢国权[9]通过 EMD 算法对沪深 300 指数的日收益进行了多尺度分解，更精确地提取了不同频率的分量，证明了 EMD 在金融时间序列分析中的广阔应用前景。王文波等[10]应用 EMD、混沌分析与 NN 进行组合分析，有效提高了金融时间序列预测的精度。Islam 等[11]对金融时间序列进行了多维 EMD，与小波算法相比，EMD 具有更优的预测效果。然而，张承钊和潘和平[12]指出 EMD 本身是存在内部缺陷的，随着输入时间序列的不断更新，其右端点对应的 EMD 结果是不断变化的，并不稳定。因此，他们提出了 FEMD 算法，并对沪深 300 指数和澳大利亚股票指数进行了实证分析，结果表明 FEMD 具有更高的预测精度。

在预测模型的构建中，特征提取是最重要的环节之一。Tsai 和 Hsiao[13]提出了三种特征提取方法：PCA、遗传算法（genetic algorithm，GA）和决策树，结果表明含 PCA 的组合模型效果更优。PCA 又称主成分分析，最初由 Pearson[14]在 1901 年提出，Hotelling[15]在 1933 年将其进行了发展。PCA 实际是一种正交线性变换，在消除变量间相关性的同时，充分考虑变量的统计特征，对主要信息进行综合有效的特征提取并达到数据降维效果。Roll 和 Ross[16]将 PCA 应用于套利定价理论的实证研究。徐国祥和杨振建[17]构建了组合模型 PCA-GA- SVM，并应用此模型对沪深 300 指数和前五大成分股的走势进行了有效分析。刘飞虎和罗晓光[18]构建了基于 PCA 和 RBF（radical basis function，径向基函数）神经网络的商业银行财务风险评价模型。

在金融时间序列的分析中，对人工智能中的模式识别理论也有着广泛应用。其中，KNN 于 1967 年由 Cover 和 Hart[19]提出，是一种简单而有效的非参数模式识别方法，可根据样本特征信息，直接进行模式识别。KNN 既可用于分类决策，也可用于回归预测，因此受到广泛关注，并衍生出一系列基于 KNN 的改进算法和其他组合集成的算法。Teixeira 和 de Oliveira[20]将 KNN 分类与技术分析工具相结合，提出了一个关于股票自动交易技术的新方法，并在智能预测系统中进行了有效的实际应用。王利等[21]提出了一种基于剪辑最近邻法的股市趋势的模式识别法，该方法利用样本集对其本身进行剪辑，筛选不同类交界处的样本，清理不同类的边界，过滤类别混杂样本，让类边界更加清晰。这样在减少样本数的同时，还能提高识别率。

在以上文献中，各个研究都是从模型结构的整体角度出发，在核心算法的基础上对模型进行优化改进。例如，徐国祥和杨振建[17]提出的 PCA-GA-SVM 模型，是在 SVM 算法的基础上，采用 PCA 和 GA 进行数据降维与动态更新模型参数，从而提高 SVM 算法的稳定性和预测精度。刘飞虎和罗晓光[18]提出的 PCA-RBF 模型，是在 RBF 算法的基础上，采用 PCA 来简化输入指标，找出主要的评价指标，从而解决模型输入样本缺乏的问题。在本文中，我们把对历史数据的特征提取过程看作预测模型构建的关键步骤，沿着历史数据输入、特征提取和局部非参数相似性预测的思路来进行预测模型的构建。实际上，我们可以把金融时间序列的特征提取过程理解为对其进行信号分析和信息融合的一个过程。因此本文将 FEMD、PCA 和 KNN 进行集成，提出了一种基于复杂的非线性特征提取过程和加权 KNN 的预测模型，简称 FEPK 模型。该模型首先采用 FEMD 对金融时间序列进行信号分解，其次用 PCA 对分解后生成的信息冗余的高维本征模态函数（intrinsic mode function，IMF）序列进行降维，最后将提取的特征输入一种新的利用 PCA 输出的加权 KNN 算法进行回归预测。FEPK 模型的整体结构是具有创新性的，包括了由 FEMD 和 PCA 组成的非线性特征提取过程，以及基于 PCA 输出的加权 KNN 回归预测。特别地，FEPK 模型中的特征提取过程对于金融时间序列的处理来说具有更强的适应性、全面性和正交性。同时，FEPK 模型预测采用了以 PCA 负荷作为权重的加权 KNN 算法，能够更合理、更有效地分类，具有更优秀的预

