

基于机器学习的风险预测方法研究

彭 岩¹ 王万森¹ 王旭仁¹ 涂序彦²

(首都师范大学信息工程学院 北京 100037)¹ (北京科技大学信息工程学院 北京 100083)²

摘 要 预测是用科学的方法和手段对事物的发展趋势和未来状态进行估量的技术。为了弥补传统方法和技术的不足,各种机器学习技术越来越多地应用于预测的研究中。讨论了在风险预测这一特定领域,应用基于案例的推理(CBR, Case based Reasoning)、支持向量机(SVM, Support Vector Machine)以及人工神经网络(ANN, Artificial Neural Network)等机器学习方法来进行预测的技术。同时,以我们的工作为基础,详细论述了在信贷风险预测和工程评标中基于机器学习预测模型的应用。

关键词 机器学习, 预测, 基于案例的推理, 支持向量机, 人工神经网络

Research of Machine Learning-based Risk Prediction

PENG Yan¹ WANG Wan-sen¹ WANG Xu-ren¹ TU Xu-yan²

(College of Information Engineering, Capital Normal University, Beijing 100037, China)¹

(College of Information Engineering, USTB, Beijing 100083, China)²

Abstract Prediction technology is to predict the develop tendency and future state of a even with scientific method and means. In order to remedy the insufficient of traditional approach, various machine learning technical has been applied in forecast research. This paper discussed the use of machine learning in the risk prediction field, such as Supports Vector Machine, Case based Reasoning as well as Artificial Neural Network. At the same time, based on our own work, the application of machine learning-based prediction model used in credit risk prediction and foundation project bid was discussed in detail.

Keywords Machine learning, Prediction, CBR, SVM, ANN

1 引言

预测技术是根据一定事物的运动和变化规律,用科学的方法和手段对该事物的发展趋势和未来状态进行估量,做出定性或定量的评价。现代预测方法有200多种,但常用的只有20种左右,如直观法、因果链法和比较类推法等定性预测方法,趋势外推法、指数平滑法和回归分析法等定量预测方法,其应用面遍及经济、社会、军事、科技等各个方面。随着社会经济与科技的迅速发展,越来越多的人认识到精确预测对于资源配置、生产安排及风险防范的重要性。风险预测分析中,虽然传统的基于统计的方法应用广泛,但存在着对前提条件要求过于严格或结果不够精确等诸多缺陷。如多元判别分析模型(MDA),它要求数据服从正态分布和等方差,所以很不适合变化多端并且复杂的环境;再比如逐步回归法,由于算法不能够对发展环境下的显著关系集合的复杂性充分地进行建模,因此结果不够精确。近年来,为了弥补传统方法的不足,人们将注意力转移到应用各种机器学习技术来进行风险预测的研究与相应软件系统的研发上。目前,基于机器学习

的方法在风险预测中大显身手,专家和学者们做了许多有益的尝试,并且逐渐进入实用阶段。

本文讨论了在风险预测这一特定领域,应用机器学习来进行风险预测的几种方法:基于案例的推理(CBR, Case based Reasoning)、支持向量机(SVM, Support Vector Machine)以及人工神经网络(ANN, Artificial Neural Network)等方法。选择这几种方法是因为它们自身有着许多优点,可以从多个不同角度弥补传统方法的不足。同时,以我们的工作为基础,详细论述了基于机器学习的智能预测模型的设计与实现。

2 基于机器学习的风险预测方法

机器学习是一门研究怎样用计算机来模拟或实现人类学习活动的学科。它是人工智能中最具有智能特征的前沿研究领域之一。基于机器学习的预测方法试图利用人类思维的特点来处理一些复杂的预测与评估问题,并且已经被成功地运用在许多领域,如风险投资预测、股市预测和招标评标等等。

