

因子分析和支持向量机的信息系统风险评价

伍浏阳

(唐山师范学院 科研处,河北 唐山 063000)

摘 要: 风险评价是信息系统管理中的关键技术,针对信息系统风险的影响因子多、小样本、时变性等特点,以提高风险评价的精度为目标,提出一种因子分析法和支持向量机的信息系统风险评价模型(FA-CS-SVM). 首先采用因子分析法对风险指标体系进行处理,提取公共指标,消除指标之间相关性,然后将公共指标作为支持向量机的输入向量进行训练,并采用布谷鸟搜索算法寻找最合理机参数,构建信息系统风险评价模型,最后利用仿真实例验证其可行性和合理性. 结果表明,FA-CS-SVM 提高了信息系统风险评价的精度,评价结果可以为信息系统管理员进行风险管理提供科学的参考依据.

关键词: 因子分析;神经网络;信息系统;风险分析

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2016)02-0144-05

Risk Analysis of the Information System by Using Factor Analysis and Support Vector Machine

WU Liu-yang

(Department of Scientific Research, Tangshan Normal University, Tangshan 063000, China)

Abstract: Risk assessment is the key technology in information system management. In view of the risk factors characteristics of small sample, time variability, etc., a risk assessment method (FA-CS-SVM) for the information system by using factor analysis and improved support vector machine is proposed to improve the accuracy of risk assessment. Firstly, factor analysis method is used to deal with the risk index system to extract the public index system and eliminate the correlation among the indexes, and secondly public indexes are used as input vector of support vector machine for training which cuckoo search algorithm is used to find the most reasonable parameters for support vector machine, and the risk evaluation model of information system is constructed, and the feasibility and rationality of the model is verified in the end. The results show that FA-CS-SVM can improve the accuracy of the information system risk evaluation, and the evaluation results can provide scientific reference for the management of information system.

Key words: factor analysis; neural network; information system; risk analysis

1 引言

随着信息技术的不断发展,社会信息化程度日益提高,信息系统广泛应用于商业、政府以及企业中. 由于信息系统具有开放性,信息技术给人们工作、生活带来便利的同时,也给社会各组织带来前所未有的威胁^[1]. 风险的分析和评价有利于掌握信息系统当前以及将来存在的风险,有助于做好应对措施,

防患于未然,因此信息系统的风险评价成为信息安全领域研究中的重点^[2].

针对信息系统的风险评价技术,各行各业的专家和学者投入了系统、深入的研究,涌现出了许多有效的信息系统风险评价模型^[3]. 当前信息系统风险评价模型有定性分析、定量分析以及两者的组合,其中定性分析主要有 OCTAVE 法^[4],根据专家知识和经验对信息系统风险等级进行评价,由于信息系统

收稿日期: 2015-04-24; 修回日期: 2015-06-15

基金项目: 河北省科技计划项目(3220319D);唐山师范学院科学研究基金项目(2014D03)

风险是多种因素综合作用的结果,因此变化十分复杂,而定性分析结果具有盲目性,评价准确性低、结果不可靠^[5].定量分析主要有模糊理论、扩展攻击树、灰色理论等^[6-8],通过建立信息系统风险评价问题的相应数学模型,然后进行评价,而信息系统风险具有时变性、非线性,难以建立精准的数学模型,因此应用受到一定的限制^[9].组合法综合了定性分析法和定量分析法的优点,出现了层次分析法、神经网络、模糊理论和支持向量机等融合^[10-12],较好地描述风险值与评价指标之间的变化关系,且不需要了解系统的先验知识,具有自学习能力,获得了较理想的评价结果.在实际应用中,神经网络要求的信息系统风险评价样本规模大,如果不能满足该条件,就会得到“过学习”的风险评价结果^[13];支持向量机(SVM)的学习性能好,不会出现神经网络的“过学习”现象,其评价结果与参数直接相关,而当前参数选择没有统一理论指导,信息系统风险评价精度有待进一步提高^[14].

为了改善信息安全风险的评价精度,针对信息系统风险的影响因子多、小样本、时变性等特点,本文提出一种因子分析(factor analyze, FA)法和布谷鸟搜索(cuckoo search, CS)算法优化支持向量机的信息系统风险评价模型(FA-CS-SVM),并采用具体应用实例对可行性和评价结果进行对比分析.

