基于 PCA 和 AP 的嵌套式 KNN 金融时间序列预测模型

唐 黎[1]，潘和平[2,3]，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；
3. 重庆金融学院 智能金融研究中心，重庆 400067)

摘　要： 本文提出一种金融时间序列预测的数据降维与信息融合计算智能模型-PANK 模型。该模型由三个部分组成：(1) 主成分分析(Principal Component Analysis，PCA)，用于减少冗余信息；(2) 仿射传播聚类(Affinity Propagation，AP)，用于找到聚类中心和相应的聚类作为特征提取；(3) 嵌套式 k-最邻近元(Nested k-Nearest Neighbor，Nested KNN)用于回归预测。PANK 模型先采用滑动窗口技术截取最近期的金融时间序列作为输入数据，再经过 PCA 减少冗余信息，提取富含有效信息的主成分，并将其放入 AP 中进行聚类，最后采用两层嵌套式 Nested KNN 预测。本文特别提出了一种新的嵌套式 Nested KNN，可以有效解决 KNN 中的两个主要问题：计算量大和不均衡样本问题。通过对模型在欧元兑美元汇率和中国沪深 300 股指上的实证，结果表明 PANK 预测模型可达到 80% 的最佳命中率。

关键词： 金融时间序列；主成分分析；仿射传播聚类；最邻近元；金融市场预测
中图分类号： F832.5　**文献标识码：** A　**文章编号：** 1003-5192(2019)01-0091-06　**doi：** 10.11847/fj.38.1.91

A Financial Time Series Prediction Model Integrating Principal Component Analysis, Affinity Propagation and Nested K-Nearest Neighbor Regression

TANG Li[1], PAN He-ping[2,3], YAO Yi-yong[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*; 3. *Intelligent Finance Research Center, Chongqing Institute of Finance, Chongqing* 400067, *China*)

Abstract: This paper advances a complex computational intelligence model for financial time series prediction, called PANK for PCA-AP-Nested KNN. As an information fusion and computational intelligence model, PANK integrates Principal Component Analysis (PCA), Affinity Propagation Clustering (AP), and a nested reformulation of k-Nearest Neighbor regression (Nested KNN). PANK model uses a sliding window to capture a certain length of recent time series data, then applies PCA to reduce the dimensionality of the data and transform the time series into principal components with rich information as the input for AP clustering, and uses Nested KNN for prediction modeling. The original KNN is updated to Nested KNN which can tackle the large amount of computation and disequilibrium samples problem of original KNN. Two specific PANK models are trained and tested on EUR/USD exchange rate and Chinese HS300 index with 15-year historical data, achieving best hit rate of 80%, higher than other reference models based on KNN.

Key words: financial time series; principal component analysis; affinity propagation; k-nearest neighbor; financial market prediction

1　引言

金融时间序列预测是一个极具挑战性的理论和技术问题，具有重要的经济意义，近年来引起了金融界和计算机界的广泛关注。金融时间序列概率可预测性已被大量文献所证实。在有效市场假说(EMH)的基础上，Box 和 Jenkins[1] 提出了 ARIMA 模型，Engle[2] 提出了 ARCH 模型，Bollerslev[3] 提出了 GARCH 模型，认为金融市场的价格波动有可预测性，但并非是价格本身。而一些跨学科的研究人员采用从其他不同角度继续研究金融市场预测，包括常见的混沌理论模型[4]，支持向量机(SVM)的预测模型[5]，神经网络(NN)的预测模型[6]，和基于 k-最邻近元(KNN)的预测模型[7]。

在这些常用的预测模型中，Cover 和 Hart[8] 提出的 KNN 可以处理高维数据，并能够简单直观地

收稿日期： 2018-01-06
基金项目： 国家社会科学基金资助项目(17BGL231)

