中国农业大学学报 2002, 7(6): 97～100
Journal of China Agricultural University

基于 SVM 分类的预警系统①

刘广利② 邓乃扬
(中国农业大学管理工程学院)

摘 要 将 SVM 理论与预警理论相结合, 提出了一个基于 SVM 的宏观经济预警系统, 并应用于我国棉花产量增长率的预警。与已有预警系统比较, 该预警系统在预警概化能力上有着明显的优势。

关键词 预警系统; 支持向量机(SVM); 概化能力
中图分类号 F 323.11

An Early Warning System Based on SVM Classification

Liu Guangli, Deng Naiyang
(Management Engineering College, China Agricultural University, Beijing 100083, China)

Abstract An early warning system based on Support Vector Machine classification (SVM) was designed with SVM and early warning theory. It has been shown that this early warning system has an obvious advantage of generalization by testing China cotton production.

Key words early warning system; support vector machines; generalization

基于数据的机器学习的基本思想是从观测数据出发寻找统计规律, 并利用这些规律对未来数据进行预测, 这就为宏观经济预警研究开辟了一条新的途径。据此, 王建成[1]提出了概率模式分类的宏观经济预警方法, 与传统的预警方法相比在预警效果方面有了较大的改进。但该方法有明显的缺陷: 需要事先知道先验概率、条件概率、后验概率, 以及概率密度和误判时的损失, 实际上, 如果这些条件已知, 分类问题就是一个简单的运算; 另外, 方法的基础是统计学的渐进理论, 所以小样本条件下就不能保证好的概化能力。

本文中提出了一个基于 SVM 分类的预警系统, 并应用于我国棉花生产预警。与已有预警系统比较, 该预警系统在预警概化能力上有着明显的优势。具体包括: 不用估计概率密度而直接根据样本和问题本身求出最优决策函数; 解决了小样本下的"过学习"; 以结构风险最小化为优化目标。

1 基于 SVM 分类的预警理论

根据统计学习理论, 如果样本数据服从某种分布, 要使机器的实际输出与理想输出之间的偏差尽可能小, 则机器不应遵循经验风险最小化原理, 而应遵循结构风险最小化原理, 即错误概率的上界最小化。SVM 正是这一理论的具体实现。

SVM 是从线性可分情况下的最优超平面发展而来的, 所谓最优超平面就是要求超平面不

收稿日期: 2002 05 09
①国家自然科学基金资助项目
②刘广利, 北京清华东路 17 号 中国农业大学(东校区) 214 信箱, 100083

但能将两类正确分开, 而且使分类间隔最大; 使分类间隔最大实际上就是对概化能力的控制, 这正是 SVM 的核心思想所在。构造最优超平面可转化为下面的最优化问题:

$$\min \Phi(w) = \| w \|^2 \tag{1}$$

约束条件为 $y_i(w \cdot x_i + b) \geq 1,\ i = 1, 2, \cdots, l$。其中: w 为全向量, $\Phi(w)$ 为向量函数。x_i 为第 i 个训练样本, $y_i = \pm 1$, b 为常数。

利用 Lagrange 优化方法可以把上述最优超平面问题转化为其对偶问题:

$$\max W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \tag{2}$$

约束条件为 $\sum_{i=1}^{l} \alpha_i y_i = 0,\ \alpha_i \geq 0,\ i = 1, 2, \cdots, l$。其中 α_i 为 Lagrange 乘子。

式(2)是一个不等式约束下二次函数寻优的问题, 存在唯一解。解中只有一部分(通常是少部分) α_i 不为 0, 对应的样本就是支持向量。求解上述问题, 得到最优分类函数。

$$f(x) = \operatorname{sgn}\{(w \cdot x) + b\} = \operatorname{sgn} \sum_{I} \alpha_i y_i (x_i \cdot x) + b$$

其中 I 表示支持向量。

在线性不可分的情况下, 可以在条件中增加一个非负的松驰项 ξ_i, 式(1)变为

$$y_i(w \cdot x_i + b) + \xi_i \geq 1,\ i = 1, 2, \cdots, l$$

将目标改为

$$\Phi(w, \xi) = \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{l} \xi_i$$

