Support Vector Machine Based Methods for

金融工程的支持向量机方法

Financial and Engineering Problems

曹丽娟　王小明　著

3

上海财经大学出版社

国家自然科学基金项目研究成果

金融工程的支持向量机方法

Support Vector Machine Based Methods for Financial and Engineering Problems

曹丽娟　王小明　著

上海财经大学出版社

图书在版编目（CIP）数据

金融工程的支持向量机方法/曹丽娟、王小明著.
—上海：上海财经大学出版社，2007.2
ISBN 978-7-81098-761-5/F・707

Ⅰ.金…　Ⅱ.①曹…②王…　Ⅲ.向量计算机-算法理论-应用-金融学-英文
Ⅳ.F830-39

中国版本图书馆 CIP 数据核字（2006）第 141336 号

□责任编辑　姜　勇
□封面设计　周卫民
□版式设计　孙国义

JINRONG GONGCHENG DE ZHICHI XIANGLIANGJI FANGFA
金融工程的支持向量机方法
曹丽娟　王小明　著

上海财经大学出版社出版发行
（上海市武东路 321 号乙　邮编 200434）
网　　址：http：//www.sufep.com
电子邮箱：webmaster@sufep.com
全国新华书店经销
江苏省句容市排印厂印刷装订
2007 年 2 月第 1 版　2007 年 2 月第 1 次印刷

787mm×960mm　1/16　15.25 印张　324 千字
印数：0 001—1 200　定价：29.00 元

序

支持向量机（Support Vector Machine，简称为 SVM）是在 20 世纪 90 年代由 V. Vapnik 等人研究并迅速发展起来的一种基于统计学习理论（Statistical Learning Theory）的机器学习算法。它通过寻求结构风险最小化来实现实际风险最小化，从而在样本量较少的情况下也能获得良好的学习效果。支持向量机算法是一个二次优化问题，因此，能保证所得到的解是全局最优的解。支持向量机具有完备的理论基础（统计学习理论）和出色的应用表现，正成为神经网络之后，机器学习领域中新的研究热点。以往困扰机器学习方法的很多问题，如模型选择与学习问题、非线性和维数灾难问题、局部极小问题等，在这里都得到了一定程度上的解决。它已经应用在模式识别、函数回归和概率密度估计等方面。

本书的一部分内容是系统介绍支持向量机的理论基础和应用方法。大部分的篇幅则是提出我们的新方法去处理金融工程中的大量问题。

本书由两部分组成。第一部分集中讨论用支持向量机解决时间序列的预测问题。时间序列的预测是回归研究中最常见的问题之一。第二部分研究用支持向量机解决分类问题与奇异点检测问题。

第一部分“回归估计中的支持向量机”包含 7 章的内容。

第一章总结了用支持向量机解决回归问题的已有的文献。通过文献的综述，我们可以知道运用支持向量机解决回归问题的重要特征和分析方向。

第二章探讨了用支持向量机来预测金融时间序列。BP 中性神经网络被用作与支持向量机相比较的一种方法。本章详细探讨了特征萃取、奇异点剔除与功能评价指标的选择问题，同时，也探讨了支持向量机的功能特征。本章的研究表明，支持向量机在金融时间序列中是一种非常有前途的分析方法。

第三章分析了在支持向量机中的特征选择问题。本书建议利用两种方法：显著性分析法与基因运算法，借助这两种方法我们可以在支持向量机中选择相关的重要特征。显

著性分析法通过分析输出对输入的敏感性来检测重要的输出的特征。基因运算法是一种选择重要特征的全局优化方法。我们的经验结果显示，显著性分析法与基因运算法在支持向量机中都能够有效地获得重要的特征。与此同时，与显著性分析法相比，基因运算法能够选择一组更为有效的特征。借助特征的选择，支持向量机能够得出更好的预测精度。