从特征空间中提取相似实例进行预测分析，由此被广泛应用于时间序列预测[9]。但是 KNN 有两个明显的缺陷会影响其性能也是众所周知的。第一，计算量过大，因为需要计算测试点与已知样本中的所有点的相似度，并找到 k 个最邻近元；第二，受不均衡样本影响，当不同类所含样本点数量差距较大时，测试点的 k 个最邻近元中，大样本类容易占多数而影响测试点的分类或者回归效果。为了解决这两个问题，本文提出了一种嵌套式 KNN，称为 Nested KNN。Nested KNN 分成两层：第一层首先计算预测点前一时点 t 与各个聚类中心的相似度，并找到其最近的聚类中心和相应的聚类；第二层计算 t 与第一层输出的聚类中的各个点的相似度，并找到 t 的 k 个最邻近元。由于第一层只需要计算 t 与各个聚类中心的相似度，所以大大减少了算法的计算量；第二层只计算 t 与同一聚类中的各个点的相似度，因此避免了受不均衡样本的影响。

Tsai 和 Hsiao[10] 指出特征提取是金融时间序列预测的关键。本文采用主成分分析(PCA)和仿射传播聚类(AP)对历史数据集进行特征提取，将输出的特征作为 Nested KNN 预测的输入。由此提出了一种将 PCA、AP 和 Nested KNN 组合起来的智能预测模型，简称为 PANK 模型(PCA + AP + Nested KNN)。该模型将用于实证预测欧元兑美元汇率及沪深 300 指数的走势。

2　PANK 预测模型的结构框架

本文在 KNN 回归的基础上，提出了 PANK 金融时间序列预测模型。PANK 模型通过 PCA 来减少原始金融时间序列中的冗余信息，生成富含有效信息的主成分，并将其输入 AP 进行聚类以找到时间序列的最优聚类方案和相应的聚类中心集，最后输入两层嵌套式 Nested KNN 进行回归预测。

一般地，我们取一段足够长的历史数据对模型进行训练和测试。首先，需要确定数据的时间框架，本文采用日数据，用日作为基本时间尺度。时间序列 $X(t)$ 表示 t 时间(一天)的数据，包含了四个价格分量：开盘价 $X.O(t)$、最高价 $X.H(t)$、最低价 $X.L(t)$、收盘价 $X.C(t)$ 和交易量 $X.V(t)$。在本文中，我们只考虑 $X.C(t)$，因此在后面的论述中 $X(t)$ 仅包含 $X.C(t)$，在后续研究中我们将会加入更多的分量。

对于任意 $X(t)$，定义当日相对收益率

$$RR(t,\lambda) = \frac{X(t) - X(t-\lambda)}{X(t-\lambda)} \tag{1}$$

其中 λ 表示预测步长，最基本预测步长为 1，因此在没有具体说明时，本文用 $RR(t)$ 表示 $RR(t,\lambda)$。取足够长的历史相对收益率数据

$$DR(t,T) = (RR(t), RR(t-1), \cdots, RR(t-(T-m)+1)) \tag{2}$$

其中 t 表示所取得的数据中最近的时间点；T 表示全部数据的总天数；$m << T$ 表示截取数据时的滑动窗口的宽度，$DR(t,T)$ 也可以表达为 $DR(t,m)$。

一般的 PANK 预测模型可以表达为

$$\text{PANK: } RR(t+\lambda) = NKNN(C(t),k) \tag{3}$$

其中 $NKNN$ 表示嵌套式 KNN；$C(t)$ 是 AP 聚类后生成的最优聚类结果，是预测模型提取的特征集；k 是模型参数；$RR(t+\lambda)$ 是模型输出的预测日的相对收益率。具体地，PANK 模型使用 PCA 和 AP 聚类生成特征集 $C(t)$，因此，PANK 模型可以进一步地表达为

$$\text{PANK: } RR(t+\lambda) = NKNN\{FE[AP(PCA(DR(t,m)))],k\} \tag{4}$$

其中 $FE(\ )$(Feature Extraction，FE)表示聚类提取特征值的过程。

3　PANK 预测模型的三个组成部分及其算法

PANK 模型由 PCA、AP 和 Nested KNN 三个部分组成。模型首先采用 PCA 提取富含原始数据信息的主分量，并输入 AP 进行信息传播聚类，生成特征集 $C(t)$，最后输入 Nested KNN 进行回归预测。下面将依次介绍 PANK 模型的各个组成部分及其算法流程。