对于信贷风险性,应考虑的因素按性质大体可归纳为财务、非财务、现金流与信用支持4个方面^[1]。其中财务因素是

到稿日期:2008-05-29 本文研究受到国家自然科学基金项(60773130),国家教育部人文社会科学“十一·五”规划项目(07JC630034)及北京市属市管高校人才强教计划项目资助。

彭 岩(1967—),女,副教授,主要研究方向为人工智能理论及应用,E-mail: pengyan@mail. cnu. edu. cn;王万森(1953—),男,教授,主要研究方向为智能科学技术;王旭仁(1972—),女,副教授,主要研究方向为智能网络技术;涂序彦(1935—),男,教授,主要研究方向为人工智能理论与大系统控制论。

根本因素,其财务状况的评估好坏是偿债能力的关键。在评估中将通过各种财务指标,包括短期偿债能力的指标(如流动比率、速动比率、现金比率)和长期偿债能力指标(如资产负债比率、产权比率等)来反映。现金流量是还款能力的主要标志,要根据借款人的即期现金流量,并与同期偿还的对外债务进行比较,看出是否能足以偿还到期债务。非财务因素用以评价其对现金流量等财务指标的影响方向和程度。信用支持为借款人的信用支持它贷款的偿还提供第二还款来源。

本文讨论将机器学习技术应用于风险分类研究。

2.1 基于案例的推理

基于案例的学习是基于类比的学习的进一步发展,特别适用于较难发现规律性知识、也不容易找到因果模型的领域,这种方法参考了在学习和问题解决过程中先前情况所起的作用,将过去解决问题的经验、方法和过程用于对当前相似问题的求解,其中案例是一定时间和空间中已解决或尚未解决的事件的抽象。设计一个 CBR 系统必须依次解决案例的表示、案例的组织、案例的检索、案例的调整等一系列问题。其中,案例的组织与管理是关键技术,因为案例组织策略与案例检索是对应的,它直接影响案例的检索效率。Aarmodt 和 Plaza 认为 CBR 是一个循环过程^[2],且由 4 个阶段组成:(1)相似案例的检索;(2)重用检索出的案例来得到问题的解决方案;(3)如果有必要修正提出的方法;(4)保留解决方案,并产生新的案例。

通常情况下,检索出来的案例与当前问题仅仅是一定程度的匹配。这时检索出来的案例只是作为一个参考,或者使用一定的修正算法来修正检索出来的案例,使其匹配当前问题。正是由于系统中存在这样的用于修正案例的适应性公式和规则存在,使得系统具有自学习的特点。随着案例的增多,系统处理新问题的能力越来越强。

Hongkyu 和 Ingoo 提出了一个带有多阶段的结构化模块^[3,4]。它一共包括解决问题的 4 个步骤(培训、测试、调试、预测)和 3 种数据输入类型(培训的、测试的、预测的)。这种集成化的模块包括 3 种方法:区别分析、神经网络和基于预测的案例,而且它已经被应用到企业破产预测实验中。这种集成化的方法与独立的方法相比预测的准确性非常高。Jha, Hui 和 Foo^[5]将 CBR 和 ANN 结合在一起研究出了一种在线智能错误诊断方案,可以检测到消费者服务的数据库信息。通过用 CBR 运行收集到的信息就可以获得相配的故障和修复案例。通过研究这些案例我们发现,与传统的 CBR 相比,这种复杂的基于推理的案例和在线智能错误诊断的神经网络方法具有更高的准确性。

Jarillo, Succi 和 W. Pedrycz^[6]利用 CBR, ANNs, RI 等机械学习方法建立了一个软件成果预测系统,并从以下 3 个方面来进行比较:准确性、性能值、配置。他们的结论是 ANN 具有很高的精确度,而 RI 的精确度则相对较低。Kim 和 Han^[7]为公司的债券额提出了一种复杂的数据挖掘模型。这是利用 SOM 和 CBR 相结合的方法,不仅利用传统 CBR 方法为以往问题的具体知识建立案例库及相关方法,而且利用来自案例库以外的知识进行索引。研究表明,SOM/CBR 是一个有效的案例索引方法,而且能够显著地提高预测准确率。Essam 和 Ahmed^[8]研究了一些投标钢铁建筑的建材公司的案例。基于消费者的需求和以往建设项目的经验,这些

