

基于投影寻踪和支持向量机的模式识别方法^{*}

赵晓翠, 王来生

(中国农业大学 理学院, 北京 100083)

摘 要: 由于支持向量机(SVM) 在处理小样本、高维数及泛化性能强等方面的优势, 提出了一种基于投影寻踪(PP) 和支持向量机的模式分类方法。利用 PP 方法把高维数据转换到低维子空间, 同时用加速遗传算法获得最佳投影方向和投影值, 揭示了高维数据的结构特征, 然后在低维空间中用 SVM 对特征向量进行分类识别, 并将其应用到银行信贷风险评估中。选用 2005 年度 80 家贷款申请企业的数据样本, 对该模型进行验证, 通过与神经网络模型的比较, 证实了该方法用于模式识别的有效性 & 优越性。

关键词: 支持向量机; 投影寻踪; 信贷风险评估

中图法分类号: TP391. 4 文献标识码: A 文章编号: 1001-3695(2007) 02-0086-03

Pattern Classification Method Based on Projection Pursuit and Support Vector Machines

ZHAO Xiao-cui, WANG Lai-sheng

(College of Science, China Agriculture University, Beijing 100083, China)

Abstract: Based on the high performance of Support Vector Machine(SVM) in tackling small sample size, high dimension and its good generalization, this paper proposed a pattern classification method based on Projection Pursuit(PP) and SVM. Through applying PP based on RAGA, and the optimum projection direction and the value of project function can be obtained. Then the SVM is used for classification. In the end, this method is used in commercial banks. The financial data specimens of 80 enterprises in 2005 are selected in our country to test this model, empirical results show that the new proposed model is more advantageous than neural network model.

Key words: Support Vector Machine(SVM); Projection Pursuit(PP); Credit Risk Assessment

投影寻踪(Projection Pursuit, PP)^[1] 是国际统计界于 20 世纪 70 年代中期发展起来的一种新的、有价值的高新技术, 是统计学、应用数学和计算机技术的交叉学科。它是用来分析和处理高维观测数据, 尤其是非正态、非线性高维数据的一种新兴统计方法。其基本思想是将高维数据投影到 1 ~3 维子空间上, 寻找能够反映高维数据的结构或特征的投影, 以达到研究分析高维数据的目的。目前, 它已在多元统计分析中得到广泛应用。

支持向量机(Support Vector Machine, SVM) 是一种新型的统计学习方法, 是借助最优化方法解决机器学习问题的新工具, 其理论基础是统计学习理论。它最初于 20 世纪 90 年代由 Vapnik^[2] 提出, 由于其出色的学习性能, 该技术已成为机器学习界的研究热点, 并且在很多领域都得到了成功的应用, 如文本分类、银行信贷风险评估、图像识别等。将投影寻踪方法与支持向量机有机结合, 可望有效地解决模式识别问题。

1 基于 PP 特征提取和 SVM 的模式分类方法

1.1 投影寻踪算法理论^[3]

PP 是模式识别中一种有效的特征提取方法, 其目的是用较

少数量的特征对样本进行描述, 以降低特征空间的维数, 同时又能保留所需要的识别信息, 达到研究和分析高维数据的目的。

在具体操作上, PP 是一个寻优过程, 可用如下的迭代过程表示:

(1) 投影指标的归一化处理。设样本集中各指标的值为 $\{ x^*(i, j) \mid i=1, \dots, n, j=1, \dots, p \}$, 其中 $x^*(i, j)$ 为第 i 个样本的第 j 个指标值。为统一各指标值的变化范围, 可采用极值归一化处理:

$$x(i, j) = [x^*(i, j) - x_{\min}(j)] / [x_{\max}(j) - x_{\min}(j)] \quad (1)$$

式中, $x_{\max}(j)$, $x_{\min}(j)$ 分别为第 j 个指标值的最大值和最小值。

(2) 构造投影指标函数 $Q(a)$ 。PP 方法就是把 p 维数据 $\{ x^*(i, j) \mid j=1, \dots, p \}$ 综合成以 $a = \{ a(1), a(2), a(3), \dots, a(p) \}$ 为投影方向的一维投影值 $Z(i)$, 即