3.1　主成分分析(PCA)

主成分分析(Principal Component Analysis，PCA)是 1901 年由 Pearson 提出并于 1933 年由 Hotelling发展[11,12]，是一种常用的减少冗余信息，对大数据进行降维的方法。在应用 PCA 提取主分量之前，需要采用滑动窗口(窗口宽度为 m)技术截取历史数据[13]，形成预测模型的输入，并形成预测模型训练输入输出数据集

$$DT(t,T-m) = \{D \rightarrow R\} \tag{5}$$

$$D = (DR(t-1,m)\ DR(t-2,m) \cdots DR(t-(T-m),m))' \tag{6}$$

$$R = (RR(t)\ RR(t-1) \cdots RR(t-(T-m)+1))' \tag{7}$$

PANK 模型的第一步是 PCA 变换减少冗余数据，获取主分量，这实际也是一个奇异值分解过程。首先对历史数据矩阵 D 进行标准化处理和奇异值分解，可得矩阵 Z

$$Z = U\Sigma V^{\mathrm{T}} \tag{8}$$

其中 U 和 V 都是正交矩阵，分别是 ZZ^{T} 和 $Z^{\mathrm{T}}Z$ 的特征向量矩阵。Σ 是非负矩形对角矩阵，其左上角的子矩阵的对角元素 $\lambda_i(i=1,2,\cdots,r)$ 为 ZZ^{T} 的特征值。由此可以得到数据转换矩阵

$$P = Z^{\mathrm{T}}U = V\Sigma^{\mathrm{T}}U^{\mathrm{T}}U = V\Sigma^{\mathrm{T}} \tag{9}$$

其中矩阵 P 的各列依次为各个主成分。

实际上,在时间序列组成的矩阵中,存在信息冗余,其信息主要集中于前面部分主分量上。因此,我们可根据对主成分累积贡献率(Cumulative Contribution Rate, CCR)的约束来提取前 $l(l<<r)$ 个主分量组成新的低维矩阵 P^* 取代原始数据矩阵 D 作为预测输入。一般地,设定 CCR 必须高于预设的阈值(如 85%)

$$CCR_l=\left(\sum_{i=1}^{l}\lambda_i\right)/\left(\sum_{i=1}^{r}\lambda_i\right)>85\% \tag{10}$$

3.2 仿射传播聚类(AP)

仿射传播聚类(Affinity Propagation, AP)是 2007 年 Frey 和 Dueck [14] 在 Science 上提出的一种快速有效的新聚类算法,该算法事先将所有的数据点都看作是可能的聚类中心,通过不断的循环迭代,最终获得包含有一系列聚类中心和相应的聚类集的聚类方案。在含有 N 个数据点的样本中,AP 算法首先计算每两个数据点之间的相似度,并组成相似度矩阵 $S_{N\times N}$。在此基础上,每个数据点都被看作是潜在的聚类中心(称为 exemplar),然后通过迭代循环,在信息的搜集和传递过程中不断竞争,最后产生一系列最优的聚类中心,并将各个数据点分配到最相似的聚类中心所代表的类中,形成 AP 最优的聚类结果 [14]。下面将具体说明 AP 的算法流程。

