支持向量机决策树在隐患预警模型中的应用①

闫晓静[1,2], 于　放[2], 孙　咏[2], 肖卡飞[1,2], 王　嵩[2]

[1](中国科学院大学, 北京 100049)
[2](中国科学院 沈阳计算技术研究所, 沈阳 110168)

摘　要: 危化企业的安全监控数据具有社会价值, 对安全隐患进行实时精确的预测是预警研究的热点, 本文从人、设备、环境和管理四个维度出发, 对安全生产隐患预警的相关指标进行分析, 构建隐患预警指标体系, 在此基础上, 构建了自底向上的基于支持向量机的决策树多分类预警模型, 实现对安全等级的的准确分类并用于预警未来的安全生产状态, 通过与自顶向下的多分类模型比较, 证实本文所采用的预警模型具有较好的实时性和精确度, 满足对预警模型的基本要求.

关键词: 预警模型; 支持向量机; 决策树

Risk Early-Warning Model Based on SVM Decision Tree

YAN Xiao-Jing[1,2], YU Fang[2], SUN Yong[2], XIAO Ka-Fei[1,2], WANG Song[2]

[1](University of Chinese Academy of Sciences, Beijing 100049, China)
[2](Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China)

Abstract: The security monitoring data of Dangerous chemicals business has great social value, especially real-time accurate prediction of the security risk has become a hot warning research. From the view of four dimensions which are people, equipment, the environment and management, this article analyzes the relevant indicators of safety hazards warning, constructs the bottom-up decision tree based on multi-classification SVM warning model, constructs a bottom-up decision tree SVM multi-classification model based on early warning, to achieve the security level of accurate classification and for future production safety status warning. By comparison with more top-down classification model, it confirms that early warning model used in this paper has better performance in real-time and accuracy, and meets the basic requirements of early warning models.

Key words: early-warning model; SVM; decision tree

1　引言

我国危化企业规模庞大, 安全指数参差不齐, 并且涵盖的危化品种类繁多. 因此, 造成从产品的生产、存储、运输到经销、使用等环节中隐患重重. 其中, 受生产力发展水平、从业人员素质、安全生产条件等因素的制约, 危化品生产安全事故发生情况尤其突出. 目前, 在危化品生产中, 越来越多地增加了安监预警机制, 通过利用采集到的生产状态监督数据, 归纳隐患状态特征, 建立满足需求的预警模型, 对未来生产数据进行隐患预警[1]. 但安全生产预警技术仍处于起步阶段, 预警指标单一, 例如: 只考虑危险源、利用仪器实时监测等, 未形成客观全面的预警体系[2], 而且长期监测的数据资源比如生产状态数据、隐患排查整改以及持续不断完善的安全评估标准等未能充分利用. 而这些数据基础为利用数据挖掘技术进行预警技术研究提供了条件.

随着人工智能技术的研究不断深入, 智能模型如神经网络[3]、支持向量机[4]、分类树等被广泛应用在各行各业的风险预警中, 周健等将多元判断分析预警模型应用在煤矿灾害方面并得到了较好的验证[5]; 张星联等提出了基于神经网络的蔬菜农药残留风险预警模型[6], 人工神经网络模拟人类大脑处理信息, 充分利用了神经网络的优点, 增强了网络的自适应性; 但是收敛速度较慢而且目标函数存在局部极小点; 金珠研

① 收稿时间:2016-05-16;收到修改稿时间:2016-06-16　[doi: 10.15888/j.cnki.csa.005589]

究了改进的支持向量机分类算法在人因事故安全评价中的应用[7], 支持向量机克服了人工神经网络的缺点, 具有全局最优解, 预测能力较强,但是难以提取预警规则. 本文采用支持向量机决策树的预警模型对危化企业隐患状态进行分类预测, 达到实时预警的效果.