测性能。在实证中，FEPK 模型对真实的沪深 300 指数的历史数据进行了预测，结果表明 FEPK 模型具有良好的预测效果。

2 模型构建与组成部分

2.1 模型的总体结构

在对金融时间序列进行分析预测前，需要选取时间序列的时间尺度。本文的分析主要针对日数据，因此采用日作为基本时间尺度，即任意时间 t 对应一天。时间序列 $X(t)$ 表示第 t 天的数据，包含了开盘价 $O(t)$ 、最高价 $H(t)$ 、最低价 $L(t)$ 、收盘价 $C(t)$ 四个价格分量和一个交易量 $V(t)$ 。在本文的分析中，我们暂时只使用 $C(t)$ 作为预测分量，因此有 $X(t)$ 只包含 $C(t)$ ，后续论文中我们将会加入更多分量进行预测研究。

一般地，采用滑动窗口技术取一段足够长的历史数据 $\mathrm{DX}(t)$

$$\mathrm{DX}(t,N)=\big(X(t-(N-w)+1),\cdots,X(t-1),X(t)\big) \tag{1}$$

其中，t 为数据中最近的时间；N 为数据的总天数；$w<<N$ 为滑动窗口宽度；$\mathrm{DX}(t,N)$ 也可以表达为 $\mathrm{DX}(t,w)$ 。对于任意的 $X(t)$ ，定义当日的相对收益率为

$$R(t,\lambda)=\frac{X(t)-X(t-\lambda)}{X(t-\lambda)} \tag{2}$$

其中，λ 为预测步长，并设定其基本值 $\lambda=1$ ，本文在没有其他额外说明的情况下将使用 $R(t)$ 表示 $R(t,\lambda)$ 。相应地，有历史相对收益率序列

$$\mathrm{DR}(t,w)=\big(R(t-(N-w)+1),\cdots,R(t-1),R(t)\big) \tag{3}$$

根据预测模型的工作流程，一般的 FEPK 预测模型可以表达为

$$\mathrm{FEPK}:R(t+\lambda)=\mathrm{KNN}(F(t),k) \tag{4}$$

其中，$F(t)$ 为历史数据（ t 时间以前）经过 FEMD 和 PCA 以后，提取的特征集；k 为模型参数，表示 KNN 回归中与测试点最相近的 k 个最邻近元。为了更具体地展示预测模型的工作流程，本文可以展开特征提取过程，因此一般的 FEPK 预测模型可以更具体地表达为

$$\mathrm{FEPK}:R(t+\lambda)=\mathrm{KNN}(\mathrm{FE}(\mathrm{PCA}(\mathrm{FEMD}(\mathrm{DR}(t,w)),\alpha)),k) \tag{5}$$

其中，$\mathrm{FE}(\bullet)$ 为特征集的提取（ feature extraction，FE ），注意其中的 $\mathrm{FEMD}(\mathrm{DR}(t,w))$ 为 FEMD 的输入是采用滑动窗口技术动态截取的 t 时间前面长度为 w 的时间序列。

FEPK 预测模型的一般表达式可以直观地表达出该模型的整体结构是有别于现有的其他模型的，其核心在于非线性特征提取过程和局部非参数相似性预测的结合，而并不从整体上对一个核心算法进行改进。下面我们将具体介绍 FEPK 模型的两大组成部分。

2.2 非线性特征提取过程

2.2.1 FEMD

EMD 是一种能够有效处理非线性、非平稳时间序列的分解技术。但在 EMD 中，随着新数据的不断输入，其端点分解会不稳定[12]，因此本文将引入一种更适合处理金融时间序列的分解方法 FEMD。该方法先采用滑动窗口技术对原始历史时间序列进行捕获，再输入 EMD，这既能满足金融时间序列分析的实时性需求，又能增加预测模型的鲁棒性。其具体的分解流程如下。

先采用滑动窗口技术截取历史相对收益率序列 $\mathrm{DR}(i-1,w),\left[i=t,t-1,\cdots,t-(N-w)+1\right]$，输入 EMD 后生成 IMF，再将 IMF 作为模型训练的输入，$R(i),\left[i=t,t-1,\cdots,t-(N-w)+1\right]$ 作为模型训练的输出，因此可以得到模型训练的输入输出数据集 $\mathrm{DT}(t,N-w)$

$$\mathrm{DT}(t,N-w)=\left\{\begin{array}{ccc}\mathrm{EMD}(\mathrm{DR}(t-1),w) & \rightarrow & R(t)\\ \mathrm{EMD}(\mathrm{DR}(t-2),w) & \rightarrow & R(t-1)\\ \vdots & \vdots & \vdots\\ \mathrm{EMD}(\mathrm{DR}(t-(N-w)),w) & \rightarrow & R(t-(N-w)+1)\end{array}\right\} \tag{6}$$