在金融时间序列经过 PCA 后,将选出的主成分组成样本空间,计算每两个样本点 i 和 j 之间的相似度 $S(i,j)$。在本文中,以欧氏距离作为相似度测度,后续的研究中将探索更适合金融时间序列相似度的测度方法,有

$$S(i,j)=-\|i-j\|^2 \tag{11}$$

同时生成了由所有的相似度 $S(i,j)$ 组成的相似度矩阵 S。

为了找到最合适的聚类中心,AP 算法在相似度矩阵 S 的基础上,对每个样本点进行信息搜集和传递,并进行迭代循环。在每一个迭代循环过程中,对于任意一个潜在的聚类中心 e,都从任意一个样本点 i 搜集信息 $R(i,e)$,同时,也为点 i 从潜在的聚类中心 e 搜集信息 $A(i,e)$

$$R(i,e)=S(i,e)-\max_{j\neq e}\{A(i,j)+S(i,j)\} \tag{12}$$

$$A(i,e)=\min\{0,R(e,e)+\sum_{j\neq i,e}\max(0,R(j,e))\} \tag{13}$$

其中 $R(i,e)$ 表示点 e 对点 i 的吸引度(responsibility)或者说是点 e 适合作为点 i 的聚类中心的程度。$A(i,e)$ 表示点 i 对点 e 的归属度(availability)或者说是点 i 选择点 e 作为其聚类中心的适合程度。整个过程中 $R(i,e)$ 和 $A(i,e)$ 不断被计算迭代直到 $R(i,e)+A(i,e)$ 达到最大值,点 e 才是选出来的最适合点 i 的聚类中心。

需要指出的是,在相似度矩阵的对角线上存在非常重要的偏向参数 p,它表示每个点被选作聚类中心的倾向性。一般地,设定 p 的初始值 p_m 为相似度矩阵 S 中元素的中值,下降步幅 [15]

$$p_{step}=0.01p_m/0.1\sqrt{h+50} \tag{14}$$

在迭代循环的过程中,若聚类个数收敛到某个值 h 时,以 p_{step} 逐渐减小 p,并继续迭代,以获得不同聚类个数的不同聚类方案。为了在不同的聚类方案中,选取一个最优的,可引入能够有效反映聚类结构中的类内紧密性和类间分离性的 Silhouette 指标 [15]。假设样本空间被分成了 r 个聚类 $C_i(i=1,2,\cdots,r)$,可以计算点 x^* 的 Silhouette 指标

$$Sil(x^*)=\frac{\min\{d(x^*,C_i)\}-a(x^*)}{\max\{a(x^*),\min\{d(x^*,C_i)\}\}} \tag{15}$$

其中 $d(x^*,C_i)$ 表示聚类 C_j 中的点 x^* 与另一聚类 $C_i(i\neq j)$ 中所有的点之间的平均不相似度,$a(x^*)$ 表示聚类 C_j 中的点与同聚类中其他所有点的平均不相似度。由此,可以计算出整个样本空间中所有点的 Silhouette 指标的平均值 $Sil_{average}$

$$Sil_{average}=\text{mean}\{\sum_{t=1}^{N}Sil(x^*)\} \tag{16}$$

$Sil_{average}$ 值可以有效反映聚类结果的质量,$Sil_{average}>0.5$ 表示各个不同聚类间具有明显的可分离性,$Sil_{average}$ 值越大表示聚类质量越好 [15]。

3.3 嵌套式 k-最邻近元(Nested KNN)

作为一种简单、直观、有效的非参数模式识别方法,KNN 既能用于分类,又能用于回归,因此被广泛应用 [16,17]。但是 KNN 算法有两个非常明显的缺点: 计算量过大和受不均衡样本影响。为了改进 KNN 算法的这两个不足之处,本文特别提出了一种嵌套式的 KNN 算法,称为 Nested KNN。