就得到广义最优分类超平面。广义最优分类面的对偶问题与线性可分情况下几乎完全相同, 只是条件变为: $C \geq \alpha_i \geq 0,\ i = 1, 2, \cdots, l$。于是, 构建最优超平面的问题就转化为二次规划问题。

对非线性问题, 可以通过非线性变换转化为某个高维空间中的线性问题, 在变换空间求最优分类面。上面的对偶问题只涉及训练样本之间的内积运算, 在高维空间只需进行内积运算, 而这种内积运算是可以用原空间中的函数运算来实现的。因此, 只要一种核函数 $K(x_i, x_j)$ 满足 Mercer 条件, 它就对应某一变换空间中的内积。因此 SVM 的决策函数就变成

$$f(x) = \operatorname{sgn} \sum_{I} \alpha_i y_i K(x_i \cdot x) + b$$

其中 $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$。选择不同的核函数就可以生成不同的支持向量机。常用的核包括: 多项式核, $K(x, y) = [s(x \cdot y) + c]^d$, 其中 c 和 d 为任意整数, s 为常数; 高斯(径向基函数)核, $K(x, y) = \exp\{-g \| x - y \|^2\}$; 二层神经网络核, $K(x, y) = \tanh[s(x \cdot y) + c]$。

根据上面的分析可知, SVM 学习算法最终归结为求解二次规划的问题。常用的 SVM 学习算法包括 SVM-light, SMO, Chunking 等。本文中实例采用 SVM-light, 算法的基本思想为: 设计有效算法(在一定的约束条件下, 最小化目标函数的一阶近似), 选择满足可行下降方向的"工作集"(Working set), 包含 q 个非 0 分量; 分解原最优化问题。之所以可以对原问题进行分解, 是因为可以利用大多数 SVM 学习问题的 2 个特性, 即: 一是支持向量相对于训练样本而言数量很少, 二是很多支持向量的 Lagrange 乘子分量在上界 C 上。

笔者认为, 宏观经济预警过程可以看作是一个模式识别的过程。预警就是把未知警度的新预警样本与已知警度的预警标准样本进行比较辨别, 从而确定新预警样本所归属于的预警模

式类别的过程。

根据上述理解, 基于 SVM 分类的预警系统构架如下:

知识获取子系统, 由 SVM 分类算法构成, 这是预警系统的核心; 知识库系统, 由学习获取的可以不断更新的权值和参数构成, 知识库是预警知识的载体, 也是报警部件的先导部件; 报警子系统, 相当于模式识别系统中的错误检查系统, 在新的预警数据的驱动下, 经过决策函数的计算, 得到报警结果; 人机交互界面, 是为用户使用该预警系统提供的程序接口, 界面的输入信息是训练集原始数据和相关参数, 输出信息是报警信号。

2　应用实例

文献[2]和[3]对我国棉花产量增长率进行了预警设计, 建立了警兆指标体系, 并对每一个指标与警情变量进行了时差相关分析, 筛选了 12 个指标, 构建了预警计量模型。在此基础上, 对 1989 年和 1990 年的棉花产量增长率进行了外推预警, 结果表明 1989 年增长率介于中警和重警之间; 而 1990 年增长率是轻警。与实际警度相比, 1989 年预警通过检验, 而 1990 年则不能通过检验。

王建成[1]根据文献[2]的 12 个指标, 利用附件信息检验方法筛选了 3 个指标: 棉花播种面积增长率、纺织工业产值增长率和棉粮收购价比, 用贝叶斯分类器对棉花产量增长率进行了预警研究。训练集是 1953- 1988 年的 36 组样本数据, 用所有的样本来设计预警分类器, 同时用所有的样本来检验预警分类器的效果。从预警结果来看, 36 个样本点中有 2 个与原警度不同, 结论是"预警判别效果不错"。但是, 研究表明[4], 过分追求错分率目标会导致"过学习", 从而使得预警方法的概化能力降低。

为了验证 SVM 预警方法的概化能力, 笔者利用文献[1]的贝叶斯分类器对 1989 年和 1990 年的棉花产量增长率进行外推预警, 结果表明两年均为无警; 而实际 1989 年是有警的, 预警外推错误率为 50%。