其中，对 $\mathrm{DR}(i-1,w),\left[i=t,t-1,\cdots,t-(N-w)+1\right]$ 进行 EMD 是为了得到一系列 IMF。鉴于在一般的软件（如 Matlab）中都有 EMD 的具体算法，这里就不赘述。

经过 EMD 后，可以得到

$$\mathrm{DR}(t-1)=\left(\sum_{j=1}^{n}c_j\right)+r \tag{7}$$

也可以表示为

$$\mathrm{DR}(t-1)=\begin{pmatrix}\mathrm{IMF}_1(t-1,w)\\ \vdots\\ \mathrm{IMF}_n(t-1,w)\end{pmatrix}+r \tag{8}$$

原始序列被分解为了 n 个 IMF 分量（一般地，取 $n\leqslant 5$）和一个残差量 r。因此可以将式（6）改写为

$$\mathrm{DT}(t,N-w)=\{D\rightarrow R\} \tag{9}$$

$$D=\begin{pmatrix}\mathrm{IMF}_1(t-1,w) & \cdots & \mathrm{IMF}_n(t-1,w)\\ \mathrm{IMF}_1(t-2,w) & \cdots & \mathrm{IMF}_n(t-2,w)\\ \vdots & & \vdots\\ \mathrm{IMF}_1(t-(N-w),w) & \cdots & \mathrm{IMF}_n(t-(N-w),w)\end{pmatrix} \tag{10}$$

$$R=\begin{pmatrix}R(t) & R(t-1) & \cdots & R(t-(N-w)+1)\end{pmatrix}^{\mathrm{T}} \tag{11}$$

注意在式（10）中，矩阵 D 每一行的 IMF 序列都是一个高维数组，必然存在信息冗余，将影响预测的稳定性及精度。因此，我们将采用降维技术对 D 进行降维并消除冗余信息。

2.2.2 PCA 技术

在机器学习中，被广泛使用的降维技术包括：PCA、LDA（linear discriminant analysis，线性判别分析）、LLE（locally linear embedding，局部线性嵌入）和 LE（Laplacian eigenmaps，拉普拉斯特征映射）。但对比发现，这四种技术所追求的降维后的目标是不同的[22]：①PCA 追求降维后能够尽可能最大化地保持原始数据的潜在信息；②LDA 追求降维后能够尽可能容易地区分数据点；③LLE 追求降维后能够尽可能保持原有的流形结构；④LE 和 LLE 的思想类似，追求在降维后的空间中，相关点能够尽可能地靠近，以反映数据原有的流形结构。对于金融时间序列预测而言，预测输入的特征集需要在低维的情况下尽可能地保持原有数据的内在信息，以保证预测模型的鲁棒性和提高预测的精度。因此，本文将采用 PCA 对 FEMD 生成的高维数据矩阵 D 进行降维并提取特征集。

PCA 的主要过程在于将原来高维度的数据集映射到低维空间，转化为少数的富含信息量的主成分。从数学表示来看，PCA 实际就是用奇异值分解来实现数据降维的。其主要步骤如下。

首先对经过 FEMD 后得到的数据矩阵 D 进行标准化变换和奇异值分解，可得矩阵 Z

$$Z=U\Sigma W^{\mathrm{T}} \tag{12}$$

其中，U 和 W 分别为 ZZ^{T} 和 $Z^{\mathrm{T}}Z$ 的特征向量矩阵；Σ 为一个非负矩形对角矩阵，其左上角子矩阵的对角元素为矩阵 ZZ^{T} 的非零奇异值 σ_i，$i=1,2,\cdots,r$。由此，可以得到数据转换矩阵

$$Y = Z^{\mathrm{T}}U = W\Sigma^{\mathrm{T}}U^{\mathrm{T}}U = W\Sigma^{\mathrm{T}} \tag{13}$$