公司进行了成本预测并依据预测尽量缩减成本。为了提高利用 CBR 对获得案例预测的精确度,在实验中每个新的项目都被分成许多子问题来解决。系统为每个子问题找到最相似的案例模型,并经 ANNs 为每个子问题提出的解决方法整合在一起,就形成了新项目的完整解决方案。

2.2 支持向量机

支持向量机 Support Vector Machines (SVMs) 是 Vapnik^[9]等人提出的机器学习算法。由于其出色的学习性能尤其是泛化能力,从而引起了人们对这一领域的极大关注。支持向量机采用线性模型通过某个非线性的从向量到一个高维空间的映射来划分非线性的分类边界线。在新空间构建的线性模型能够描绘一个原始空间的非线性的判别边界。在这个新空间里,构建了一个最佳分离超平面(OSH)。因此,SVM 被认为是一个特殊种类线性模型的算法,是最大的边缘超平面。最大的边缘超平面给了决策等级间的最大分离。SVM 方法已经被用于信贷、时间数列预测、保险索赔欺骗预测等金融领域中。

以文献^[10]中的研究为例,研究中所用的数据库来自韩国的信贷担保机构。从 2000—2002 年的破产案例中搜集了 1888 个公司案例,其中包括了 944 个破产的和 944 个非破产的案例,并以随机的顺序排列。其步骤如下:第一步,将多维财政比率的数值减少到两个系数,使用了主分量分析法(PCA)和计算系数分数法(所有公司使用的系数附加成本等于或大于 0.5)。第二步,使用系数分数法在 SVM 图形模型中比较训练的性能。第三步,在破产预测模型中逐步逻辑回归,选择财政比率。最后,比较 SVM, BPN, MDA 和 Logit 方法的预测准确度。

研究结果表明,把输入矢量绘制到一个高维的特性空间中,SVM 能够把复杂的问题(复杂的决策表面)转变成为能用线性判别函数的较简单的问题。在实际的应用中,SVM 能够达到与 BPN 类似或者更好的性能,但 SVM 需要的分类训练数据量少。同时,SVM 能够有效防止过拟合等问题。

2.3 人工神经网络

人工神经网络(ANN, Artificial Neural Network)是对生物神经网络系统的模拟,其信息处理功能是由网络单元的输入输出特性(激活特性)、网络的拓扑结构(神经元的联接方式)所决定的。它由许多互相连接的神经元组成,神经元可以计算所有输入的加权和。如果这个和值大于某个初始值,则产生一个输出。这个输出可以作为其他神经元一个兴奋或者抑制的输入。这个过程一直进行下去,直到神经网络中有一个或者多个输出产生为止。ANN 对问题的求解方式与传统方法不同,它是经过训练来解答问题的。训练一个 ANN 是把同一系列的输入例子和理想的输出作为训练的“样本”,根据一定的训练算法对网络进行足够的训练,使得 ANN 能够学会包含在“解”中的基本原理。训练完成后,可以用来求解相同的问题。

神经网络计算 n 个输入的加权和,如果这个和值大于一个域值 u ,则产生一个输出。否则不产生输出。ANN 的一般训练过程是:首先初始化一系列随机权值,然后用训练集去训练这个网络,通过训练不断调整权重,从而可以学习数据集合中的复杂关系。有许多算法可以训练神经网络,这些算法各有千秋,BP 算法是研究人员最常用的算法。