2 基础知识

2.1 支持向量机

支持向量机通过寻找最优分类超平面将各类精确分开, 通过最大边缘超平面实现最优分类[8], 主要分两种情况:

1) 对于线性可分问题, 设给定数据集 D 为 $(X_1,y_1),(X_2,y_2),...,(X_{|D|},y_{|D|})$, X_i 是训练元祖, 具有类标号 y_i, $y_i\in\{+1,-1\}$, 学习就是构造函数, 尽可能正确划分样本且最大化分类间隔, 分类超平面记为:

$$W\cdot X+b=0 \tag{1}$$

其中, W 是权重向量, $W=\{w_1,w_2,...,w_l\}$; l 是属性个数; b 是标量. 在约束条件:

$$y_i\left[(W\cdot X)+b\right]\geq 1, i=1,2...l \tag{2}$$

计算最大边缘 $\frac{2}{\|W\|}$, 所以最优超平面就是在约束下求得:

$$\min\varphi(W)=\frac{1}{2}\|W\|^2 \tag{3}$$

利用拉格朗日公式转换为对偶形式:

$$\begin{aligned}&\min Q(\alpha)=\sum_{j=1}^{l}\alpha_j-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)\\&s.t.\sum_{j=1}^{l}\alpha_j y_j=0\quad j=1,2,...,l,\alpha_j\geq 0, j=1,2,...,l\end{aligned} \tag{4}$$

最终求得最优分类超平面 $W^*\cdot X+b^*=0$, 最优权值向量为 W^*, 最优偏倚为 b^*, 最优分类函数为:

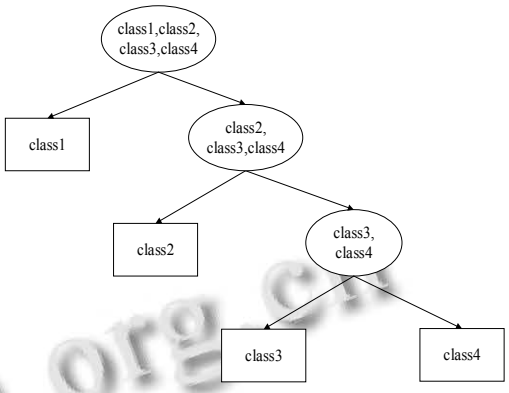
$$f(x)=\operatorname{sgn}\left\{\left(\sum_{j=1}^{l}\alpha_j^*y_j(x_j\cdot x_i)\right)+b^*\right\}, x\in R^n \tag{5}$$

2) 对于非线性问题, 主要是将输入向量映射到高维特征空间, 然后在该特征空间求得最优分类超平面. 假设 Rd 为输入空间, 则存在映射 $\phi: Rd\to H$, H 为高维空间, Vapnik 指出, 核函数 $K(x_i,x_j)$ 满足 Mercer 条件, 则存在某一变换空间中的内积[7], 于是引入核函数解决高维空间计算问题. 最终非线性问题的分类函数为:

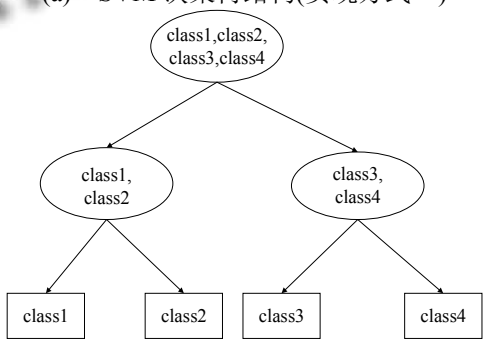
$$f(x)=\operatorname{sgn}\left(\sum_{i=1}^{l}y_i\alpha_i^*\phi K(x_i,x)+b^*\right) \tag{6}$$

2.2 SVM 决策树分类器

对于一个多分类问题, 任何两类都是两两可分的, 于是 SVM 二分类器与二叉树相结合构造的分类器能够巧妙的用二分类的方法解决多分类问题, 在SVM决策树分类器中, 从根节点开始, 对每一层节点都进行 SVM 二分类, 依次进行下去, 直到叶子节点能把所有的类别都单独分离完成为止. 经典的SVM决策树分类器有两种实现形式[9]: 1)在此类 SVM 决策树分类器中, 根节点表示所有类元素的集合, 通过一次SVM二分类器将其分为一个单独类别所有元素的叶子节点和包含其他所有类别的子树根节点, 再对子树根节点进行依次深入SVM二分类, 最终得到每个叶子节点均表示一个单独类别, 如图 1(a); 2)在此类 SVM 决策树分类器中, 从根节点起, 通过一次 SVM 二分类, 将所有类别元素分成两个子类集合, 然后分别对它们进行SVM二分类, 直至每一个叶子节点表示一个单独的类别, 如图 1(b).