其中，矩阵 Y 的列依次为各个主成分。

实际上，在时间序列组成的矩阵 Z 中，其信息量主要集中于前面部分特征维上。因此，我们可根据对主成分的累计贡献率（cumulative contribution rate，CCR）进行约束来选取 r 个主成分中的前 $p << r$ 个，并组成新的低维矩阵。一般地，可约束 CCR 必须大于一个预设的阈值（如 85%）

$$\mathrm{CCR}_p = \left(\sum_{i=1}^{p}\sigma_i\right) \Big/ \left(\sum_{i=1}^{r}\sigma_i\right) > 85\% \tag{14}$$

各主成分对应的方差贡献率为

$$\mathrm{VCR}_i = \sigma_i \tag{15}$$

由此，矩阵 Y 对应的新的低维矩阵 Y_p 为

$$Y_p = Z^{\mathrm{T}}U_p = W\Sigma_p^{\mathrm{T}} \tag{16}$$

Y_p 就是 PCA 提取的低维数据矩阵，将和各主成分对应的方差贡献率 $\mathrm{VCR}_i\left(i=1,2,\cdots,p\right)$ 作为特征集输入下一步预测。

2.3 基于 PCA 输出的加权 KNN 算法

KNN 是一个经典的非参数算法，它不需要为自变量设定任何具体的函数，仅依赖于数据本身，就可以通过匹配历史序列中最相似的 k 个最邻近元来进行回归预测。本文提出了一种基于 PCA 输出的加权 KNN 算法。该算法将 PCA 提取的各主成分对应的方差贡献率作为权重，对选取的 k 个最邻近主成分进行信息融合，能够尽可能地体现各个邻近元的信息含量及其影响。因此，这种加权式的 KNN 算法比简单的 KNN 更为合理，具有更好的回归预测效果。其具体算法如下。

将 PCA 后得到的矩阵 Y_p、各主成分对应的 $\mathrm{VCR}_i(i=1,2,\cdots,p)$ 和预测点 $x(t+\lambda)$ 前一点 $x(t)=\mathrm{DR}(t,N)$ 作为 KNN 算法预测的输入，可构建预测模型

$$x(t+\lambda) = \mathrm{KNN}(x(t), Y_p, \mathrm{VCR}_i, k) \tag{17}$$

首先，计算点 $x(t)$ 与矩阵 Y_p 中任意点 $x_i(i=1,2,\cdots,p)$ 的相似度

$$S\left(x(t), x_i\right) = -\| x(t) - x_i \|^2 \tag{18}$$

在本文中，我们采取欧式距离作为相似度测度，后续论文中将重点关注更适合于金融时间序列相似度测度的函数。将得到的 S 进行排序，找到前 k 个最大的 S 值和最相似的 k 个最邻近元 $x_j(j=1,2,\cdots,k)$，其中 $k<p$。注意到本文提出的基于 PCA 输出的加权 KNN 算法需要将选出来的 k 个最邻近主成分对应的 $\mathrm{VCR}_j(j=1,2,\cdots,k)$ 考虑为权重系数并计算加权均值，因此，计算预测点 $x(t+\lambda)$

$$x(t+\lambda) = \sum_{j=1}^{k}\mathrm{VCR}_j \bullet x_j \tag{19}$$

其中，$x(t+\lambda)=\mathrm{DR}(t+\lambda,N)$，由式（3）可以输出 $R(t+\lambda)$。改进的 KNN 算法意味着在对最邻近主成分进行信息融合时，充分考虑了原始信息的包含量，具有更全面的信息输入，这比较于简单的 KNN 来讲可以有效提高模型的预测精度。

3 模型的结构参数及效能测度

3.1 模型的主要结构参数

根据式（5）可知，在 FEPK 模型中有四个主要的结构参数：w，α，λ，k。w是滑动窗口宽度；α是 IMF 所取的层数；λ 是预测步长；k 是最邻近元数量。对于一个具体的 FEPK 预测模型，也可以将式（5）表达为

$$\text{FEPK}: R(t+\lambda)=\text{KNN}\left\{\text{PCA}^{*}\left[\text{FEMD}\left(\text{DR}(t,w)\right),\alpha\right],k\right\} \tag{20}$$