第四章考察特征的提取问题，作者建议在支持向量机中借助主成分特征分析、核主成分和独立主成分分析来进行特征的提取。主成分将已有的相关特征转化为独立不相关的特征。核主成分是一种非线性的主成分分析法。独立主成分分析法将已有的输入线性转换为统计意义上相互独立的特征。实验的结果表明，借助主成分特征分析、核主成分和独立主成分分析进行特征萃取的时候，支持向量机的预测精度要高于不进行特征萃取的情形。在三种分析法中，独立主成分分析的精度是最高的。

第五章研究三种动态的支持向量机，即 c-ASVMs、ε-DSVMs 与 ASVMs，被应用于非平稳的时间序列分析中。放弃在标准的支持向量机中的固定系数，我们用适应性参数去处理数据中的结构变化问题，与较远的训练数据相比，较近的训练数据被赋予了更高的权数。深入的实验结果表明，在时间序列的预测领域，动态的支持向量机要比标准的支持向量机有更高的预测精度。

第六章建立一个将支持向量机与自组织方法相结合的混合系统，目的是用它去预测非平稳的时间序列。基于“分割与合并”的思想，我们将一个复杂的系统分割成一系列简单的问题，每一个简单的问题都可以借助支持向量机来解决。混合系统等价于广义的专家系统。混合系统的大小自动取决于树结构的路径。我们将混合的系统应用于 13 种不同的时间序列的预测问题上。在所有的情形，与标准支持向量机相比，混合系统都能够产生更好的结果，更快的收敛以及借助更少的支持向量。

第七章提出一个训练支持向量机的并行序列最小优化方法。我们是借助信息传递接口思想来开发并行序列最小优化方法的。具体地说，并行序列最小优化方法首先将整个训练数据集分割为更小的子集，然后用多个 CPU 处理器去分析每一个分隔的数据子集。实验结果表明，在运用了多个处理器之后，并行序列最小优化方法能够大大提高数据处理的速度。

第二部分“分类与奇异性侦察中的支持向量机”包含 3 章的内容。

第八章用支持向量机来解决债券等级的评估问题。债券等级的评估是一个多级分类问题。我们应用了支持向量机解决多级分类问题的常用三种方法，即“一对所有”、“一对一”以及“有向的非循环图”(DSVM)。我们将支持向量机的分析效果与后向神经网络方法、逻辑回归方法和有序概率回归方法作了比较。我们在固定收益的投资证券(FISD) 数据库中建立了一个真实的美国债券数据库。实验表明，支持向量机的结果要明显好于前三种基础性的预测方法。在三种以支持向量机为基础的方法中，DAGSVM

的结果最好。

第九章提出了一个改进的支持向量机奇异性侦察方法。我们用它将正常数据与奇异的数据分隔开来。在Scholkopf等人（2001）原创性的论文中，他们提出了一种奇异性侦察方法。我们改进了他们的方法。我们的方法考虑了训练集中的奇异数据，并将正常的训练数据集与异常的训练数据集区分了开来。通过检验人为与真实的数据集，实验表明，与Scholkopf等人（2001）的方法相比，我们的方法显著改善了模型的结果。原来的方法对奇异性的数目比较敏感，随着数目的增加，侦察的绩效会降低。

最后一章总结了所有的结果并对支持向量机的研究作了展望。

Preface

Developed based on the statistical learning theory, support vector machines (SVMs) are a novel and robust technique for solving various data mining problems such as classification, regression estimation and novelty detection.

This book is composed of two sections. The first section focuses on applying SVMs to solve one type of regression estimation problems: time series forecasting. The second section deals with applying SVMs to solve the classification and novelty detection problems.