ANN 是一种平行分布模式处理系统,由于它的模式识别、自学能力、容错能力能很好地处理含有数量缓变的信息。而对涉及到的一些定性因素,其有关信息通常用符号表示,这恰恰是专家系统擅长的,可利用知识和推理来解决。神经网络最初是由 Tam 和 Kiang^[11], Dutta 和 Shekhar^[12] 建议用于银行破产预测。他们利用 3 层 BP 神经网络来训练网络,根据输入到网络的一些样本提供一套权重,在网络训练之后,可以将任何新输入(公司)划分为破产或业绩良好。

3 我们的工作

3.1 基于神经网络组合的风险预测

在信贷风险分析中,一般选取包括盈利能力比率、营运能力比率、短期偿债能力比率、长期偿债能力比率等指标作为输入。输出定义可以按需要分为 2~5 类,既可以简单地分为正常贷款和关注贷款,也可逐级分为好、一般、合格、不佳、恶化。神经网络模型结构则要根据所选取的指标数和输出分类数来决定。图 1 为我们所设计的信贷风险预测模型结构,采用 SOM-PNN 集成神经网络进行商业银行信贷风险预警。其中 SOM 模型选用二维拓扑结构,由两层神经层组成。输入层中的神经元数量与其所包含的变量数一致,其功能是获取数据(在我们的应用中是各种财务比率指标)。PNN 结构动态生成。该模型结合了 SOM, PNN 的优点,使得风险预测准确率得以提高。同时,这个模型可以通过来自各种不同时期的财会数据来研究一个公司的演进、确定公司在同行竞争中所处的位置等等。图 2 为预测结果,其中圆圈标出的为破产公司,其余为有偿付能力的公司。

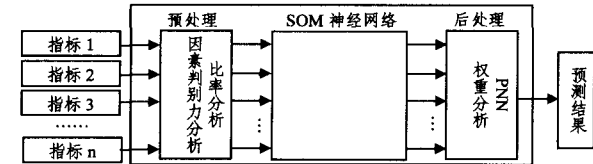


图 1 基于 SOM-PNN 的预测模型

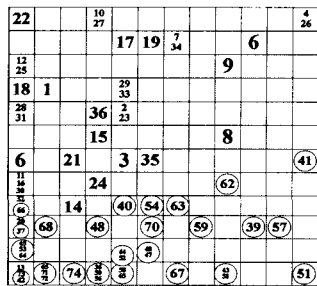


图 2 预测结果

神经网络模型是分布自由的,而且对实际问题适用的,特别是当变量是从未知分布取出和协方差结构不相等(在企业失败样本中的常态)时,神经网络能够提供良好的分类准确性。

3.2 基于神经网络与层次分析法的预测

针对建设工程评标预测工作,我们将层次分析法(AHP)与自组织特征映射神经网络(SOM)相结合,构建了针对分散的和高维的样本数据的决策模型,如图 3 所示。

该决策模型有两大部分:模型的组成和决策流程。其中模型的组成包括决策标准、AHP 和 SOM 3 部分;决策流程由

“评价指标体系”开始沿箭头所指方向,直至到达“SOM”,经 SOM 评价输出决策结果。

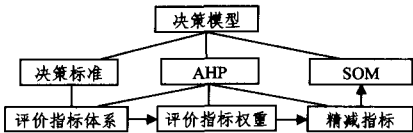


图 3 AHP-SOM 决策模型示意图

我们选取建设工程评标体系中 15 项综合评价指标,每项指标由专家进行考评,并量化处理为 $[0, 1]$ 之间的数值,分值越高表明单个指标的效能越高。以 15 项指标的评价分值作为 SOM 网络的输入,网络的期望输出分为 3 类:优秀企业(1~0.8)、一般企业(0.8~0.6)、较差企业(0.6~0)。预测结果如图 4 所示。