其中，$\text{PCA}^{*}=\text{FE(PCA)}$ 为降维提取预测输入特征的过程。

3.2 预测模型的效能测度

在预测模型的效能测度上，我们通常采用能够测度实际值与预测值偏差的指标，如平均绝对百分比误差（mean absolute percentage error，MAPE）、均方根误差（root-mean-square error，RMSE）和平均绝对误差（mean absolute deviation，MAD）。但基于对金融市场风险和交易策略的考虑，本文将采用能够测度预测方向准确性的指标来进行实证结果比较，因此可以采用命中率[23]（hit rate，HR）

$$\text{hit rate}=\frac{1}{n}\sum_{i=1}^{n}d(i),\quad d(i)=\begin{cases}1, & R_i\times R^{*}>0\\ 0, & R_i\times R^{*}<0\end{cases} \tag{21}$$

其中，R_i 为相对收益率的真实值；R^{*} 为预测值；n 为数据点的总量。

4 历史数据与特征提取

对股票市场的历史数据进行有效的分析和预测可以为投资者的交易决策与投资策略提供依据，具有重要的现实意义。本文将采用滑动窗口技术对股市的历史时间序列进行截取，然后输入 FEMD 并获取一系列 IMF 分量，再利用 PCA 降维技术提取主成分及相应的方差贡献率作为特征集，最后将特征集输入 KNN 进行信息融合预测。

4.1 历史数据集

我国的基准股票指数是沪深 300 指数，它能够反映我国沪深股市的整体趋势。本文将选取沪深 300 指数的历史时间序列进行实证分析。通过国泰安数据库，我们选取了沪深 300 指数从 2007 年 1 月 4 日到 2017 年 7 月 28 日的真实历史价格时间序列。整个历史序列由 2571 个交易日的数据组成。为了有效训练数据和测试模型，我们将整个历史序列按照 8∶2 的比例分成了两个数据集：样本内训练数据集和样本外测试数据集。

4.2 沪深 300 指数的 FEMD

FEMD 的目的是要把非线性、非平稳的时间序列分解成若干个 IMF 分量和一个残差项。分解生成的 IMF 分量都是平稳的，且频率不同。在进行 FEMD 之前，我们需要利用滑动窗口技术对历史数据进行截取，再输入 FEMD。需要指出的是，在训练模型的过程中，滑动窗口宽度 w 和 FEMD 后 IFM 所取的层数 α 的取值为模型的超参数。w 的取值不能过大也不能过小。过大会导致分解延迟，影响预测的实时性能；过小又可能没有足够多的极值点，从而不能分解得到 IMF 序列，并且 FEMD 过程中是存在计算误

差的，而且随着分解层数的增加，误差增加。但是分解层数越多，分解出的 IMF 分量平稳性就越好，越能够增加预测精度。为了选择可靠的超参数，同时避免训练集和验证集划分对模型性能的影响，本文将使用 *K*-折交叉验证[24]（通常可做 10-折交叉验证）评估每种超参数组合下模型的预测精度，将历史数据集均分为 10 个子集，每次用 1 个作为测试集，其余 9 个作为训练集，重复 10 次，最后取平均命中率进行比较评估，验证结果如表 1 所示。结合考虑验证结果和预测结果的导向性，当滑动窗口宽度取 300 天，IMF 取 3 层时，效果最佳。图 1 是滑动窗口宽度为 300 天时，沪深 300 指数的 FEMD 结果图，包含 5 个 IMF 分量和一个残差项。可以看出，随着分解层数的增加，分解出的 IMF 分量的频率从高到低，逐渐降低，而最后的残差项反映的是一个平均趋势。

表 1　10-折交叉验证评估超参数 w 和 α 组合下 FEPK 模型训练的预测精度

滑动窗口宽度 w	命中率				
	$\alpha=1$	$\alpha=2$	$\alpha=3$	$\alpha=4$	$\alpha=5$
100	0.7067	0.6881	0.7137	0.6805	0.6781
150	0.7019	0.6942	0.6975	0.6869	0.7018
200	0.7113	0.7131	0.7044	0.7133	0.6982
250	0.7142	0.6971	0.7020	0.6956	0.6974
300	0.7177	0.6984	0.7211	0.6976	0.7190
350	0.6989	0.6735	0.7103	0.7011	0.7050