2 信息系统安全风险指标

信息系统风险是一个相对概念,怎么样的系统是安全的,通常通过信息系统安全风险指标进行评价,因此评价指标是否科学、合理将直接影响评价结果的优劣.根据国内外的相关研究以及评价指标选择的原则和要求,建立如图1所示的信息系统风险评价指标体系.

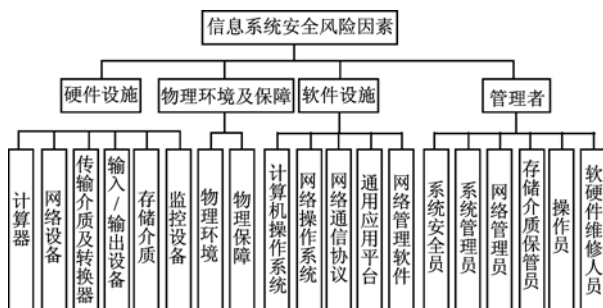


图1 信息系统风险的指标

3 信息系统风险评价模型

3.1 因子分析法

在信息系统风险评价过程中,每一种指标对风险评价影响大小不同,不同指标之间可能存在相关性,因此直接采用上述指标评价信息系统的风险,有时难以反映信息系统风险的真实情况,对评价精度产生不利影响.FA法可以将关系比较复杂的指标归结为少数几个无关的综合指标,降低指标数量,减少SVM的输入向量,提高建模效率.设总共有 p 个原始信息系统风险评价指标,那么对 $m < p$ 个公共指标(X_i)和一个特有指标(ε_i)进行线性加权,就可以描述信息系统风险的原始评价指标,即:

$$T_i = \omega_{i1}X_1 + \omega_{i2}X_2 + \cdots \omega_{im}X_m + \varepsilon_i \quad (1)$$

$$i = 1, 2, \dots, n$$

式中, n 表示样本单位; ω_{ij} 表示第 i 个指标在第 j 个公共指标上的系数.

FA法的工作步骤如下所示.

(1) 对输入数据(t_{ij})进行标准化,消除变量之间的量纲差异,计算相关系数矩阵判断.

$$\begin{cases} R = (r_{ij})_{p \times p} \\ r_{ij} = \frac{\sum_{k=1}^n (t_{ki} - \bar{t}_i)(t_{kj} - \bar{t}_j)}{\sqrt{\sum_{k=1}^n (t_{ki} - \bar{t}_i)^2 (t_{kj} - \bar{t}_j)^2}} \\ i, j = 1, 2, \dots, p \end{cases} \quad (2)$$

式中, $\bar{t}_j$ 为第 i 行数据的均值; r_{ij} 为相关系数.

(2) 根据KMO量值分析变量之间的相关性,构造因子变量,根据 $| \lambda I - R | = 0$ 得到方差贡献率和累积方差贡献率,具体如下:

$$\frac{\lambda_i}{\sum_{k=1}^p \lambda_k}, i = 1, 2, \dots, p \quad (3)$$

$$\frac{\sum_{k=1}^i \lambda_k}{\sum_{k=1}^p \lambda_k}, i = 1, 2, \dots, p \quad (4)$$

式中, λ_i 表示相关系的数矩阵特征值.

(3) 根据特征的累计贡献率确定主要评价指标,一般选择累计贡献率处理85% ~ 90%范围的前 m 个指标,其载荷矩阵具体为

$$\begin{cases} A = (a_{ij})_{m \times p} \\ a_{ij} = \sqrt{\lambda_i} l_{ij} \\ i, j = 1, 2, \dots, p \end{cases} \quad (5)$$

式中, l_{ij} 表示第 i 个指标与原变量 T_j 间的系数.

(4) 将指标变量变为原始变量的线性组合, 得到:

$$\begin{cases} X_i = l_{i1}T_1 + l_{i2}T_2 + \cdots + l_{ip}T_p \\ i = 1, 2, \cdots, m \end{cases} \quad (6)$$

3.2 SVM

SVM 是建立在结构风险最小化理论的算法, 较好地解决传统算法的训练误差小、测试误差大的难题, 平衡了学习精度和学习能力之间矛盾, 在小样本条件下仍然可得到较好泛化性能. 设训练样本集为 (x_i, y_i) , 那么回归模型为

$$y = w_1x_1 + w_2x_2 + \cdots + w_mx_m + b \quad (7)$$

根据 SVM 的工作原理, 可以得到:

$$f(x) = \langle w \cdot x \rangle + b \quad (8)$$

式中, w 为权向量, b 为阈值.