笔者采用文献[1]的 3 个指标和 36 个训练样本对我国棉花产量增长率进行 SVM 预警。决策函数为 $f(x)=\operatorname{sgn}\left(\sum_i \alpha y_i K(x_i \cdot x)+b\right)$, 核函数为 $K(x,y)=x \cdot y$。运算过程表明默认参数 $C=0.003\ 1$, 迭代步 467, 样本错分率为 13/36, 支持向量个数为 28。预警模型文件保存了参数 b 和 α_i 的计算结果。对 1989 年和 1990 年的产量增长率进行外推预警, 结果表明两年均通过预警检验, 外推错误率为 0。其中, 1989 年的 $\sum_i \alpha y_i K(x_i \cdot x)+b=-0.138\ 42$, 1990 年则为 0.620 80。可以看出, 本文中预警的外推效果要比文献[1]的效果好。

应该指出, 上述基于 SVM 分类的预警样本错分率虽然为 13/36, 但如果调整 $C=0.5$, 训练样本的错分率就变为 0, 即可以实现比文献[1]更低的样本错分率, 也就是说, 一个预警分类器的优劣不能单纯以预警样本的错分率来评价。事实上, 预警分类器的概化能力才是应该追求的最终目标[5,6]。

为了进一步验证 SVM 预警方法的概化能力, 下面的试验把训练集分成 2 部分, 其中前 26 个作训练样本, 用来设计预警分类器; 后 10 个做测试样本, 求其概化错误。参数的选取与上面试验相同。结果表明该预警系统具有较高的概化能力(表 1)。

表 1 我国棉花产量增长率 SVM 预警外推结果

年度	分类函数的计算结果	预测警度	实际警度	年度	分类函数的计算结果	预测警度	实际警度
1979	1.598 50	+ 1	+ 1	1984	1.742 60	+ 1	+ 1
1980	3.333 10	+ 1	+ 1	1985	- 3.354 90	- 1	- 1
1981	- 0.348 21	- 1	+ 1 *	1986	- 1.996 30	- 1	- 1
1982	2.302 40	+ 1	+ 1	1987	0.339 49	+ 1	+ 1
1983	1.383 30	+ 1	+ 1	1988	- 0.290 37	- 1	- 1

注: * 表示预警出现错误。

3 结 论

SVM 分类器具有很好的概化性能。

预警指标选择是机器学习中的一个十分重要的问题。指标选择可以提高预警系统的概化能力, 并能提高运行速度。事实上, 预警模式的维数达到一定的程度, 也会因为引入了太多的特征而导致性能退化。另外, 预警指标可能出现线性相关和冗余, 对预警分类效果产生影响。指标选择的目的就是建立最优预警指标子空间, 以提高机器运行速度, 改善预警效果。

许多文献[7~9]就 SVM 分类的特征选择问题进行了研究, 产生了一些新的思路, 对于进一步优化 SVM 预警系统有着推进作用, 笔者将就基于 SVM 分类的经济预警指标选择问题做进一步的研究。

参 考 文 献

1 王建成. 我国宏观经济预警系统的预警方法研究: [学位论文]. 杭州: 浙江大学, 1999

2 顾海兵, 陈 璋. 中国工农业经济预警. 北京: 中国计划出版社, 1992. 53~207

3 顾海兵, 俞亚丽. 未雨绸缪—宏观经济问题预警研究. 北京: 经济日报出版社, 1993. 30~165

4 Scholkopf S, Burges C J C, Smola A J. Advances in Kernel Methods: Support Vector Learning. Cambridge: M IT Press, 1999. 43~45

5 边肇祺, 张学工. 模式识别(第 2 版). 北京: 清华大学出版社, 2000. 284~303

6 Cristianini N, Shawe-Taylor J. Introduction to Support Vector Machines. Cambridge: Cambridge University Press, 2000. 52~76

7 Weston J, Elisseeff A, Scholkopf B. Use of the lo-norm with linear models and kernel methods. BIOWulf Technical Report, 2001. 12p

8 Chapelle O, Vapnik V, Bousquet O, et al. Choosing kernel parameters for support vector machines. Machine Learning, http: //citeseer. nj nec. com/384397. html, 2000

9 Weston J, Mukherjee S, Chapelle O, et al. Feature selection for svms. Advances in Neural Information Processing Systems, http: //citeseer. nj nec. com/326502. html, 2000