对于一段足够长的、包含 N 个样本点的历史时间序列,可构建基于原始 KNN 的预测模型

$$x(t+\lambda)=KNN(N,k) \tag{17}$$

首先,用欧氏距离作为相似度测度,计算测试点 $x(t)$ 与任意样本点 $x_n(n=1,2,\cdots,N)$ 的相似度

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

将 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 $k<N$。由此,可以输出预测点 $x(t+\lambda)$

$$x(t+\lambda)=\frac{1}{k}\sum_{j=1}^{k}x_j \tag{19}$$

针对 KNN 的两大不足之处，本文特别在以上模型的基础上，将 KNN 改进为两层嵌套式算法 Nested KNN。该算法由三个函数: *NKNN*、*NKNN*1 和 *NKNN*0 组成。其中 *NKNN* 作为主函数，对应整个算法的输入-输出，并调用 *NKNN*1 找到与输入测试点最相似的聚类中心和相应的聚类; *NKNN*1 再调用 *NKNN*0 在前一步输出的聚类中找到与测试点最相似的 k 个最邻近元用于回归预测。具体如下:

函数 *NKNN*1，输入预测点 $X_{t+\lambda}$ 的前一点 $X_t = DR(t,T)$ 和 AP 聚类输出的 r 个类 $C_i(i=1,2,\cdots,r)$ 的集合 C 和各个聚类中心 $e_i(i=1,2,\cdots,r)$ 的集合 E，输出与 X_t 最相似的聚类中心 $e_{nearest}$ 和相应的聚类 $C_{nearest}$，即

$$(C_{nearest}, e_{nearest}) = NKNN1(X_t, C, E, k=1) \quad (20)$$

计算 X_t 与任意聚类中心 $e_i(i=1,2,\cdots,r)$ 之间的相似度 $S(X_t, e_i)$。当 $S(X_t, e_{nearest})$ 为最大值时，$e_{nearest}$ 就是与 X_t 最相似的聚类中心，由此，可以得到以 $e_{nearest}$ 为代表的聚类 $C_{nearest}$，并将 X_t 归为此类。

函数 *NKNN*0，输入 X_t 和 $C_{nearest}$，输出 $X_{t+\lambda}$，即

$$(X_{t+\lambda}) = NKNN0(X_t, C_{nearest}, k) \quad (21)$$

计算 X_t 与 $C_{nearest}$ 中任意点 X_j^* 之间的相似度 $S(X_t, X_j^*)$。将 $S(X_t, X_j^*)(j=1,2,\cdots,k)$ 排序，找到前 k 个最大的 S 值和最相似的 k 个最邻近元 $X_j^*(j=1,2,\cdots,k)$。由此计算得出

$$X_{t+\lambda} = \frac{1}{k}\sum_{j=1}^{k} X_j^* \quad (22)$$

其中参数 k 直接影响输出结果。在实际的测试中，具体的不同样本都对应于不同的最优 k 值，因此，为了得到更好的预测效果，本文将通过模型训练，找到最优的 k 值。

函数 *NKNN*，输入 X_t、C 和 E，先调用 *NKNN*1 输出 $C_{nearest}$ 和 $e_{nearest}$，再调用 *NKNN*0 输出 $X_{t+\lambda}$，即

$$(X_{t+\lambda}) = NKNN(X_t, C, E, k) \quad (23)$$

其中有 $X_{t+\lambda} = DR(t+\lambda, T)$，由此可以得出 $RR(t+\lambda)$。

在 Nested KNN 算法中，*NKNN*1 只需要计算测试点 X_t 与各个聚类中心之间的相似度，因此较 KNN 算法极大地减少了计算量。同时在 *NKNN*0 中，只需要计算测试点 X_t 与其最相似的聚类中心所在聚类 $C_{nearest}$ 中的各个点之间的相似度，因此有效避免了不均衡样本的问题。

