

基于距离和时序的层次粒度支持向量回归机

王 珏^{1,2}, 乔建忠¹, 林树宽¹

(1. 东北大学 信息科学与工程学院, 辽宁 沈阳 110819 ; 2. 沈阳农业大学 信息与电气工程学院, 辽宁 沈阳 110866)

摘 要 : 针对目前粒度支持向量回归机的粒划算法只考虑了距离因素, 引入时序因素, 提出适用于金融时间序列的基于距离和时序的层次粒度支持向量回归机(DTHGSVR). 该方法首先将训练样本通过核函数映射到高维空间, 并在该特征空间中进行初始粒划. 然后, 通过衡量样本粒与当前回归超平面的距离以及当前样本粒时序的综合因素, 找到含有较多回归信息的粒, 并通过计算其半径、密度及时序信息进行深层次的动态粒划. 如此循环迭代, 直到没有粒需要进行深层划分为止. 最后, 对不同层次的粒进行回归训练. 采用提出的基于距离和时序因素的层次粒度支持向量回归机对基金净值进行预测, 实验结果表明回归的泛化性有所提高.

关 键 词 : 粒度支持向量回归; 时序; 金融时间序列; 预测; 泛化性

中图分类号: TP 181

文献标志码: A

文章编号: 1005-3026(2016)07-0942-05

Hierarchical Granular Support Vector Regression Based on Distance and Temporal

WANG Jue^{1,2}, QIAO Jian-zhong¹, LIN Shu-kuan¹

(1. School of Information Science & Engineering, Northeastern University, Shenyang 110819, China ; 2. College of Information and Electric Engineering, Shenyang Agricultural University, Shenyang 110866, China. Corresponding author: QIAO Jian-zhong, E-mail: kelly_wang13@126.com)

Abstract : Only distance factor is considered in the granular algorithm of granular support vector regression. Temporal factor was introduced simultaneously in granular algorithm. Hierarchical granular support vector regression based on distance and temporal factors(DTHGSVR) was proposed which is applicable for financial time series. The training samples were mapped into the high-dimensional space by mercer kernel, and the samples were divided into some granules initially. Then, the granules which have more regression information was found by measuring the distances between the granules and regression hyperplane and the granule's temporal factor. By computing the radius, density of granules and the temporal factor, the deeper hierarchical granulation process was executed until no granules was needed to be granulated. Finally, those granules in different granulation levels were trained by SVR. Fund net was forecast by the hierarchical granular support vector regression based on distance and temporal factors. Experimental results showed the generalization performance of regression had been improved.

Key words : granular support vector regression ; temporal ; financial time series ; forecasting ; generalization

支持向量机是基于统计学习的 VC 维理论和结构风险最小原理基础上的, 它能很好地解决小样本、非线性以及高维模式识别问题. 近年来支持向量机越来越多地被应用到回归问题中. 求解回归问题最终归结为求解一个凸二次规划问题, 当

数据规模增大时, 数据的训练效率急速下降, 这是制约其应用的主要原因. 很多学者在相关方面进行了深入的研究^[1-6], 并提出了一些改进方法, 其中较有效的一种方法为粒度支持向量机. 文献[7]首次将粒度计算理论同支持向量机相结合,

提出粒度支持向量机(GSVM). GSVM 的基本思想是在原始样本空间中对样本以某种聚类算法划分为若干粒 ,然后使用部分重要样本代替粒进行学习 ,从而得到最终分类器 .粒划算法的好坏直接决定了粒度支持向量机(GSVM)的精度 .此后许多学者对粒划算法进行研究 .文献[8]提出了基于聚类的 GSVM ,根据样本间的相似度将数据集划分为多个粒 ,对粒进行学习 ,从而得到最终的分类或回归面 .文献[9]提出了基于样本与近似最优超平面距离的粒划算法 .文献[10]提出了动态粒度支持向量机 ,根据密度和半径来确定信息粒的重要程度 ,对重要程度不同的信息粒进行不同层次划分 .上述文献提出的粒划算法 ,在粒划过程中只考虑了距离因素 .本文在文献[10]的基础上 ,将时序信息引入粒划过程 ,综合距离及时序信息衡量粒的重要程度 ,对重要的粒进行较多层次的划分 ,对次要的粒进行较少层次的划分 .最后根据不同层次、不同细化程度的粒训练得到回归平面 ,从而进一步提高了 GSVM 的泛化性 .