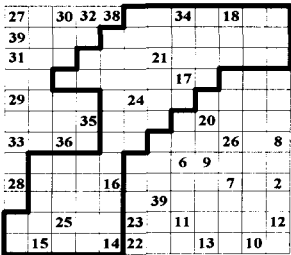


图 4 模型测试结果示意图

通过实验测试表明,模型消除了众多决策因子之间的重叠信息,并将复杂问题的决策简单化,提高了决策的准确性和效率。

结束语 应用预测方法和技术进行科学预测,首先要明确预测目标,根据决策任务提出所预测的目标、时间范围、对预测的技术要求(如结果的准确程度、时限、计量单位、对预测报告的要求)等。同时,收集和研历史数据。最后,选择预测方法并建立预测模型。本文研究了几种基于机器学习的预测方法和技术,并以我们在信贷风险预测和工程评标中的具体工作为对象,研究了基于机器学习的预测模型的应用。在今后的研究中,可进一步探索模型的拓扑结构,以进一步扩展模型的应用领域并提高决策的准确性。

参考文献

- [1] 杨保安,徐晶,查秋野.基于人工智能贷款风险分类的系统设计[J].管理工程学报,2003,17(2):45-48
- [2] Aarmodt A, Plaza E. Case-based reasoning: foundational issues, methodical variations and system approaches. AI Communications, 1994, 7(1): 39-59
- [3] Jo Hongkyu, Han Ingoo, Lee Hoonyoung. Bankruptcy prediction using case-based reasoning, neural networks, and discriminant analysis. Expert Systems with Applications, 1997, 13(2): 97-108
- [4] Jo Hongkyu, Han Ingoo. Integration of case-based forecasting, neural network, and discriminant analysis for bankruptcy prediction. Expert Systems with Applications, 1996, 11(4): 415-422
- [5] Jha G, Hui S C, Foo S. A hybrid case-based reasoning and neural network approach to online fault diagnosis//Proc. 3rd International ICSC Symposia on Intelligent Industrial Automation (IIA'99) and Soft Computing (SOCO'99). Genoa, Italy, 1999: 376-381

了提高,并且在测试精度相同的情况下,RSVM的支持向量会相对较少(如 breast 数据)。这样,对于大数据来说,减少了存储要求,提高了运算速度。RSVM 一般情况比 FSVM 的测试情况要好。

(2)多类样本实验

提出粗糙支持向量机模型主要是考虑了数据样本之间的等价性关系,这同样适应于多类样本的分类。在多类实验中,我们分别用多类分类中的 1-v-1 和 1-v-a 方法进行了测试,核函数同样选取了 3 种核函数来进行实验。

在多类样本实验中,不论是 1-v-1 方法还是 1-v-a 方法,iris,auto,vehicle 3 种数据在 3 种核函数的测试下,我们所提的粗糙支持向量机(RSVM)模型相对于原始 SVM 和 FSVM 的测试效果都有了明显的提高;machine 数据在 polynomial 和 RBF 核函数下,glass 数据在 RBF 测试下,RSVM 的测试结果也较 SVM 和 FSVM 有所提高。

(3)平均测试精度

我们将测试数据样本所得的测试精度进行平均,结果如图 1—图 3 所示。

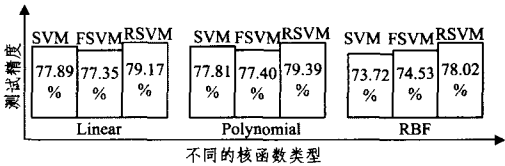


图 1 两类实验对比结果

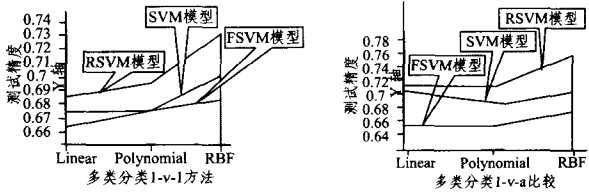


图 2 多类实验对比结果(1-v-1) 图 3 多类实验对比结果(1-v-a)

从上面各图可以看出,将数据结果进行平均以后,不论是两类实验结果还是多类实验结果,粗糙支持向量机(RSVM)的测试精度始终比 SVM 和 FSVM 的测试精度高。

表 2 核函数为 polynomial 时的测试精度(两类)