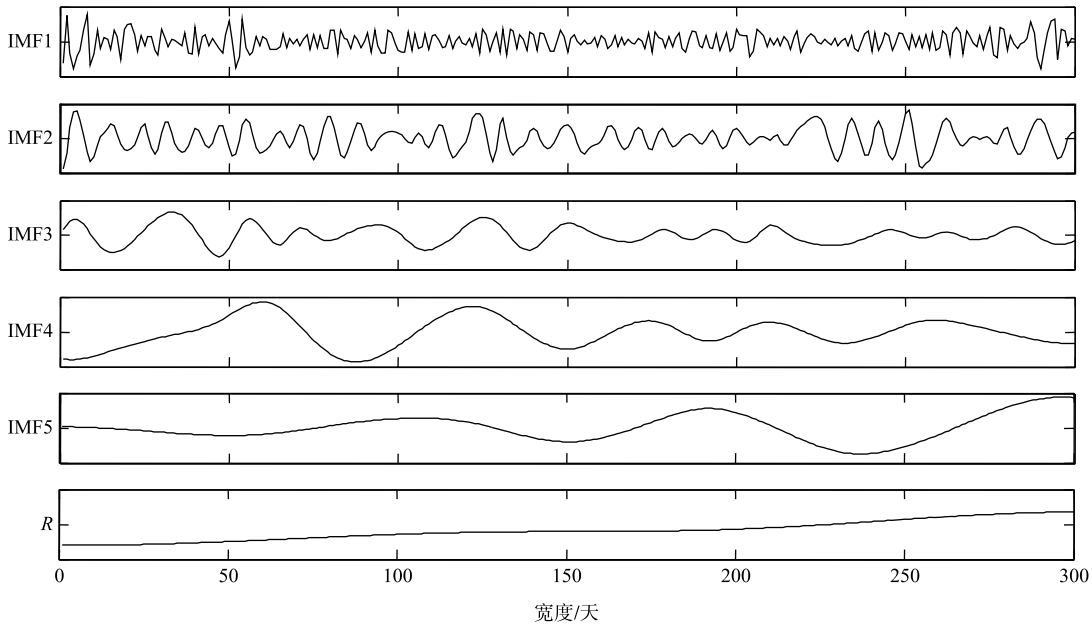

图 1　沪深 300 指数的一个滑动窗口数据（300 天）的 FEMD 结果图

4.3　沪深 300 指数的 PCA 降维及特征提取

经过 FEMD 后，我们得到的是一个由 IMF 序列组成的具有冗余信息的高维数组，需要对其进行降维处理。采用 PCA 技术进行降维，可获得富含信息的主成分及其相应的方差贡献率和 CCR。沪深 300 指数训练数据集经过 PCA 降维以后产生的各个主成分的方差贡献率和 CCR 如表 2 所示。可以看到在降维后得到的前 6 个主成分的 CCR 就达到了 87.9372%，超过了 85%的常用标准。取前 20 个主成分时，

其 CCR 更是达到了 99.99%。为了更好地训练模型，提高预测精度，我们将保留前面 20 个主成分及其对应的方差贡献率作为沪深 300 指数的特征集，输入 KNN 进行信息融合预测。

表 2 沪深 300 指数的主成分方差贡献率及 CCR

主成分序号	方差贡献率	CCR	主成分序号	方差贡献率	CCR
1	52.8302%	52.8302%	11	1.1612%	96.5340%
2	15.4979%	68.3281%	12	0.9606%	97.4946%
3	7.5220%	75.8501%	13	0.7868%	98.2814%
4	5.2103%	81.0604%	14	0.6228%	98.9042%
5	3.8890%	84.9494%	15	0.4421%	99.3463%
6	2.9878%	87.9372%	16	0.2910%	99.6373%
7	2.3447%	90.2819%	17	0.1915%	99.8288%
8	1.9695%	92.2514%	18	0.1058%	99.9346%
9	1.6975%	93.9489%	19	0.0402%	99.9748%
10	1.4239%	95.3728%	20	0.0152%	99.9900%

5 模型实证与分析

我们构建了一个具体的 FEPK 预测模型 FEPK_CSI 300_D1 预测沪深 300 指数（CSI 300）$t+1$ 日线收益率，根据式（20），该模型可以具体表达为

$$R(t+\lambda)=\mathrm{KNN}\left\{\mathrm{PCA}^{*}\left[\mathrm{FEMD}\left(\mathrm{CSI\,300\_D1\_DR}(t,w)\right),\alpha\right],k\right\} \tag{22}$$

表 3 显示了 FEPK_CSI 300_D1 预测模型在样本外测试的命中率，在 $w=300$ 和 $k=4$ 时，达到了最佳命中率 0.7542（75.42%）。因此，对于预测沪深 300 指数的 $t+1$ 日线收益率来说，FEPK_CSI 300_D1 是一个预测效能优秀的预测模型。