1 粒度支持向量回归机

粒度支持向量回归机(GSVR) ,将粒度计算理论和统计学习理论结合起来 ,总体思想是通过某种粒划算法将样本集分为多个组 ,在每组样本中选取中心样本进行训练得到回归超平面^[11] .这样 SVR 的训练样本减少 ,从而提高训练效率 .这样做可能会使经过聚类粒划的样本集的分布与原始样本集完全不同 ,从而导致泛化性能的降低 .针对这些问题 ,文献[10]提出了改进方法 ,在核空间中 ,粒划过程在不同层次下进行划分代替 ,这样做能使重要的样本尽可能多地留在样本集中 ,从而尽量保证样本粒划前后分布的一致性 .粒划后 ,样本集的规模减小 ,训练过程缩短 ,学习的泛化能力提高 .本文在文献[10]的基础上 ,将时序信息引入粒划过程 ,提出适用于金融时间序列的基于距离和时序因素的层次粒度支持向量回归机 .

2 基于距离和时序因素的层次粒度支持向量回归机(DTHGSVR)

依据支持向量回归机的几何意义 ,位于回归间隔边界上的点最有可能成为最终的支持向量 ,这些样本对回归平面很重要 ;而位于分类间隔内部的点成为支持向量的可能性较小 ,对回归平面的影响较小 .目前大多粒度支持向量机的粒划算

法 ,都是依据距离因素进行粒划 ,而对于时间序列来说 ,离预测点越近的样本点 ,对最终回归平面的贡献越大 .基于这样的思想 ,参考文献[10] ,对粒划算法进行了改进 ,既考虑样本点的距离因素又考虑了样本点的时序特征 ,提出了适合金融时间序列的基于距离和时序因素的层次粒度支持向量回归机 .

2.1 相关概念及定义

参考文献[10]中定义 ,假定原始训练集 $T = (X , Y) = \{ \{ (x_i , y_i) \} \mid i = 1 , \dots , l \}$, $x_i \in \mathbf{R}^n$, $y_i \in \mathbf{R}$,经过非线性映射 φ ,样本在高维空间 \mathbf{R}^n 中表示为 $T = \{ \{ (\varphi(x_i) , y_i) \} \mid i = 1 , \dots , l \}$,将样本划分为 k 个粒 G_1 , \dots , G_k ,其中 , $G_i = \{ \{ (\varphi(x_{ij}) , y_{ij}) \} \mid i = 1 , \dots , k \}$, $j = 1 , \dots , n_i$ (n_i 为第 i 个粒中样本个数) .将每个粒看作一个超球 ,其中心和半径定义如下 .

定义 1 (核空间粒超球的中心及半径)粒划后的样本集形成的每一个 N 维样本粒 X_i 称为一个粒超球(粒超球记作 X_i) ,其中 μ_i 称为中心(粒心) , r_i 称为半径 ,计算方法如下 :

$$\mu_i = \frac{1}{n_i} \sum_{p=1}^{n_i} \varphi(x_p) = \sqrt{\frac{1}{n_i^2} \left(\sum_{p=1}^{n_i} \varphi(x_p) \right)^2} = \frac{1}{n_i} \sqrt{\sum_{p=1}^{n_i} \sum_{q=1}^{n_i} K(x_p , x_q)} , \tag{1}$$

$$r_i = \max_{x_s \in G_i} (\varphi(x_s) - \mu_i) = \max_{x_s \in G_i} \sqrt{ (\varphi(x_s))^2 - 2 \varphi(x_s) \mu_i + \mu_i^2 } = \max_{x_s \in G_i} \sqrt{ K(x_s , x_s) - \frac{2}{n_i} \sum_{p=1}^{n_i} K(x_s , x_p) + \frac{1}{n_i^2} \sum_{p=1}^{n_i} \sum_{q=1}^{n_i} K(x_p , x_q) } . \tag{2}$$