这样可以得到如下最优化问题:

$$\begin{aligned} \min \varphi(w) &= \frac{1}{2} \|w\|^2 \\ \text{s. t.} & \begin{cases} y_i - \langle w \cdot x \rangle - b \leq \varepsilon \\ \langle w \cdot x \rangle + b - y_i \leq \varepsilon \\ i = 1, 2, \cdots, l \end{cases} \end{aligned} \quad (9)$$

引入松弛变量, 并将训练误差作为约束条件, 可以得到:

$$\begin{aligned} \min \varphi(w) &= \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n (\xi_i^* + \xi_i) \\ \text{s. t.} & \begin{cases} y_i - \langle w \cdot x \rangle - b \leq \varepsilon + \xi_i \\ \langle w \cdot x \rangle + b - y_i \leq \varepsilon + \xi_i^* \\ i = 1, 2, \cdots, l \end{cases} \end{aligned} \quad (10)$$

式中, C 为惩罚系数.

为了加快学习效率, 引入 Lagrange 乘子法对式 (11) 进行求解.

$$\begin{aligned} \min_{\alpha, \alpha^*, \beta, \beta^*} \min_{w, b} Q &= \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n (\xi_i^* + \xi_i) - \\ &\sum_{i=1}^l \alpha_i (\varepsilon + \xi_i - y_i + \langle w \cdot x \rangle + b - \\ &\sum_{i=1}^l \alpha_i^* (\varepsilon + \xi_i^* + y_i - \langle w \cdot x \rangle - b) - \\ &\sum_{i=1}^l (\beta_i \xi_i + \beta_i^* \xi_i^*) \end{aligned} \quad (11)$$

对于一个非线性回归问题, 通过选择适当的内积函数 $K \langle x_i, x_j \rangle$ 可以变为线性回归问题, 此时目

标函数为

$$\begin{aligned} \min_{\alpha, \alpha^*, \beta, \beta^*} \min_{w, b} Q &= \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n (\xi_i^* + \xi_i) - \\ &\sum_{i=1}^l \alpha_i (\varepsilon + \xi_i - y_i + \langle w \cdot x \rangle + b - \\ &\sum_{i=1}^l \alpha_i^* (\varepsilon + \xi_i^* + y_i - \langle w \cdot x \rangle - b) - \\ &\sum_{i=1}^l (\beta_i \xi_i + \beta_i^* \xi_i^*) \end{aligned} \quad (12)$$

最后得到 SVM 的回归模型为

$$f(x) = \sum_{x_i \in S_0} (\alpha_i - \alpha_i^*) k \langle x_i \cdot x \rangle + b \quad (13)$$

通常选择径向基核函数构建 SVM 模型, 其定义为

$$K(x_i, x) = e^{-\frac{1}{2\sigma^2} \|x - x_i\|^2} \quad (14)$$

式中, σ 表示核函数宽度.

参数 C 和 σ 直接影响 SVM 复杂程度和学习效果, 为了提高 SVM 的学习能力, 本文选择寻优能力强的 CS 算法对参数 C 和 σ 进行寻优, 建立精度更高的信息风险评价模型.

3.3 构建信息系统的风险评价模型

风险评价一直是信息系统管理研究中关注的重点, 针对支持向量机应用于信息系统风险评价过程中的参数优化问题, 以及系统指标数量多、相关性严重等特点, 提出了 FA-CS-SVM 的信息系统风险评价模型, 其工作思路: 首先建立信息系统风险评价的指标体系, 然后采用因子分析法对风险指标体系进行处理, 选择公共指标, 消除指标间相关性, 最后将公共指标作为 SVM 的输入向量, 采用训练集进行学习, 通过布谷鸟搜索算法寻找最合理参数, 具体流程如图 2 所示.

4 信息系统风险评价算例分析

4.1 数据源

本文以某企业的信息系统作为研究对象, 以该系统的 19 个评价指标作为输入, 以实际的信息系统风险值作为期望输出, 风险值的变化曲线如图 3 所示, 选择前 100 样本作为训练集建立基于 FA-CS-SVM 的信息系统风险评价模型, 最后 50 个样本作为测试集, 测试 FA-CS-SVM 的风险评价效果.