4　PANK 预测模型的结构参数和效能测度

4.1　PANK 预测模型的结构参数

在构建一个具体的 PANK 预测模型时，需要设定三个关键的模型结构参数: m, λ, k。其中 m 表示截取历史金融时间序列的滑动窗口宽度; λ 表示预测未来时间序列的步幅; k 表示所取最邻近元的个数。由此，可以将(3)式和(4)式中的 PANK 模型表达为

$$\text{PANK: } RR(t+\lambda) = NKNN\{AP^*\ [PCA(DR(t,m))], k\} \quad (24)$$

其中 $AP^* = FE(AP)$ 表示聚类提取预测输入的过程。

4.2　PANK 预测模型的效能测度

对于时间序列预测模型的效能测度，我们常见的指标有均方根误差(Root Mean Square Error, RMSE)[18]、平均绝对百分比误差(Mean Absolute Percentage Error, MAPE)[19]和平均绝对误差(Mean Absolute Difference, MAD)，用来测度实际值与预测值偏差。然而，对于量化投资而言，只有在预测趋势方向有了显著的正确率后，这些效能测度指标才有意义。因此，本文对于具体 PANK 预测模型的效能测度将采用衡量预测趋势方向正确性的命中率(Hit Rate, HR)[20]

$$HR = \frac{1}{n}\sum_{i=1}^{n} hr_i, hr_i = \begin{cases} 1, & RR_i \times RR_i^* \geq 0 \\ 0, & RR_i \times RR_i^* < 0 \end{cases} \quad (25)$$

其中 RR_i 和 RR_i^* 分别是相对收益率的真实值和预测值，n 是样本点的总数目。

5　PANK 预测模型的实证分析

浮动汇率制度在世界市场有效，大家对外汇市场广泛关注，为了避免外汇市场风险，提出了各种汇率趋势预测的模型[21,22]。与此同时，中国的股市，如沪深 300 股指等也受到了研究者的关注[23]。在本文的模型实证部分，我们构建了两个具体的 PANK 预测模型，分别对欧元兑美元汇率和沪深 300 指数真实的历史数据进行预测实证。

5.1　PANK_EURUSD_D1 预测欧元兑美元汇率日线收益率

PANK_EURUSD_D1 预测模型针对欧元兑美元汇率日线收盘价数据，对 $t+1$ 日线收益率进行预测，该模型可从(24)式具体化为

$$\text{PANK\_EURUSD\_D1: } RR(t+1) = NKNN\{AP^*\ [PCA(EURUSD\_D1\_DR(t,m))], k\} \quad (26)$$

模型训练和测试的历史数据集由 2002 年 11 月 1 日至 2017 年 11 月 24 日期间的 3909 个交易日数据组成，前面时段的 3128 个数据点用于样本内训练，后面时段的 781 个数据点用于样本外检验。表 1 显示了 PANK_EURUSD_D1 预测模型的样本外检验预测命中率结果，其中在 $m=25$ 和 $k=5$ 时，取得了最高的命中率 0.80(80%)，这表明

PANK_EURUS_D1 模型是一种性能优良的汇率预测模型。

表1 PANK_EURUSD_D1 预测欧元兑美元汇率 $t+1$ 日线收益率的命中率

滑动窗口宽度 m	命中率(HR) $k=1$	$k=2$	$k=3$	$k=4$	$k=5$
10	0.67	0.75	0.77	0.76	0.72
15	0.73	0.76	0.78	0.75	0.77
20	0.75	0.76	0.76	0.76	0.73
25	0.74	0.72	0.75	0.73	0.80
30	0.74	0.76	0.77	0.77	0.77
35	0.70	0.69	0.74	0.72	0.76

5.2 PANK_HS300_D1 预测沪深 300 指数日线收益率

PANK_HS300_D1 预测模型针对沪深 300 指数日线收盘价数据,预测 $t+1$ 日线收益率,该模型可从(24)式具体化为

$$\begin{aligned}&\text{PANK\_HS300\_D1: } RR(t+1)\\&=NKNN\{AP^{*}[PCA(HS300\_D1\_DR(t,m))],k\}\end{aligned}\tag{27}$$

模型训练和测试的历史数据集由 2002 年 1 月 4 日至 2017 年 7 月 28 日期间的 3775 个交易日数据组成,前面时段的 3020 个数据点用于样本内训练,后面时段的 755 个数据点用于样本外检验。表 2 显示了 PANK_HS300_D1 预测模型的样本外检验预测命中率结果,其中在 $m=25$ 和 $k=3$ 时,取得了最高的命中率 0.80(80%)。该实证结果表明 PANK_HS300_D1 模型仍然具有良好的预测效能。