根据定义 1 , N 维空间中任一样本 $\varphi(x_j)$ 到第 i 个粒超球 G_i 的距离为

$$d(\varphi(x_j) , G_i) = \sqrt{ K(x_j , x_j) - \frac{2}{n_i} \sum_{p=1}^{n_i} K(x_j , x_p) + \frac{1}{n_i^2} \sum_{p=1}^{n_i} \sum_{q=1}^{n_i} K(x_p , x_q) } . \tag{3}$$

定义 2 (粒到超平面的距离) N 维空间的粒 G_i 到超平面 $f : y = w \cdot \varphi(x) + b$ 的距离为

$$d(G_i , f) = \frac{\frac{1}{n} \sum_{k=1}^{n+1SV} \sum_{j=1}^{n+1SV} a_j y_j K(x_j , x_k) + b}{\sqrt{\sum_{j=1}^{n+1SV} \sum_{k=1}^{n+1SV} a_j a_k y_j y_k K(x_j , x_k)}} - r_i . \tag{4}$$

其中 , SV 是支持向量集合 .

定义 3 (信息粒)假设存在粒心和半径分别

为 μ_i 和 r_i 的粒 G_i 近似的回归超平面为 $f: y = w \cdot \phi(x) + b$ 若粒 G_i 到回归平面 f 的距离 $d(G_i, f)$ 大于或等于 $\eta - 2r_i$ (其中 η 是 SVR 回归间隔), 则粒 G_i 为信息粒。

由信息粒的定义可知,若粒与超平面 f 的回归间隔区域有重叠或在间隔之外,此粒为信息粒。否则,粒 G_i 为非信息粒。由于传统 SVR 模型中支持向量在超平面间隔边界上,且其与超平面的距离为 η 。因此,信息粒包含支持向量的可能性较大,非信息粒包含支持向量的可能性较小。

定义 4 (粒密度)假设存在粒 $G_i = \{x_{ij}\}$ ($j = 1 \dots n_i$) 其粒心和半径分别为 μ_i 和 r_i , 所含样本数为 n_i , 则粒 G_i 的密度 ρ_i 定义为

$$\rho_i = \frac{n_i}{\sum_{j=1}^{n_i} d(\mu_i, x_{ij})} = \frac{n_i}{\sum_{j=1}^{n_i} \sqrt{\frac{1}{n_i} \sum_{p=1}^{n_i} \sum_{q=1}^{n_i} k(x_p, x_q) - \frac{2}{n_i} \sum_{p=1}^{n_i} k(x_p, x_{ij}) + \sum_{q=1}^{n_i} k(x_{ij}, x_{ij})}} \quad (5)$$

定义 5 (动态粒化个数)假设存在第 L 层的信息粒 $G'_{L,j}$, 针对信息粒 $G'_{L,j}$ 在第 L 层的动态粒划个数为

$$k'_{L,j} = \left\lfloor \frac{r'_L \times \rho'_{L,j}}{d} \right\rfloor \quad (6)$$

其中 d 为动态粒化参数,用来控制数据集的粒划进程,通过网络搜索的方式设置 d 的取值。通过反复试验,当训练集样本个数大于 100 时,动态粒化参数分别取 $[1.5, 2, 2.5]$ 进行搜索;当训练集样本个数小于 100 时, d 分别取 $[1, 1.25, 1.5]$ 进行搜索。直到所有信息粒的粒划个数均为 1 时,动态粒划过程停止。

2.2 基于距离和时序因素的粒划算法

2.2.1 基本思想

粒化算法的总体思想是,在进行粒划的过程中尽可能将重要的样本保留在样本集中。相比简单使用核空间邻近法进行粒划,本文提出的粒化算法首先按照核空间邻近法进行粒化,然后将粒分为信息粒和非信息粒。对信息粒,根据其距离因素(密度、半径)和时序因素再进行粒划,重复粒化过程,直至不能再粒划。

2.2.2 时序信息的表示

对于金融时间序列,往往离预测点近的样本对预测结果影响较大。为了使时序信息在粒划过程中得到体现,本文引入 $\frac{1}{1 + \exp(a - 2ax/n)}$ 这一