Data	Polynomial					
	SVM	FSVM	全精度等价性(100%)	变精度等价性(90%)	变精度等价性(80%)	变精度等价性(70%)
breast	0.9474(26)	0.9649(25)	RSVM:0.9649(24)	RSVM:0.9474(25)	RSVM:0.9474(25)	RSVM:0.9474(24)
pima	0.7865(302)	0.7760(292)	RSVM:0.7865(300)	RSVM:0.7865(304)	RSVM:0.7917(304)	RSVM:0.7917(305)
heart	0.7432(89)	0.7432(87)	RSVM:0.7432(89)	RSVM:0.7838(88)	RSVM:0.7432(87)	RSVM:0.7432(90)
bupa	0.6353(198)	0.6118(175)	RSVM:0.6353(198)	RSVM:0.6353(197)	RSVM:0.6353(196)	RSVM:0.6353(196)

结束语 支持向量机能否很好地分类,一个重要方面就是其模型能否有效提取数据的分布与关联信息。粗糙集理论方法的成功应用使人们对数据的等价类信息给予关注,很多研究者试图把粗糙集理论方法与支持向量机结合起来。但所用的结合方法都是分别使用粗糙集和支持向量机处理数据。本文首次把粗糙集信息利用结合到支持向量机模型中,使得该模型有更强的信息提取和利用能力。所做的实验证明了我们所提的 RSVM 模型优于传统的 SVM 和 FSVM 模型,大大提高了测试精度。

参 考 文 献

[1] Cortes C, Vapnik V. Support Vector Networks. Machine learning, 1995, 20(3): 273-297
[2] Vapnik V N. Estimation of Dependencies Based on Empirical Da-

ta, Berlin: Springer-Verlag, 1982

[3] Hsu Chihwei, Lin Chihjen. A comparison of methods for multi-class support vector machines. IEEE Transactions on Neural Networks, 2002, 13(2): 415-425
[4] Huang H P, Liu Y H. Fuzzy support vector machines for pattern recognition and data mining. Int'l Journal of Fuzzy Systems, 2002, 4(3): 826-835
[5] Lee Ki-Young, Kim Dae-Won. Possibilistic support vector machines. Pattern Recognition, 2005, 38(8): 1325-1327
[6] Bo Liefeng, Jiao Licheng, Wang Ling. Working Set Selection Using Functional Gain for LS-SVM. IEEE Transactions on Neural Networks, 2007, 18(5): 1541-154
[7] Pawlak Z. Rough set. International Journal of Information and Computer Science, 1982, 11(5): 341-356

(上接第 207 页)

[6] Jarillo G, Succi G, Pedrycz W, et al. Analysis of Software Engineering Data Using Computational Intelligence Techniques // Proceedings of the 7th International Conference on Object-Oriented Information Systems. Calgary, AB, 2001
[7] Kim Kyung-Sup, Han Ingoo. The cluster-indexing method for case-based reasoning using self-organizing maps and learning vector quantization for bond rating cases. Expert Systems with Applications, 2001, 21(3): 147-156
[8] Changchien S W, Linb Ming-Chin. Design and implementation of a case-based reasoning system for marketing plans, 2005, 28(1): 43-53

[9] Vapnik V N. The Nature of Statistical Learning Theory. Springer, 1995
[10] Min J H, Lee Young-Chan. Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. Expert Systems with Applications, 2005, 28: 603-614
[11] Tam K Y, Kiang M Y. Managerial Applications of Neural Networks; The Case of Bank Failure Predictions. Management Science, 1992, 38: 926-947
[12] Dutta S, Shekhar S. Bond Rating: A Non-Conservative Application of Neural Networks // Proceedings of the IEEE International Conference on Neural Networks. San Diego, New York: IEEE Press, 1989: 443-450