$$Z(i) = \sum_{j=1}^p a(j) x(i, j), \quad i=1, \dots, n \quad (2)$$

根据 $\{ Z(i) \}$ 的一维散布图进行分类。为了在多维指标中找到数据的结构组合特征, 在综合投影指标值时, 要求投影值 $Z(i)$ 的散布特征应满足局部投影点尽可能密集, 最好凝聚成若干个点团, 而在整体上投影点团之间尽可能散开。因此, 投影指标函数可以表达成

$$Q(a) = S_Z D_Z \quad (3)$$

式中, S_Z 为投影值 $Z(i)$ 的标准差; D_Z 为投影值 $Z(i)$ 的局部密度。

$$S_Z = \sqrt{\sum_{i=1}^n (Z(i) - E(Z))^2 / (n - 1)}^{0.5} \quad (4)$$

$$D_Z = \sum_{i=1}^n \sum_{j=1}^n (R - r(i, j)) u(R - r(i, j)) \tag{5}$$

式中, $E(z)$ 为序列 $\{Z(i) \mid i = 1, \cdots, n\}$ 的平均值; R 为局部密度的窗口半径, 一般可以取值为 $0.1S_Z$; $r(i, j) = |Z(i) - Z(j)|$; $u(t)$ 为一个单位阶越函数, 当 $t \geq 0$ 时, 其值为 1, 当 $t < 0$ 时其值为 0。

(3) 优化投影指标函数, 确定最佳投影方向。不同的投影方向反映不同的数据结构特征, 最佳投影方向就是尽最大可能反映高维数据某类特征结构的投影方向。因此可以通过求解投影指标函数最大化问题来估计最佳投影方向, 即

$$\max Q(a) = S_Z D_Z \tag{6}$$

$$\text{s. t. } \sum_{j=1}^p a^2(j) = 1 \tag{7}$$

这是一个以 $a(j)$ 为优化变量的复杂非线性优化问题, 用常规优化方法处理较困难。模拟生物优胜劣汰规则与群体内部染色体信息交换机制的加速遗传算法(RAGA)^[4, 5] 是一种通用的全局优化方法, 用它来求解上述问题十分简便和有效。目标函数达到极值时得到最佳投影方向 a 。

1.2 支持向量机原理^[6]

SVM 是一种基于结构风险最小化原理的机器学习方法。它用于模式分类的基本思想是, 通过求解凸二次规划问题, 寻找一个最优超平面, 使得它能够尽可能多地将两类数据点正确地分开, 同时使分开的两类数据点距离分类面最远。

设线性问题的训练样本集 $T = \{(x_i, y_i) \mid i = 1, 2, \cdots, l\}$, 其中 $x_i \in R^n, y_i \in \{1, -1\}$, 分类超平面为 $(w \times x) + b = 0$ 。

等比例调节 w 和 b 将分类超平面方程归一化, 使得所有满足 $y_i = 1$ 的下标 i , 有 $(w \times x_i) + b \geq 1$; 所有满足 -1 的下标 i , 有 $(w \times x_i) + b \leq -1$ 。此时相应的两条极端直线的距离为 $\frac{2}{\|w\|}$ 。

所以最大化间隔可转换为如下最优化问题:

$$\min_{w, b, \xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l \xi_i \tag{8}$$

$$\text{s. t. } y_i((w \times x_i) + b) + \xi_i \geq 1, i = 1, \cdots, l; \xi_i \geq 0, i = 1, \cdots, l \tag{9}$$

其中使式(2) 等号成立的输入向量称为支持向量, ξ 为松弛变量, C 为惩罚参数。

利用 Lagrange 优化方法可以将上述最优化问题转换为其对偶问题:

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l y_i y_j \alpha_i \alpha_j (x_i \times x_j) - \sum_{j=1}^l \alpha_j \tag{10}$$

$$\text{s. t. } \sum_{i=1}^l y_i \alpha_i = 0 \tag{11}$$

$$0 \leq \alpha_i \leq C, i = 1, \cdots, l \tag{12}$$

通过求解式(3) 可得决策函数为

$$f(x) = \text{sgn}(\sum_{i=1}^l y_i \alpha_i^* (x_i \times x) + b^*)$$

其中 $b^* = y_j - \sum_{i=1}^l y_i \alpha_i^* (x_i \times x_j)$ 。

对于非线性问题, 超平面无法将两类分隔开, 因此通过引进核函数技术将低维的输入空间通过非线性变换映射到高维的特征空间, 在特征空间中求解分类问题。Vapnik 将这种非线性映射函数称为核函数, 并且 $K(x, x') = \Phi(x) \times \Phi(x')$, Φ 为某个映射。引入核函数后, 以上各式的内积都可以用核函数代替, 此时 SVM 的决策函数为

$$f(x) = \text{sgn}(\sum_{i=1}^l y_i \alpha_i^* K(x_i \times x) + b^*)$$

1.3 模式分类实现过程

首先应用 PP 将指标体系(高维数据) 投影到一维子空间上, 然后应用 SVM 方法建立模式分类模型, 求出最优超平面, 最后利用最优超平面分析数据集。

2 基于 PP 和 SVM 的商业银行信贷风险实证研究

近年来, 随着金融的全球化趋势及金融市场的波动性加剧, 商业银行的风险管理一直是国际、国内金融界关注的焦点^[7]。作为我国银行业主体的国有银行不仅其不良贷款额约占我国银行总不良贷款额的 80%, 而且其不良贷款率也很高, 面临的不良贷款问题非常严峻, 已成为我国金融体系稳定运行的主要隐患。如何化解和防范信贷风险, 继而避免由此引发的银行呆账、坏账及信用危机, 已是摆在中国金融业面前的迫切课题。因此, 建立贷款企业的准确审核模型, 对提高银行信贷资产的安全性并减少不良资产的产生有着极其重要的意义。

下面我们将根据科学的指标体系, 采用 PP 进行特征提取, 然后再利用 SVM 对特征向量进行分类识别, 对被评级单位履行经济责任的能力以及可信任程度进行客观公正的评价, 并确定其信用等级^[8]。

2.1 指标与数据

选取我国某国有银行 2005 年度 80 家贷款申请企业的数据样本, 按照他们的财务状况、经营状况以及过往的信用记录分为信用好和信用差两个小组, 其中信用好的小组是由那些如期还本付息的贷款企业构成; 信用差的小组是由那些有可能拒绝还贷的企业构成。训练集是由每个小组中的 30 个样本构成, 因此训练集含有 60(75%) 个样本, 余下的 20(2.5%) 个样本(信用好的有 9 个, 信用差的有 11 个) 构成测试集。

我们将考虑能反映贷款企业的还款能力、赢利能力、经营效率、经营周转率和资金结构等七个财务比率, 它们分别是:

$$\text{负债比率 } x_1 = \frac{\text{负债总额}}{\text{资产总额}}; \text{主营业务净利润率 } x_2 = \frac{\text{净利润}}{\text{主营业务收入}};$$

$$\text{净资产收益率 } x_3 = \frac{\text{销售收入净额}}{\text{净资产}}; \text{流动比率 } x_4 = \frac{\text{流动资产}}{\text{流动负债}};$$

$$\text{速动比率 } x_5 = \frac{\text{速动资产}}{\text{流动负债}}; \text{存货周转率 } x_6 = \frac{\text{产品销货成本}}{\text{平均存货}};$$

$$\text{应收账款周转率 } x_7 = \frac{\text{销售净额}}{\text{应收账款总额}}$$

实际检测中, 可以根据具体要求, 通过一定的方法从中进行选择, 确定最后应用于检测的指标。

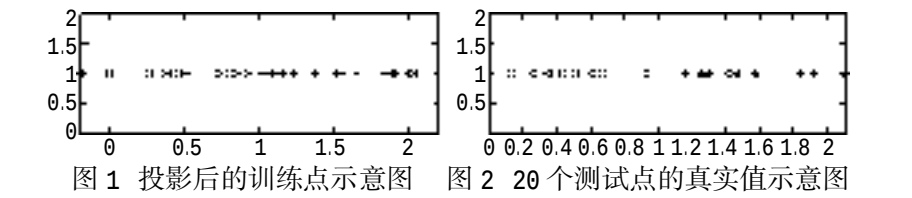
2.2 基于 SVM 和 PP 的检测

运用 MATLAB 7.0 编程^[8] 实现加速遗传算法(RAGA)。当可变惩罚系数 $\alpha = 10$ 时, 得到最佳投影方向 $\alpha = (0.7517, -0.3858, 0.1048, 0.1560, 0.4089, 0.2655, 0.1127)$, 此时式(6) 对应的最大值为 1.9430。