In the first section, there are 7 chapters. In Chapter 1, a comprehensive review of the research in the application of SVMs for regression estimation is presented with the purpose of identifying the key characteristics of SVMs in regression estimation and research directions in this area. In Chapter 2, the feasibility of applying SVMs to financial time series forecasting are examined with a back-propagation (BP) neural network used as a benchmark. Issues such as indicators extraction, outliers removal, performance measure criteria selection, training SVMs, and investigating the functional characteristics of SVMs are addressed. This study shows that SVMs provide a promising alternative tool to the BP neural network in financial time series forecasting. In Chapter 3, the issue of feature selection is discussed. Two methods, saliency analysis (SA) and genetic algorithm (GA), are proposed to be used with SVMs for selecting important features. SA identifies the important features by measuring the sensitivity of the output of SVMs to the inputs, while GA selects a feature subset in a global optimization way. Our experimental study shows that both SA and GA are effective in SVMs for selecting important features. GA can select a better feature subset than SA in the real financial tasks. By using the selected features from SA and GA, SVMs can achieve higher generalization performance and converge faster. In Chapter 4, the issue of feature selection is further discussed. Principal component analysis (PCA), kernel principal component analysis (KPCA) and independent component analysis (ICA) are proposed to SVM for feature extraction. PCA linearly transforms the original inputs into new uncorrelated features. KPCA is a nonlinear PCA developed by using the kernel method. In ICA, the original inputs are linearly transformed into features which are mutually statistically independent. The experiment shows that SVM by feature extraction using PCA, KPCA or ICA can perform better than that without feature extraction. Furthermore, among the three methods, there is the best performance in KPCA feature extraction, followed by ICA feature extraction. In Chapter 5, three dynamic SVMs, namely, c-ASVMs, ε-DSVMs, and ASVMs, are proposed by incorporating the non-stationary characteristic of time series into SVMs. Instead of fixed parameters in the standard SVMs, adaptive parameters which will place more weights on recent training data points than distant training data points are used to deal with structural changes in the data. An extensive experimental study demonstrates that the dynamic SVMs are more effective in time series forecasting than the standard SVMs. In Chapter 6, a hybrid system formed by combining SVMs with SOM is developed to forecast non-stationary time series. Generated using the idea of "divide-and-conquer" —dividing a complex problem into simpler problems whose solutions are combined to yield a solution to the complex problem, the hybrid system is equivalent to the generalized mixture of experts (ME). Its size is automatically determined by a tree-structured architecture. The hybrid system is applied to thirteen different time series forecasting problems. In all the cases, the hybrid system generalizes better, converges faster, and uses fewer support vectors than the single SVMs models. Chapter 7 proposes one parallel implementation of SMO for training SVM. The parallel SMO is developed using message passing interface (MPI). Specifically, the parallel SMO first partitions the entire training data set into smaller subsets and then simultaneously runs multiple CPU processors to deal with each of the partitioned data sets. Experiments show that there is great speedup on the adult data set and the MNIST data set when many processors are used. There are also satisfactory results on the Web data set.

The second section contains another 3 chapters. Chapter 8 proposes using support vector machine (SVM) to deal with bond rating which is actually a multi-class classification problem. The three commonly used methods for solving multi-class classification problems in SVM, "one-against-all", "one-against-one", and directed acyclic graph

SVM (DAGSVM) are used. The performance of SVM is compared with several benchmarks including Backpropagation (BP) neural network, logistic regression and ordered probit regression. One real U. S. bond data is collected based on the Fixed Investment Securities database (FISD) and the Compustat database. The experiment shows that SVM significantly outperforms the benchmarks. Among the three SVM based methods, there is the best performance in DAGSVM. Furthermore, an analysis of features shows that the generalization performance of SVM can be further improved by performing feature selection. Chapter 9 proposes a modified support vector novelty detector (SVND) for novelty detection which addresses the problem of detecting outliers from normal data patterns. While the original SVND (Schölkopf et al. , 2001) attempts to estimate a function to separate the region of normal data patterns from that of outliers based on normal data patterns, the modified SVND generalizes it to take into account the outliers in the training set by separating both the normal training data patterns and the outliers from the origin with maximal margin. By examining artificial and real data sets, the experiment shows that there is a significant improvement in the performance of the modified SVND in comparison with the original SVND. Furthermore, the original SVND is sensitive to the outliers, with the performance deteriorating when there are outliers used in the training set. Finally, Chapter 10 gives conclusions and recommendations to the researchers in SVMs.

