第 33 卷第 2 期　　河 北 科 技 大 学 学 报　　Vol. 33, No. 2
2012 年 4 月　　Journal of Hebei University of Science and Technology　　Apr. 2012

文章编号：1008-1542(2012)02-0171-04

机器学习与知识发现在高校公共突发事件智能预警系统中的应用

仇计清[1]，李晓华[2]，苏连青[1]

(1. 河北科技大学理学院，河北石家庄　050018；2. 河北科技大学党政办公室，河北石家庄　050018)

摘　要：机器学习与知识发现能够帮助人工智能系统对现象或信息之间的因果关系产生更深入的认识，有利于提高智能决策支持系统的工作效率，有利于提高使用者和机器之间的默契程度。就机器学习与知识发现技术应用于高校公共突发事件智能预警系统进行了探讨。

关键词：机器学习；知识发现；高校公共突发事件；智能预警系统

中图分类号：TP393　　**文献标志码**：A

Application of machine learning and knowledge discovery in intelligent early warning system of public emergencies in institutions

QIU Ji-qing[1], LI Xiao-hua[2], SU Lian-qing[1]

(1. College of Sciences, Hebei University of Science and Technology, Shijiazhuang Hebei 050018, China; 2. Administration Office, Hebei University of Science and Technology, Shijiazhuang Hebei 050018, China)

Abstract: Machine learning and knowledge discovery help the artificial intelligence system obtain more profound cognition of the causal relationship between phenomena and information, help IDSS(intelligence decision support system) enhance the working efficiency, and help improve the coordination between the users and machines. This paper discusses the application of machine learning and knowledge discovery to the intelligent early warning system of public emergencies in institutions.

Key words: machine learning; knowledge discovery; public emergencies in institutions; intelligent early warning system

高校公共突发事件是指在高等学校及其周边突然发生，造成或者可能造成重大伤亡、重大财产损失，能引发高校内部及社会连锁反应和严重后果，危及高校公共安全，对学校发展、社会稳定产生剧烈负面影响的自然、社会及群体性事件。预警是指在事件发生前进行预先警告，即高校突发公共事件职能部门对将来可能发生的危险进行事先的预报以提请相关当事人的注意。预警机制是指能灵敏、准确地昭示风险前兆，并能及时提供警示的机构、制度、网络、举措等构成的预警系统，其作用在于超前反馈、及时布置、防患于未然，从而最大限度地降低由于高校突发公共事件的发生对人民的生命财产造成的损失。高校公共突发事件智能预警系统的构建原则主要是及时、全面、高效和引导，达到及时预防、降低损害、保证安全和促进发展的功能。

1　机器学习与知识发现在智能预警中的作用

预警的实质就是利用科学手段对系统运行的“异常值”进行预见和警示，这种预见和警示的依据就是系

收稿日期：2011-11-21；责任编辑：李　穆
基金项目：国家自然科学基金资助项目(71040012)
作者简介：仇计清(1956-)，男，河北井陉人，教授，博士，主要从事复杂系统预测与优化控制方面的研究。

统运行过程中的大量输入数据和系统状态数据，这些大量乃至海量数据中蕴含了反映系统运行状态是否正常的信息，智能预警系统就是利用计算机人工智能方法从这些数据中分析和获取有价值的信息，从而对系统运行的“异常值”进行预见和警示。这些有价值的信息称为知识，知识的获取过程称为知识发现，而机器学习是知识发现的主要手段。目前，国内外已有不少关于智能预测以及智能预警系统方面的研究工作，但大都把研究工作的重点放在所建立系统模型的准确性上。事实上，一个预警模型不但要考虑所提出的预见和警示具有较强的准确性，更应该把预警结果与系统数据相关性的提取和分析作为研究工作的重点。

作为知识发现主要手段的机器学习，其核心是学习。关于学习迄今为止还没有一个精确的、能被大多数学者公认的定义。究其原因，一是因为进行知识发现问题研究的学者来自各种不同的学科，有着不同的知识背景，对具体问题有着不同的理解。二是因为学习是多侧面、多角度、综合性的心理活动，它与记忆方法、思维习惯、感知行为等多种心理和生理活动都有密切的联系，人们至今还没有准确把握学习的生物机理与实现过程。目前在机器学习领域影响较大且具有较大认同的观点是：学习是系统中的任何改进，这种改进使得系统在重复同样的工作或进行类似的工作时，能完成得更好[1]。机器学习的研究内容就是如何利用计算机来模拟人的学习行为，使其能自动地通过学习来获得所需知识和特殊技能，不断提高实际智能系统的性能，实现智能系统的进一步完善和功能的提升。