将 α 代入式(2) 可以得到 80 家贷款申请企业的最优投影值 $Z(i)$ 。图 1 表示训练集中的 60 个数据点。图 1 中的 “*” 和 “o” 分别表示训练样本集中信用好的企业和信用差的企业。可以看出, 投影后的两类样本可以很容易地被分开。

我们构造 SVM 模型的内积核函数, 通过线性函数与多项

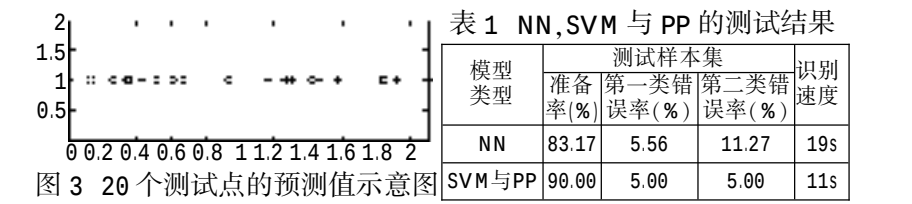
式函数及径向基函数相互比较, 最终确定采用最常用的径向基函数; δ C 等参数采用交叉验证方法(Cross Validation, CV) 确定为 $\delta=1$, $C=10^3$; 使用 SVM Classifier MATLAB Toolbox 工具包进行实验分析, 数据为 20 个测试样本点。预测结果显示如图 2 所示。



如图 2 和图 3 所示, 两幅图中的 “* ”和 “o ”分别表示测试样本集中信用好的企业和信用差的企业。图 2 表示 20 个测试点的真实值, 图 3 表示 20 个测试点的预测值。可以看出, 20 个测试点中绝大部分点被正确划分, 只有极少点划分错误。

2.3 结果分析

表 1 列出了 SVM 与 PP 模型的结果, 同时与神经网络 (NN) 所建模型的结果进行了比较。神经网络使用的是 BP 算法, 由于神经网络方法并不是一种稳定的方法, 故表 1 中神经网络的结果是 15 次的平均结果。表 1 中第一类错误是指将信用好的企业评判为信用差的企业, 第二类错误是指将信用差的企业评判为信用好的企业。可以看出, SVM 与 PP 在测试样本集中的预测准确率达到了 90%, 明显好于 NN 模型的 83. 17%; 我们还发现, SVM 与 PP 的第一类错误率和第二类错误率均小于 NN 的第一类错误率和第二类错误率; 从识别速度, 也就是指将原始数据输入计算机得到最后结果的时间上看, SVM 与 PP 也有很大提高。



3 结论

一个基本的模式识别系统主要是由数据获取、预处理、特征提取和选择以及分类决策所组成。本文主要对其后两部分进行了讨论, 提出了一种基于投影寻踪与支持向量机的模式分类方法。该方法首先利用投影寻踪对传统的指标进行特征提取, 把多维评价指标综合成一维投影指标, 然后通过支持向量机构造广义最优超平面。实验表明, 这一方法大大降低了特征空间的维数, 并有很好的分类正确率。

参考文献:

[1] 李祚泳. 投影寻踪技术及其应用进展[J]. 自然杂志, 1997, 19 (4) : 224-227.

[2] Vapnik. 统计学习理论的本质[M]. 张学工. 北京: 清华大学出版社, 2000.

[3] 黄晓荣, 付强, 梁川. 投影寻踪分类模型在工程评标中的应用[J]. 哈尔滨工业大学学报, 2004, 36(1) : 35-38.