表 3 FEPK_CSI 300_D1 预测沪深 300 指数 $t+1$ 日线收益率在不同的 w 和 k 最邻近元上的命中率

滑动窗口宽度 w	命中率							
	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$
100	0.6969	0.7153	0.7072	0.7031	0.7010	0.7521	0.7501	0.7256
150	0.7119	0.7136	0.7157	0.7428	0.7161	0.7177	0.7140	0.7265
200	0.7203	0.7224	0.7096	0.7117	0.6861	0.6776	0.6499	0.6456
250	0.6815	0.6466	0.6967	0.6989	0.7338	0.7294	0.7512	0.7490
300	0.7122	0.7256	0.7367	0.7542	0.7256	0.7145	0.6900	0.7033
350	0.7262	0.7148	0.7308	0.7125	0.7103	0.6989	0.6897	0.6852

为了更进一步对比、验证 FEPK 预测模型的有效性，我们用沪深 300 指数的历史价格时间序列同样训练和测试了 FEMD+KNN 和 KNN 预测模型。为了更为直观地对比不同模型的预测效能，我们仅选取实证结果中最高命中率进行比较，对比结果如表 4 所示。其中最高命中率是由 FEPK_CSI 300_D1 模型的预测结果得出，达到了 0.7542（75.42%）。对表 4 中各个模型的命中率进行比较分析，容易得出 FEPK

模型的预测精度优于 FEMD+KNN 模型，而 FEMD+KNN 模型的预测精度又优于 KNN 模型。这也证实了 FEMD 后得到的 IMF 分量可以更有效地展示数据特征，提高预测精度。同时说明了经过 PCA 降维处理后保留的主成分可以最大限度地保留数据的原始信息。

表 4　FEPK 模型与 FEMD+KNN、KNN 模型的预测效能对比结果

具体模型		最高命中率
沪深 300 指数 $t+1$ 日线收益率	FEPK_CSI 300_D1	0.7542
	FEMD+KNN_ CSI 300_D1	0.7294
	KNN_ CSI 300_D1	0.6963

6　结论

本文提出了一种基于复杂非线性特征提取过程和加权 KNN 回归的金融时间序列预测模型——FEPK 模型。该模型的整体结构是具有创新性的；特征提取过程对于金融时间序列来说，具有特征提取的适应性、全面性和正交性；采用以 PCA 负荷作为权重的加权 KNN 进行预测，比简单的 KNN 更合理，分类效果更好，具有更好的预测性能。在预测沪深 300 指数的实证结果上也证明了 FEPK 模型的有效性。但如何将这种可预测性与市场中的投资风险管理和交易策略相结合，还需要进一步的研究。

在后续的研究中，我们将考虑多元信息输入的情况，如增加开盘价、最高价、最低价和交易量。另外，我们会继续探索更有效的特征提取方法和预测方法，如“自编码器”“随机森林”等。

参 考 文 献

[1] Abu-Mostafa Y S，Atiya A F. Introduction to financial forecasting[J]. Applied Intelligence，1996，6（3）：205-213.
[2] Qin M J，Li Z H，Du Z H. Red tide time series forecasting by combining ARIMA and deep belief network[J]. Knowledge-Based Systems，2017，125：39-52.
[3] Engle R F. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation[J]. Econometrica，1982，50（4）：987-1007.
[4] Bollerslev T. Generalized autoregressive conditional heteroskedasticity[J]. Journal of Econometrics，1986，31（3）：307-327.
[5] Ravi V，Pradeepkumar D，Deb K. Financial time series prediction using hybrids of chaos theory，multi-layer perceptron and multi-objective evolutionary algorithms[J]. Swarm and Evolutionary Computation，2017，36：136-149.
[6] Galeshchuk S. Neural networks performance in exchange rate prediction[J]. Neurocomputing，2016，172：446-452.
[7] Sermpinis G，Stasinakis C，Theofilatos K，et al. Modeling，forecasting and trading the EUR exchange rates with hybrid rolling genetic algorithms-support vector regression forecast combinations[J]. European Journal of Operational Research，2015，247（3）：831-846.
[8] Huang N E，Shen Z，Long S R，et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis[J]. Proceedings：Mathematical，Physical and Engineering Sciences，1998，454（1971）：903-995.
[9] 丁志宏，谢国权. 金融时间序列多分辨率实证研究的 EMD 方法[J]. 经济研究导刊，2009，（6）：61-63.
[10] 王文波，费浦升，羿旭明. 基于 EMD 与神经网络的中国股票市场预测[J]. 系统工程理论与实践，2010，30（6）：1027-1033.
[11] Islam M R，Rashed-Al-Mahfuz M，Ahmad S，et al. Multiband prediction model for financial time series with multivariate empirical mode decomposition[J]. Discrete Dynamics in Nature and Society，2012，2012：21.
[12] 张承钊，潘和平. 基于前向滚动 EMD 技术的预测模型[J]. 技术经济，2015，34（5）：70-77.
[13] Tsai C-F，Hsiao Y-C. Combining multiple feature selection methods for stock prediction：union，intersection，and multi-intersection approaches[J]. Decision Support Systems，2010，50（1）：258-269.
[14] Pearson K. On lines and planes of closet fit to systems of points in space[J]. Philosophical Magazine，1901，2（11）：559-572.
[15] Hotelling H. Analysis of a complex of statistical variables into principal components[J]. Journal of Educational Psychology，