(a) SVM 决策树结构(实现方式一)



(b) SVM 决策树结构(实现方式二)

图 1 经典的 SVM 决策树分类器的两种实现形式

从 SVM 决策树分类器的构造过程来看, 构造的决策树是一棵完全二叉树, 每个叶子结点对应一个类, 非叶结点对应一个SVM分类器, 对于N类问题, 需要训练

N-1 个 SVM, 传统的“一对一”法构造的多分类器需要 N 个 SVM, “一对其他”法构造的多分类器需要 N*(N-1)/2 个 SVM, 基于树结构的多分类器具有较优的性能[10].

3 隐患预警模型构建

3.1 预警指标及样本

影响危化企业安全的因素复杂多样, 按照国家标准, 安全生产管理要从四方面着手, 因此安全隐患指标主要从人的因素, 设备的因素, 环境的因素, 和管理的因素四个方面来考虑, 预警指标由 3 个一级指标—人的行为隐患, 设备隐患, 环境隐患和管理隐患, 12 个二级指标—人员技术素质, 群体行为, 人员心理状态, 人员生理状况, 设备设计与选型, 设备购置, 设备使用, 设备维护, 外围环境, 自身环境, 条例管理, 行动管理以及 52 个三级指标构成. 通过假设检验的统计方法, 比较前两周以及安全等级发生变化当天相关指标的平均水平是否有显著性差异, 挖掘出影响安全等级的隐患特征, 有可能引发安全生产爆发危险的隐患特征包括:

设备隐患: 1)设备的使用期限越长, 发生危险的可能性越大; 2)安全保护装置的完好程度影响隐患的爆发, 3)设备在使用过程中越接近最大负荷越容易引发危险.

生产环境隐患: 1)生产过程中, 温度、压力、液位、流量、气体浓度等参数的控制严重影响易燃易爆事故的发生.

管理隐患: 1)设备、生产环境等隐患排查工作不能及时完成, 安全等级明显降级; 2)排查出的隐患不能及时整改, 安全等级降级严重时引发事故.

人的行为隐患: 工作强度和工作人员技术水平是影响安全生产的显著指标.

最终筛选出 12 个隐患预警指标作为样本属性, 分别为设备使用期限、保护装置完好程度、设备负荷、温度、压力、液位、流量、浓度、隐患排查情况、隐患整改情况、技工等级、工作强度, 按照事故类型, 隐患整改情况将安全等级设为安全, 较安全, 基本安全, 危险 4 个等级, 预警模型就是实现对检测的数据进行分类.

3.2 度量相异性

样本分类的难易程度可以通过相异性来度量, 相异性越小, 相异性值越接近 0, 相异性越大, 相异性值越接近于 1, 对于 n 类样本集{X1, X2, …,Xn},第 i 类样本为 $X^i=\left\{x_1{}^i,x_2{}^i,...,x_{ni}{}^i\right\}$, 相异度矩阵为:

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \mathrm{M} & \mathrm{M} & \mathrm{M} & & \\ d(n,1) & d(n,2) & \mathrm{L} & \mathrm{L} & 0 \end{bmatrix} \tag{7}$$

$$d(i,j)=\frac{N_i(d_i)+N_j(d_j)}{n_i+n_j}\in[0,1] \tag{8}$$

$$d_{ij}=\sqrt{\frac{\sum_{k=0}^{n_i}(x_k{}^i-c_i)}{n_i}}+\sqrt{\frac{\sum_{k=0}^{n_j}(x_k{}^j-c_j)}{n_j}} \tag{9}$$

$$c_i=\frac{\sum_{i=1}^{n}x_i}{n} \tag{10}$$

式(8)为 i 类与 j 类的相异值, 式(9) d_{ij} 表示 i 类半径与 j 类半径之和, 式(8)中 d_i 表示 Xi 类样本点到 c_i (i 类的形心)的距离和到 c_j (j 类的形心)的距离之和, $N_i(d_i)$ 表示 Xi 类样本中 d_i 大于 d_{ij} 的样本个数, $N_i(d_i)$ 值越小, 两类越多重合, 相异值越接近 0, 越难分. 这种相异性度量进一步应用于类集合间的相异性度量.