4.2 评价指标的因子分析

首先采用 DSP 8.0 软件对构建的评价指标体系进行因子分析, 得到 KMO 度量值为 0.755, 这表明

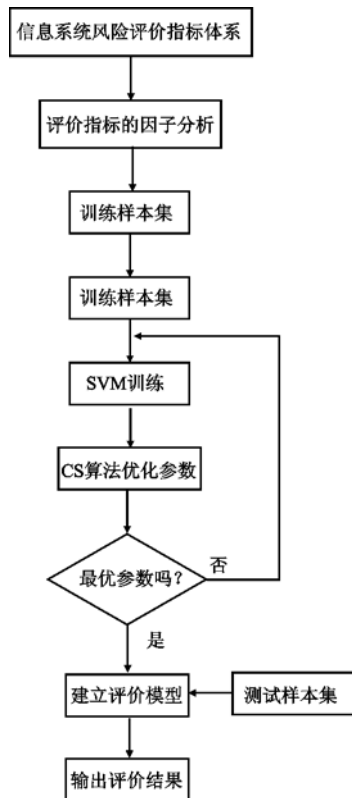


图2 FA-CS-SVM 的工作流程

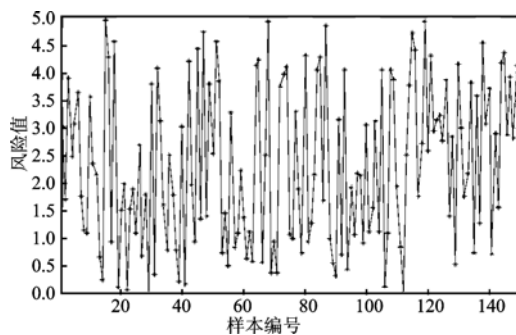


图3 某信息系统的风险值变化曲线

需要对信息系统的原始评价指标进行因子分析,得到评价指标的累计贡献率如表1所示.从表1可以清楚看出,前8个因子的累计贡献率就达到了95.638%,这表明该8个因子可以解释原始评价指标的能力,有效降低指标间的相关性,去掉了一些对评价结果贡献比较小的指标,有利于提高后续的信息系统风险评价速度.

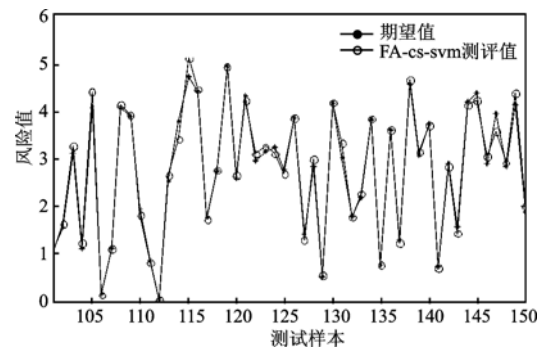
4.3 FA-CS-SVM 的评价结果分析

根据因子分析结果对信息系统风险评价训练集和测试集进行降维处理,然后采用SVM对训练集中的样本进行学习,CS算法优化SVM参数 C 和 σ ,得到最优 $C=10$, $\sigma=1.015$,测试集样本的评价结果如图4所示.从图4可知,FA-CS-SVM的信息系统风

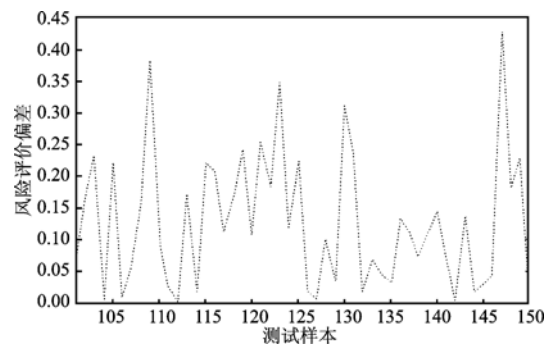
险评价效果很好,评价误差能够控制在10%有效范围以内,评价结果可靠.

表1 因子的累计贡献率

因子	特征值	贡献率%	累计贡献率%
1	10.156	36.274	36.274
2	6.556	23.414	59.687
3	3.605	12.873	72.560
4	2.028	7.244	79.804
5	1.760	5.518	85.322
6	1.279	4.567	89.889
7	1.020	3.356	93.245
8	0.750	2.394	95.638
...	...	...	...