Nomenclatures

l	training set size
$(*)$	variables with and without $*$
n	dimension of original input space;also real number
k	general real number
N	dimension of high dimensional feature space;natural number
$x \in X^n$	input and input space
$x_i \cdot x_j$	inner product between x_i and x_j
$y \in Y$	output and output space
$\bar{y}$	mean of output values
$\hat{y}$	predicted output value
w	weight vector
b	bias
$f(x)$	general real-valued function
$a_i^{(*)}$	Lagrange multipliers
$\xi_i^{(*)}$	slack variables
ξ	vector of all slack variables
h	VC dimension
H	high dimensional feature space
$\phi(x): X \rightarrow H$	mapping to high dimensional feature space
$L(y, f(x))$	loss function
$K(x_i, x_j)$	kernel function
$R(f)$	generalization error
$R_{emp}(f)$	empirical error

L	Primal Lagrangian
W	Dual Lagrangian
ε	tube size of the ε-insensitive loss function
C	regularization constant
c	constant value
θ	testing accuracy threshold
γ	margin
m	arbitrary delay
τ	embedding dimension
R	radius of the ball containing the data
$\Vert \; \Vert$	Euclidean distance
$\Vert \; \Vert_p$	p-norm
d	distance
η	learning rate
t	time index
δ	confidence,also kernel width
p	probability
p_i	parameter
VC	Vapnik-Chervonenkis
BP	back-propagation
SOM	self-organizing feature map
ERM	empirical risk minimization
SRM	structural risk minimization
QP	quadratic programming
KKT	Karush-Kuhn-Tucker
ME	mixture of experts
GA	genetic algorithm
SA	saliency analysis
DLS	discounted least squares
OLS	ordinary least squares

目　录

序　1

数学符号表　1

第一部分　回归估计中的支持向量机

第一章　回归估计中的支持向量机：一项研究　3

1.1　回归估计中支持向量机的理论　3

1.2　支持向量机中的训练算法　10

1.3　支持向量机的方法论分析　14

1.4　支持向量机的应用与效果　16

1.5　小结与未来的研究工作　20

第二章　金融时间序列预测中的支持向量机　23

2.1　引论　23

2.2　金融预测理论综述　25

2.3　数据集与数据的预处理方法　28

2.4　效果检测标准　31

2.5　实验结果　32

2.6　小结　44

第三章　支持向量机中的特征选择　47

3.1　引论　47

3.2　神经网络回归元的特征选择技术　50

3.3　特征选择方法　52
3.4　实验结果　57
3.5　小结　68

第四章　PCA、KPCA 与 ICA 在支持向量机中降低维数方面的比较　69
4.1　引论　69
4.2　特征萃取方法　71
4.3　回归估计中的支持向量机理论　75
4.4　实验结果　76
4.5　小结　84

第五章　动态的支持向量机　86
5.1　引论　86
5.2　贴现的最小二乘法　88
5.3　三种动态的支持向量机　89
5.4　实验结果　94
5.5　相关的研究工作　107
5.6　小结　107

第六章　时间序列预测中的复合系统　109
6.1　引论　109
6.2　自组织特征图　113
6.3　设计与系统学习算法　115
6.4　实验结果　116
6.5　相关的研究工作　135
6.6　小结　135

第七章　训练支持向量机中的并行序列最小优化方法　137
7.1　引论　137
7.2　简论改进的序列最小优化方法　139
7.3　并行的序列最小优化　141
7.4　实验结果　143
7.5　小结　152