3.3 自底向上构建决策树预警模型

由于支持向量机处理的是二分类问题, 因此基于支持向量机构建的决策树模型为二叉树, 各类对象作为叶子节点, 首先从所有类中选择相异性值小的两个类, 训练这两类构建一个 SVM, 生成第一个非叶结点 SVM1, 两类合并后形成新的类, 新类和剩下的类别再次进行相异度计算, 然后选择相异度小的两类类似构建 SVM2, 以此类推, 直到训练集中无类别可分, 最后的SVM构成决策树的根节点, 至此基于支持向量机的决策树就构建完成, 训练集中的类别构成叶子结点, 训练得到的SVM构成非叶结点[11]. 整个训练过程, 构造的决策树是介于一对一和一对其他多分类器之间的一棵最优二叉树, 执行效率较高.

具体构建过程如下:

Step1: 利用N个类构成具有N棵二叉树的初始集合 F={N1, N2, …,Nn}, 每棵二叉树只有一个根结点, 左右子树为空;

Step 2: 对F中各棵树对应的类进行类间相异度计

算, 构造相异度矩阵;

Step3: 根据相异度矩阵, 选择相异度值小的类对应的两棵子树, 作为新构造的二叉树的左右子树, 对左右子树对应的类别进行训练得到 SVM, 该二叉树的根结点为 SVM;

Step4: 从 F 中删除这两棵子树, 并把新的二叉树加入到集合 F 中;

Step5: 重复 2、3、4, 直到集合 F 中只有一棵二叉树为止, 该树即为自底向上构造的决策树.

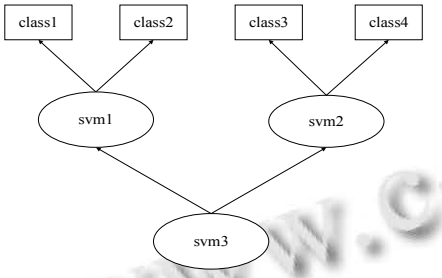


图 2 决策树训练过程

图 2 为决策树的构建过程, 决策树的构建自底向上, 每次选择相异度值小的两类进行 SVM 训练, 因为相异度值越小, 两类越相似, 越难以区分, 所以每次选择这样难以区分的类进行训练, 并合并为新类, 使得分类误差相对较少. 构建过程中, 各个类作为决策树的叶结点, 每两类训练形成一个 SVM, 作为非叶结点, N 个类共训练 N-1 个 SVM, 相比于一对一和一对其他的多分类器, 减少了 SVM 个数, 而且自底向上的构建过程使得每个类都只有一条分类路径, 相对减少训练时间.

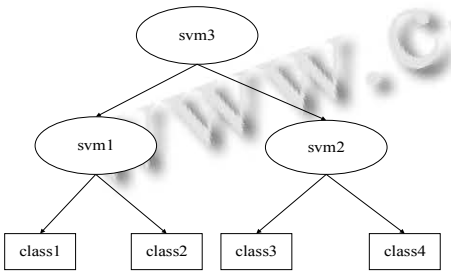


图 3 决策树测试过程

图 3 为分类过程, 利用构建的支持向量机决策树模型对选定的测试样本进行分类是一个自顶向下的过程, 从根节点的 SVM 开始分类, 若是叶节点则分类结束, 若是非叶结点, 则利用该节点进行 SVM 分类, 直到成为叶节点, 所有的叶节点构成所有类别, 分类完成.