(a) 风险期望值与评价值的变化曲线



(b) 期望值与评价值的偏差变化曲线

图4 FA-CS-SVM 的风险评价仿真结果

4.4 与其他模型的性能对比分析

为了测试FA-CS-SVM的优越性,选择FA法与交叉验证法优化SVM(FA-SVM)、没有FA法的CS算法优化SVM(CS-SVM)、多元线性回归(MLR)以及文献[15]的信息系统网络评价模型进行对比实验,它们的结果如表2所示.从表2可知,在所有信

息系统风险评价模型中,FA-SCS-SVM 评价结果的平均误差最小,评价精度最高,评价结果可以为信息系统管理员提供有科学的参考意见.

表 2 FA-CS-SVM 与对比模型的平均误差

评价模型	平均评价误差/%
MLR	11.051
文献[15]	7.903
FA-SVM	7.824
CS-SVM	7.158
FA-CS-SVM	4.255

5 结束语

信息系统风险受到许多指标的影响,复杂多变,为了提高信息系统风险评价的可靠性,使评价结果更加科学、合理,提出一种 FA-CS-SVM 的信息系统风险评价模型,从实例中结出以下结论.

(1)采用因子分析法对信息系统风险的原始评价指标进行处理,可以获得较好的公共评价指标,大幅度降低了 SVM 的输入向量维数,较好避免了“维数灾”难题,加快了信息系统风险评价建模速度.

(2)采用 SVM 对评价指标与信息系统风险值之间的变化关系进行拟合,有效描述了信息系统风险的变化趋势,并通过 CS 算法解决了 SVM 参数优化的难题,提高了信息系统风险的评价精度.

(3)对比实验验证了 FA-CS-SVM 的评价结果要优于其他评价模型,其可以为信息系统风险准确评估提供了一种科学、有效地方法.

参考文献:

- [1] 登国,张阳,张玉清. 信息安全风险评估综述[J]. 通信学报, 2004, 25(7):10-18.
- [2] 李鹤田,刘云,何德全. 信息系统安全风险评估研究综述[J]. 中国安全科学学报, 2006, 16(1):108-114.
- [3] 范红,冯登国,吴亚非. 信息安全风险评估方法与应用[M]. 北京:清华大学出版社,2006.

- [4] 王甲生,付钰,吴晓平. 基于改进 FAHP 法的信息系统安全风险评估[J]. 火力与指挥控制, 2011, 35(1): 33-36.
- [5] 付沙,杨波,李博. 基于灰色模糊理论的信息系统安全风险评估研究[J]. 现代情报, 2013, 33(7): 34-37.
- [6] 甘早斌,吴平,路松峰,等. 基于扩展攻击树的信息系统安全风险评估[J]. 计算机应用研究, 2007, 24(11): 153-156.
- [7] 王帆,霍明奎,王晓婷. 基于模糊灰度的信息系统安全风险评价与对策[J]. 情报科学, 2014, 32(1): 110-114.
- [8] 付钰,吴晓平,叶清,等. 基于模糊集与熵权理论的信息系统安全风险评估研究[J]. 电子学报, 2010, 38(7): 1489-1494.
- [9] 马丽仪,张露凡,杨宜,等. 基于模糊神经网络方法的信息系统安全风险评价研究[J]. 中国安全科学学报, 2012, 22(5): 164-167.
- [10] 党德鹏,孟真. 基于支持向量机的信息安全风险评估[J]. 华中科技大学学报:自然科学版,2010,38(3): 46-49.
- [11] 阮慧,党德鹏. 基于 RBF 模糊神经网络的信息安全风险评估[J]. 计算机工程与设计, 2011, 32(6): 2113-2115.
- [12] 王甲生,付钰,吴晓平. 基于模糊神经网络的信息系统安全风险评估研究[J]. 武汉理工大学学报:交通科学与工程版, 2011, 35(1): 51-54.
- [13] 汤昶烽,卫志农,李志杰,等. 基于因子分析和支持向量机的电网故障风险评估[J]. 电网技术, 2013, 37(4): 1039-1045.
- [14] 付钰,吴晓平,王甲生. 基于模糊-组合神经网络的信息系统安全风险评估[J]. 海军工程大学学报, 2010, 22(1): 18-23.
- [15] 付沙. 一种基于信息熵的信息系统安全风险分析方法[J]. 情报科学, 2013, 31(6): 38-42.

作者简介:

伍浏阳 男,(1981-),硕士,讲师.研究方向为信息管理系统. E-mail:35826402@qq.com