第二部分　分类与奇异性侦察中的支持向量机

第八章　支持向量机在债券评级中的应用　155

8.1　引论　155

8.2　债券评级问题　156

8.3　支持向量机分类　157

8.4　实验结果　160

8.5　小结　168

第九章　改进的支持向量奇异性侦察方法　170

9.1　引论　170

9.2　支持向量奇性侦察方法　172

9.3　改进的支持向量奇性侦察方法　173

9.4　实验结果　175

9.5　小结　181

第十章　结论与展望　183

参考文献　187

附录 A　满足 SRM 的 SVMs 回归　210

附录 B　数据集　212

附录 C　改进的 SMD 的程序编码　215

附录 D　改进的 SMD　221

Contents

Preface 1

Nomenclatures 1

SECTION 1 SUPPORT VECTOR MACHINES FOR REGRESSION ESTIMATION

Chapter 1 SVMs in Regression Estimation: a Survey 3

1.1 Theory of SVMs in Regression Estimation 3

1.2 Training Algorithms of SVMs 10

1.3 Methodologies 14

1.4 Applications and Performance 16

1.5 Conclusions and Future Works 20

Chapter 2 SVMs in Financial Time Series Forecasting 23

2.1 Introduction 23

2.2 Literature Review in Financial Forecasting 25

2.3 Data Sets and Data Preprocessing 28

2.4 Performance Criteria 31

2.5 Experimental Results 32

2.6 Conclusions 44

Chapter 3 Feature Selection in SVMs 47

3. 1 Introduction 47
3. 2 Feature Selection Techniques for Neural Network Regressors 50
3. 3 Methods of Feature Selection 52
3. 4 Experimental Results 57
3. 5 Conclusions 68

Chapter 4 A Comparison of PCA, KPCA and ICA for Dimensionality Reduction in SVMs 69
4. 1 Introduction 69
4. 2 Feature Extraction Methods 71
4. 3 Theory of Support Vector Machine (SVM) for Regression Estimation 75
4. 4 Experimental Results 76
4. 5 Conclusions 84

Chapter 5 Dynamic SVMs 86
5. 1 Introduction 86
5. 2 Discounted Least Square (DLS) 88
5. 3 Three Dynamic SVMs 89
5. 4 Experimental Results 94
5. 5 Related Work 107
5. 6 Conclusions 107

Chapter 6 A Hybrid System for Time Series Forecasting 109
6. 1 Introduction 109
6. 2 Self-Organization Feature Map (SOM) 113
6. 3 Architecture and Systematic Learning Algorithm 115
6. 4 Experimental Results 116
6. 5 Related Work 135
6. 6 Conclusions 135

Chapter 7 Parallel Sequential Minimal Optimization for the Training of Support Vector Machines 137
7. 1 Introduction 137
7. 2 A Brief Overview of the Modified SMO 139

7. 3 The Parallel SMO 141
7. 4 Experiment 143
7. 5 Conclusions 152

SECTION 2 SUPPORT VECTOR MACHINES FOR CLASSIFICATION AND NOVELTY DETECTION

Chapter 8 Using Support Vector Machine for Bond Rating 155
8. 1 Introduction 155
8. 2 Background on Bond Rating 156
8. 3 Support Vector Machine (SVM) Classifier 157
8. 4 Experimental Results 160
8. 5 Conclusions 168

Chapter 9 Modified Suppport Vector Novelty Detector Using Training Data with Outliers 170
9. 1 Introduction 170
9. 2 Support Vector Novelty Detector (SVND) 172
9. 3 Modified Support Vector Novelty Detector 173
9. 4 Experiment 175
9. 5 Conclusions 181

Chapter 10 Conclusions and Recommendations 183

References 187

Appendix A SVMs Regressor with the SRM Principle 210

Appendix B Data Sets Studied 212

Appendix C Pseudocode of the Modified SMO 215

Appendix D Modified SMO 221