4 实验证明

4.1 样本选取

实验选择 2014 年 1 月到 2016 年 1 月采集的安全生产数据, 对数据进行预处理, 选择第三部分筛选出的 12 个预警指标作为条件属性, 4 个安全等级安全、较安全、基本安全和危险作为 4 个类别, 2014 年 1 月到 2015 年 1 月的数据作为训练集, 2015 年 2 月到 2016 年 1 月的数据作为测试集.

4.2 实验结果与分析

本文采用两种方法进行模型训练实验, 一种是自底向上的决策树模型, 另一种是自顶向下的决策树模型, 两种方法均使用 libSVM 工具包完成, SVM 选择具有较强非线性映射能力的 RBF 核函数, 训练过程中不断优化惩罚因子 C 和 RBF 参数 δ.

模型训练完成, 在测试集上从三个方面评价选用模型的效果, 总体分类正确率为各类分类正确的样本总和占总体样本的数目, 反映模型的整体效果; 各类分类正确率为各类样本正确分类数目占该类测试样本数, 反映各类的分类效果; 误报率为实际为危险但分类为其他类的样本数目占分类为危险的样本数目, 这个指标对危险的预警至关重要.

分类结果评价如表 1 所示, 从表 1 可以看出本分类器能较好地分类, 在分类精度上, 本文所用模型较传统自顶向下模型准确率提升了 7%, 主要是因为自底向上错误积累相对较少; 误报率相对自顶向下的决策树模型也下降 2%, 降低了危险情况漏报的可能性, 对预警提供支持; 自底向上的决策树多分类模型, 一个类对应一条路径, 相比自顶向下的决策树多分类模型, 带权路径长度最短, 训练时间也相对缩短; 各安全级别也能明确分类, 实现对安全生产的安全状态实时预测. 需要强调的是, 实验结果与很多因素有密切关系, 例如数据集的特点, 核函数和参数的选择等.

表 1 模型分类结果

模型	总体分类正确率(%)	各类正确分类率(%)				误报率(%)	训练时间
		安全	较安全	基本安全	危险		

自底向上决策树模型	70.2	74	68.8	64.3	69	30.3	48.233
自顶向下决策树模型	62.8	70	69.1	63.6	68.7	32.3	48.399

5 总结

本文的研究重点是为安全生产监管中危化企业的隐患预警提供预警模型. 本文从多维度, 多层次考虑影响安全生产的隐患指标, 分析隐患特征, 采用基支持向量机的决策树预警模型对隐患状态分类预警, 通过对比实验, 证实本文采用模型具有很好的分类能力, 在精度和速度上都能满足需求, 使用决策树模型提取预警规则为未来隐患规则库的建成提供依据, 具有现实意义. 但是模型的效果一定程度上受样本数据的平衡性影响, 所以如何克服这方面的限制是未来研究方向.

参考文献

1 蒲洪涛.高危企业生产安全监管态势预警技术的研究与实现[硕士学位论文].哈尔滨:哈尔滨工业大学,2014.

2 张道斌.危化品企业安全生产预警预报体系现状探析.安全、健康和环境,2014,14(11):52–54.

3 Dora S, Subramanian K, Suresh S, Sundararajan N. Development of a self-regulating evolving spiking neural network for classification problem. Neurocomputing, 2016, 171.

4 Li JX, Qin YC. Feature selection for support vector machine in the study of financial early warning system. Quality & Reliability Engineering International, 2014, 30(6): 867–877.

5 周健,史秀志.矿井突水水源识别的距离判别分析模型.煤炭学报,2010,35(2):278–282.

6 张星联,张慧媛.基于神经网络的蔬菜农药残留风险预警模型研究.中国农业大学学报,2015,20(2):259–267.

7 金珠.改进的支持向量机分类算法及其在煤矿人因事故安全评价中的应用[博士学位论文].徐州:中国矿业大学,2011.

8 丁世飞,齐丙娟.支持向量机理论与算法研究综述.电子科技大学学报,2011,40(1):2–10.

9 胡俊.支持向量机与哈夫曼树实现多分类的研究.广东工业大学学报,2014.

10 赵天昀.一种改进的SVM决策树文本分类算法.情报杂志,2010,29(8):141–143.

11 乔增伟.一种基于支持向量机决策树多类分类器.计算机应用与软件,2009,26(11).