指数函数来表示样本的时序性。图 1 是 $y(x)$ 和 a 的函数关系。

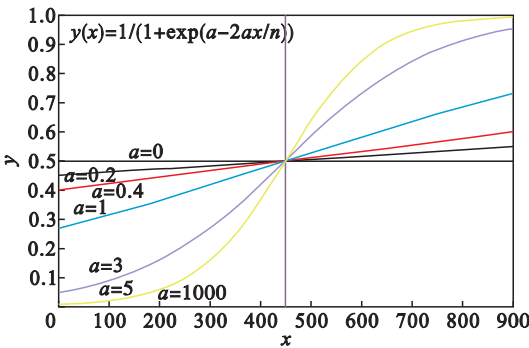


图 1 $y(x)$ 和 a 的函数关系
Fig. 1 Function relationship between $y(x)$ and a

$a > 0$ 时为单调递增函数,并且随着时间后移逐渐增大,它的范围在 0 和 1 之间。这样对于离预测点较近的样本点权重较大。

定义 6 (动态粒划个数)假设存在第 L 层的信息粒 $G'_{L,j}$, 针对信息粒 $G'_{L,j}$ 在第 L 层的动态粒划个数为

$$k'_{L,j} = \left\lfloor \frac{r'_L \times \rho'_{L,j}}{d} \times \frac{1}{1 + \exp(a - 2 \times a \times i/n)} \right\rfloor \quad (7)$$

2.2.3 算法

根据算法的基本思想,基于距离和时序因素的层次粒度支持向量回归机的算法用简单流程图描述如下。

算法 1 基于核的粒划分见图 2。

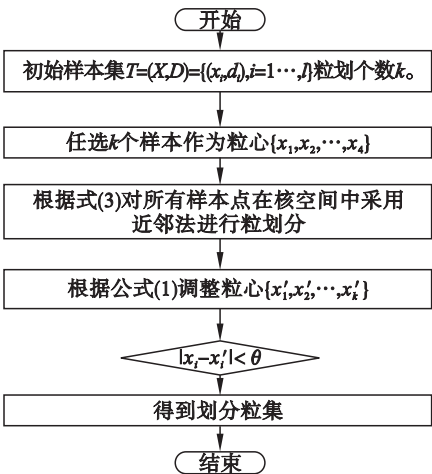


图 2 基于核的粒划分
Fig. 2 Granulation division based on kernel

算法 2 基于距离和时序因素的层次粒度支持向量回归机算法流程见图 3。此算法在计算某粒划层次上某个信息粒的动态粒划个数的算法中,采用公式 (7)。式 (7) 中,添加了时序因素

$$\frac{1}{1 + \exp(a - 2ax/n)}$$

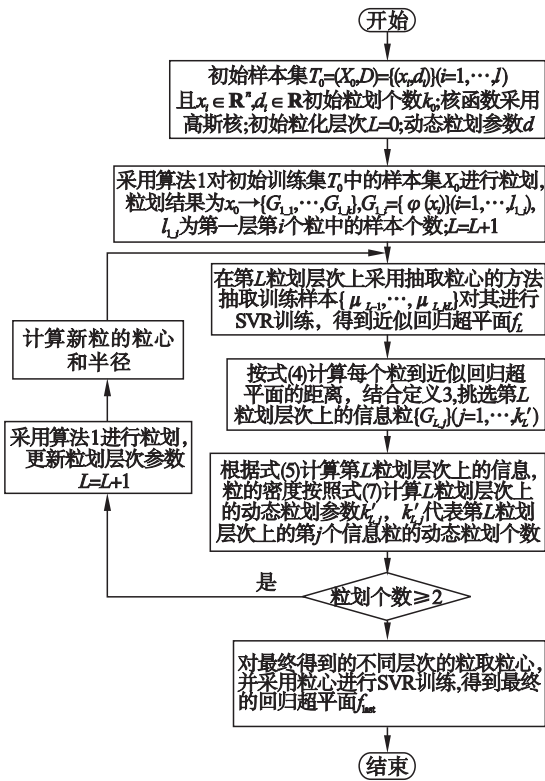


图 3 基于距离和时序因素的层次粒度支持向量回归机
Fig. 3 Hierarchical granular support vector regression
based on distance and temporal factors