表2 PANK_HS300_D1 预测沪深 300 指数 $t+1$ 日线收益率的命中率

滑动窗口宽度 m	命中率(HR) $k=1$	$k=2$	$k=3$	$k=4$	$k=5$
10	0.62	0.58	0.66	0.65	0.76
15	0.59	0.65	0.66	0.67	0.66
20	0.64	0.65	0.65	0.65	0.68
25	0.75	0.75	0.80	0.79	0.78
30	0.77	0.70	0.71	0.76	0.77
35	0.69	0.73	0.75	0.76	0.72

5.3 与 KNN 和 AP + Nested KNN 预测模型的实证比较

为了进一步对 PANK 模型的预测效能进行评价,本文另外构建了两个具体的 KNN 预测模型和两个具体的 AP + Nested KNN 预测模型与之进行比较。

$$\begin{aligned}&\text{KNN\_EURUSD\_D1: } RR(t+1)\\&=KNN\{[EURUSD\_D1\_DR(t,m)],k\}\end{aligned}\tag{28}$$

$$\begin{aligned}&\text{KNN\_HS300\_D1: } RR(t+1)\\&=KNN\{[HS300\_D1\_DR(t,m)],k\}\end{aligned}\tag{29}$$

$$\begin{aligned}&\text{AP + NKNN\_EURUSD\_D1: } RR(t+1)\\&=NKNN\{AP[EURUSD\_D1\_DR(t,m)],k\}\end{aligned}\tag{30}$$

$$\begin{aligned}&\text{AP + NKNN\_HS300\_D1: } RR(t+1)\\&=NKNN\{AP[HS300\_D1\_DR(t,m)],k\}\end{aligned}\tag{31}$$

为了更直观地对六个具体模型的预测效能进行比较,我们选取了每个模型预测的最高命中率。表 3 显示了对比结果,PANK 预测模型不管是在外汇市场还是在股指上都有着最优的预测效果,由此可以说明 PCA、AP 和 Nested KNN 在 PANK 模型中都是有效果的,能够有效改进 KNN 模型的预测效果。表 3 中不同模型的预测效能对比结果表明 PANK 预测效能优于 AP + Nested KNN,而 AP + Nested KNN 又是优于 KNN 预测模型的。这正如 Krogh 和 Vedelsby[24] 证明的一样,"当构成组合预测模型的单一模型足够精确且足够多样化时,组合预测模型一定能获得比单一模型更好的预测效果。"

表3 PANK 模型与其他相关模型的预测效能比较

	具体预测模型	最高命中率(HR)
EUR/USD	KNN_EURUSD_D1	0.69
	AP + NKNN_EURUSD_D1	0.72
	PANK_EURUSD_D1	0.80
HS300	KNN_HS300_D1	0.68
	AP + NKNN_HS300_D1	0.74
	PANK_HS300_D1	0.80

6 结论与展望

本文提出的 PANK 模型,是一种集成 PCA、AP 和 Nested KNN 算法的金融时间序列预测的计算智能模型。从整体结构上看,该模型具有 PCA + AP 的特征提取过程和 Nested KNN 回归预测两大部分,是具有创新性的。而模型中的 Nested KNN 算法是本文针对 KNN 算法本身的缺陷,提出的一种嵌套式的 KNN 改进算法。该算法由三个函数组成了两层计算:(1)在 PCA + AP 输出的聚类中心集中进行计算,并找到最相似的聚类中心及所在类。(2)在第一层输出的聚类中进行计算,并找到最相似的 k 个最邻近元进行回归预测。这样的分层计算比原始的 KNN 算法具有更有效的分类效果和更快的运算速度,从而能够更有效地对金融时间序列进行回归预测。为了验证 PANK 模型的有效性,本文在预测欧元兑美元汇率和中国基准指数沪深 300 上进行了实证,对日线收益率进行了预测。实证结果表明 PANK 模型的预测性能明显优于 KNN 和AP + Nested KNN 预测模型,在每日时间框架内,