1933，24（6）：417-441.
[16] Roll R，Ross S R. An empirical investigation of the arbitrage pricing theory[J]. The Journal of Finance，1980，35（5）：1073-1103.
[17] 徐国祥，杨振建. PCA-GA-SVM 模型的构建及应用研究——沪深 300 指数预测精度实证分析[J]. 数量经济技术经济研究，2011，28（2）：135-147.
[18] 刘飞虎，罗晓光. 基于 PCA-RBF 神经网络的商业银行财务风险评价研究[J]. 投资研究，2013，（3）：88-97.
[19] Cover T，Hart P. Nearest neighbor pattern classification[J]. IEEE Transactions on Information Theory，1967，13（1）：21-27.
[20] Teixeira L A，de Oliveira A L I. A method for automatic stock trading combining technical analysis and nearest neighbor classification[J]. Expert Systems with Applications，2010，37（10）：6885-6890.
[21] 王利，周美娇，张凤登，等. 基于剪辑最近邻法的股市趋势的模式识别分类研究[J]. 微计算机信息（测控自动化），2010，26（12/1）：237-239.
[22] CSDN.四大机器学习降维方法[EB/OL]. https://blog.csdn.net/rosenor1/article/details/52278116[2017-08-06].
[23] Pan H P，Haidar I，Kulkarni S. Daily prediction of short-term trends of crude oil prices using neural networks exploiting multimarket dynamics[J]. Frontiers of Computer Science in China，2009，3（2）：177-191.
[24] 胡局新，张功杰. 基于 K 折交叉验证的选择性集成分类算法[J]. 科技通报，2013，29（12）：115-117.

Prediction Model Based on a Nonlinear Feature Extraction and Weighted K-Nearest Neighbor

TANG Li[1]，PAN Heping[2]，YAO Yiyong[1]

（1. School of Intelligent Finance，Tianfu College of Southwestern University of Finance and Economics，Chengdu 610052，China）

（2. Business School，Chengdu University，Chengdu 610106，China）

Abstract This paper proposes an intelligence financial prediction model consists of a forward rolling Empirical Mode Decomposition（FEMD）for financial time series signal decomposition，Principal Components Analysis（PCA）for dimension reduction，and a weighted K-Nearest Neighbor for prediction. Generally，the structure of this model is original. The feature extraction process integrating FEMD and PCA is an advanced special extraction method for financial time series signal analysis. It has the adaptability，comprehensiveness and orthogonality of feature extraction. Moreover，the weighted KNN with PCA loading as weights is more reasonable and has better effect on classifying than a simple KNN，thus it has better prediction performance. The empirical results on CSI 300 prediction has confirmed that the FEPK model performs better than others.

Key words EMD，PCA，KNN，Feature extraction，Prediction

作者简介

唐黎（1984—），女，西南财经大学天府学院智能金融学院副院长、金融工程博士，研究方向：金融工程和智能金融。E-mail：tina@tfswufe.edu.cn。

潘和平（1961—），男，成都大学高等研究院教授、重庆金融学院智能金融研究中心主任教授、武汉大学遥感金融联合研究中心执行主任、博士生导师、长江学者，研究方向：金融工程和智能金融。E-mail：panhp@swingtum.com。

姚一永（1975—），男，西南财经大学天府学院副校长、副教授，研究方向：智能金融、量化交易和数据分析。E-mail：yiyongyao@yahoo.com。