[4] M L Raymer, *et al.* Dimensionality Reduction Using Genetic Algorithms[J]. IEEE Transactions on Evolutionary Computation, 2000, 4 (2) : 164-171.

[5] Friedman J H, Turkey J W. A Projection Algorithm for Exploratory Data Analysis[J]. IEEE Trans. on Computer, 1974, 23(9) : 881-890.

[6] 邓乃扬, 田英杰. 数据挖掘中的新方法——支持向量机[M]. 北京: 科学出版社, 2004.

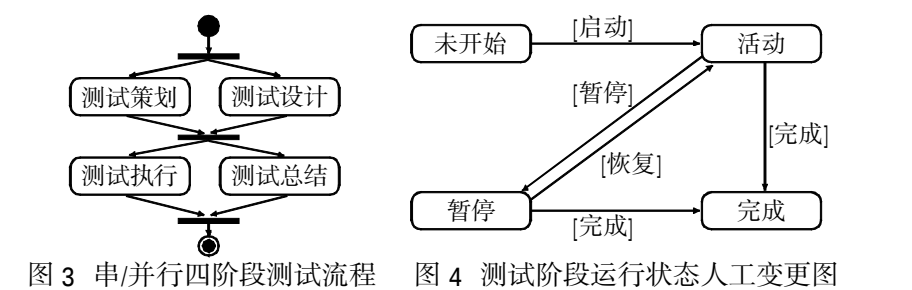
[7] 丁欣. 国外信用风险评估方法的发展现状[J]. 湖南大学学报(社会科学版), 2002, 16(3) : 140-142.

[8] 刘闽, 林成德. 基于支持向量机的商业银行信用风险评估模型[J]. 厦门大学学报, 2005, 44(1) : 72-75.

作者简介:

赵晓翠(1981-), 女, 山东东阿人, 硕士研究生, 主要研究方向为运筹与优化、支持向量机; 王来生(1954-), 男, 吉林人, 教授, 博导, 主要研究方向为运筹与优化、数据挖掘等。

(上接第 45 页) 人员对测试流程进行查看和人工干预控制。测试流程查看功能是指以图形化的方式提供对测试流程的查看, 主要包括以颜色来实时反应测试阶段的四种状态、以对话框来查看测试阶段的计划起止时间和测试流程控制项的完成情况。测试流程人工干预控制功能是指以图形化的方式对测试流程进行人工变更, 主要包括变更测试阶段计划起止时间、变更测试阶段运行状态(图 4) 和变更测试流程控制项启用状态。



5 小结

本文首先对软件测试流程的相关概念进行了阐述, 然后从概念层到实现层详细论述了测试流程定义、执行和监控机制。本文所提出的概念和方法已经在北航软件所开发的测试过程管理工具 QEsuite 2. 0 得到了很好的应用, 实现了对软件测试

项目的测试过程以及相关文档和数据的有效管理, 提高了软件测试工作的效率。

参考文献:

[1] 美] 卡耐基梅隆大学软件工程研究所. 能力成熟度模型(CMM) : 软件过程改进指南[M]. 北京: 电子工业出版社, 2001.

[2] 张春霞, 苏秦. 软件测试过程分析[J]. 计算机应用研究, 2004, 21 (5) : 46-48, 51.

[3] Workflow Management Coalition. Terminology & Glossary [Z]. WPMC-TC-1011. Issue 3. 0, 1999.

[4] Workflow Management Coalition. The Workflow Reference Model [Z]. WPMC-TC00-1003, 1995.

[5] 庞岩梅, 晏海华. 基于工作流技术的软件过程定义和追踪机制[J]. 计算机工程与应用, 2003, 39(26) : 69-72.

[6] 陶伟. 以体系结构为中心的软件产品线开发[D]. 北京: 北京航空航天大学, 1999.

作者简介:

郑小军(1980-), 男, 陕西渭南人, 硕士研究生, 主要研究方向为软件质量保证; 刘超(1958-), 男, 北京人, 教授, 硕士, 主要研究方向为软件工程; 晏海华(1964-), 男, 湖南益阳人, 副教授, 硕士, 主要研究方向为软件工程。