3 实验结果及分析

本文使用 2006 年 2 月 1 日到 2007 年 12 月 28 日时间段内共 90 周金泰基金数据。影响基金的因素如基金成交额、行业投资集中度、基金换手率及我国居民消费价格指数等数据是由中国基金网、国家统计局年鉴及各基金管理公司网站的相关数据整理获得^[12]。用前 80 周的数据进行训练, 对后 10 周的数据进行预测。

基于距离和时序的层次粒度支持向量回归机 粒化是关键的一步, 它最终决定了实际训练集的规模以及保留回归信息的多少。通过试验可知, d 增大, 动态粒个数及粒化层次均减小。当 $d = 1.25$ 时 k_0 分别取 5, 10, 15, 20, 25 时, 粒化层次分别为 6, 6, 6, 4, 4, 最终动态粒个数分别为 58, 56, 55, 54, 54。

本文提出的基于距离和时序因素的层次粒度支持向量回归机的训练时间基本与基于动态粒划算法的支持向量回归机一致, 但略大于基于聚类算法的粒度支持向量回归机, 远远小于支持向量回归机。由图 4 可以看出, 使用基于距离和时序因素的层次粒度支持向量回归机的泛化能力有所提

高, 这是因为本文提出的粒划算法综合考虑了距离和时序因素, 对样本的重要性判断更加全面, 使得重要的样本尽可能多地留在训练集中。

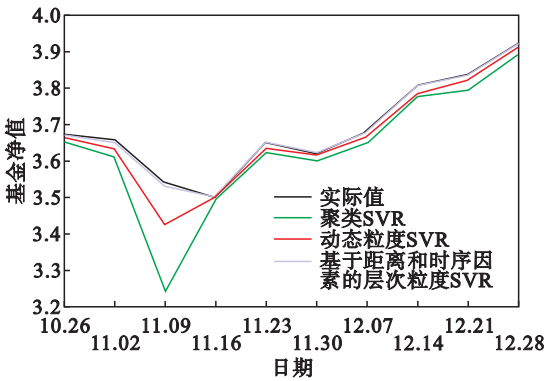


图 4 10.26 ~ 12.28 日的基金真实值及使用聚类 SVR、
动态粒度 SVR(DGSVR)、基于距离和时序因素
的层次粒度 SVR(DTHGSVR)的预测值
Fig. 4 Real value and the predict value of funds
by CSVR, DGSVR and DTHGSVR
from 10.26 to 12.28

4 结 论

针对支持向量机进行大量数据回归效率低的瓶颈问题, 本文提出了一种基于距离和时序因素的层次粒度支持向量回归机。该方法综合距离与时序因素, 对信息粒的重要程度进行判断。对重要的信息粒进行多层次的粒化, 而对不重要的粒不再进行粒化。这样, 在粒化过程完成时, 能够将重要的样本尽可能多地保留下来, 从而为提高算法的泛化性能奠定基础。在使用基于距离和时序因素的层次粒度支持向量回归模型时, 最终的回归结果与初始粒化参数 k_0 息息相关, 本文中采用的是试验法, 对初始粒化参数 k_0 的选取算法在今后还需深入研究。

参考文献：

[1] Zeng Z Q, Gao J. Simplified support vector machine based on reduced vector set method[J]. *Journal of Software*, 2007, 18 (11) : 2719 - 2727.

[2] Li D C, Fang Y H. An algorithm to cluster data for efficient classification of support vector machines[J]. *Expert Systems with Application*, 2008, 34(3) : 2013 - 2018.

[3] Hao P Y, Chiang J H, Tu Y K. Hierarchically SVM classification based on support vector clustering method and its application to document categorization[J]. *Expert Systems with Applications*, 2007, 33(3) : 627 - 635.

[4] Nath J S, Shevade S K. An efficient clustering scheme using support vector methods[J]. *Pattern Recognition*, 2006, 39 (8) : 1473 - 1480.