最佳命中率均达到 0.80(80%)。

在后续的研究中,我们的金融时间序列预测模型可以从以下方面进行改进研究:(1)将线性变换 PCA 换成一种更适合金融时间序列的非线性方法,比如"自编码器"。(2)将欧氏距离换成一个更适合金融时间序列相似性度量的方法,以提高模型预测性能。(3)将 KNN 换成更有效的非线性预测模型,比如"随机森林"。

参 考 文 献:

[1] Box G E P, Jenkins G M. Time series analysis: forecasting and control [M]. San Francisco: Holden-Day, 1970. 23-124.

[2] Engle R F. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation [J]. Econometrica, 1982, 50(4): 987-1007.

[3] Bollerslev T. Generalized autoregressive conditional heteroskedasticity [J]. Journal of Econometrics, 1986, 31(3): 307-327.

[4] 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.

[5] 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: 831-846.

[6] Galeshchuk S. Neural networks performance in exchange rate prediction [J]. Neurocomputing, 2016, 172: 446-452.

[7] Zhang N N, Lin A J, Shang P J. Multidimensional k-nearest neighbor model based on EEMD for financial time series forecasting [J]. Physica A: Statistical Mechanics and its Applications, 2017, 477: 161-173.

[8] Cover T, Hart P. Nearest neighbor pattern classification [J]. IEEE Transaction on Information Theory, 1967, 13(1): 21-27.

[9] Bannayan M, Hoogenboom G. Weather analogue: a tool for real-time prediction of daily weather data realizations based on a modified k-nearest neighbor approach [J]. Environmental Modelling and Software, 2008, 23(6): 703-713.

[10] 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.

[11] Pearson K F R S. On lines and planes of closest fit to systems of points in space [J]. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 1901, 2(11): 559-572.

[12] Hotelling H. Analysis of a complex of statistical variables into principal components [J]. Journal of Educational Psychology, 1933, 24(7): 498-520.

[13] 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.

[14] Frey B J, Dueck D. Clustering by passing messages between data points [J]. Science, 2007, 315(5814): 972-976.

[15] 王开军,张军英,李丹,等. 自适应仿射传播聚类 [J]. 自动化学报,2007,33(12):1242-1246.

[16] Arroyo J, Mate C. Forecasting histogram time series with k-nearest neighbours methods [J]. International Journal of Forecasting, 2009, 25(1): 192-207.

[17] Ghaderyan P, Abbasi A, Sedaaghi M H. An efficient seizure prediction method using KNN-based undersampling and linear frequency measures [J]. Journal of Neuroscience Methods, 2014, 232: 134-142.

[18] Schubert A L, Hagemann D, Voss A, et al.. Evaluating the model fit of diffusion models with the root mean square error of approximation [J]. Journal of Mathematical Psychology, 2017, 77: 29-45.

[19] Myttenaere A, Golden B, Grand, B L, et al.. Mean absolute percentage error for regression models [J]. Neurocomputing, 2016, 192: 38-48.

[20] 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.

[21] Chen T. Applying a fuzzy and neural approach for forecasting the foreign exchange rate [J]. International Journal of Fuzzy System Applications, 2011, 1(1): 36-48.

[22] Sermpinis G, Theofilatos K, Karathanasopoulos A, et al.. Forecasting foreign exchange rates with adaptive neural networks using radial-based functions and particle swarm optimization [J]. European Journal of Operational Research, 2013, 225(3): 528-540.

[23] 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.

[24] Krogh A, Vedelsby J. Neural network ensembles, cross validation, and active learning [A]. In Tesauro G, Touretzky D S, Leen T K, eds. Advances in Neural Information Processing Systems 7 [C]. MIT Press, Cambridge, 1995. 231-238.