机器学习的理论和方法从提出伊始，就被认为是挖掘大量复杂数据的运行模式和数据相关性的有效方法之一。近年来机器学习已开始应用于智能预测和推断[2-4]，显然利用机器学习方法也可以对动态系统的演化过程进行分析、总结和归纳进而发现知识，所获得的知识就可以用来预测系统的演化趋势，由此实现对系统行为的干预和引导。通过机器学习方法所建立的实际动态系统演化模型均可以通过学习，不断完善系统内部各因素之间的复杂、非线性、强关联的因果关系。针对智能预警系统，机器学习方法可以通过对系统历史行为的学习，获取系统状态变化规律的知识，进而预示系统演化的趋势，或者根据对系统状态的分类学习为决策者提供可靠的系统干预策略。

2 知识发现的常见结果及常用技术方法

知识发现就是通过学习从数据集(对于智能预警系统而言，可理解为系统演化过程的大量历史数据以及系统的各种指标值等)中获取有价值知识，这些知识一般称为模式。知识发现的应用范围非常广泛，可以是工业、农业、科学、经济、社会、军事、商业等领域的动态系统数据，也可以是遥感卫星观测到的地理和气象数据，知识发现作为人工智能领域中一种新兴数据处理技术和方法得到成功应用，受到了来自不同领域学者的关注[5-6]。

知识发现的常见结果有以下5种。

1)广义型知识(Generalization) 广义型知识就是通过对数据微观特性的概括和抽象，获得的能够表征其普遍特性的、高层次的中观和宏观知识。

2)分类型知识(Classification & Clustering) 分类型知识是指利用决策树、统计、神经网络、粗糙集等分类方法，从半结构化或非结构化的海量数据中，提取出的同类事物共性特征和不同事物差异特征的知识。

3)关联型知识(Association) 关联型知识是通过在项集中寻找频繁项集，进而生成强关联规则的方法，获得的反映事件之间依赖或关联关系的知识。

4)预测型知识(Prediction) 预测型知识是通过利用统计、机器学习和神经网络等方法，建立各种回归模型，对具有时间序列特性的历史及当前数据进行分析计算，预测未来发展趋势。

5)偏差型知识(Deviation) 偏差型知识是通过对同类事物的聚类分析，根据离群值获取标准类之外的特例，从而得到差异和极端特例的中微观特征描述。

知识发现的常用技术方法如下。

1) 传统数理统计方法 利用数理统计知识，对数据集建立各种随机模型，进行相关性分析、回归分析、主成分分析和贝叶斯估计等。

2) 神经网络方法 神经网络方法是通过模拟人脑，利用神经元之间同时相互作用的动态过程来完成信息处理的生物过程，仿照生理神经网络结构构造一种非线性预测模型，通过使用历史数据对模型进行训练，达到模型对某些特殊模式的识别和判断。

3）决策树方法　决策树方法是首先利用归纳算法对训练样本集进行处理，生成分类规则和决策树，然后使用测试数据集校验修正决策树，逐步得到较为完善的分类方法。

4）遗传算法　遗传算法是一种基于生物进化理论，模拟生物进化的选择、交叉及变异的迭代过程构造的一种优化计算方法。

5）近邻算法　近邻算法是指在决策系统中，当系统需要预测未来情况或进行决策时，系统自动寻找与当前情况相近的案例，从中进行筛选，获取最佳的相同解决方案。

6）粗糙集方法　粗糙集方法主要用于对智能预警系统中的不完全或不完整信息进行描述和处理，通过发现不准确数据或噪声数据的内在关联结构，对其进行分类。

3　机器学习与知识发现在智能预警系统中的应用

近几年，已有一些学者把机器学习与知识发现应用于多种智能预警系统中[7-9]，笔者重点讨论知识发现中的决策树方法在高校公共突发事件智能决策支持系统中的应用。

3.1　决策树方法的具体步骤

决策树方法是首先利用归纳算法对训练样本集进行处理，生成分类规则和决策树，然后使用测试数据集校验修正决策树，逐步得到较为完善的分类方法。决策树方法首先针对训练样本集（称为数据集的输入空间或属性空间），采用自顶而下的递归方式：从根节点开始对每个节点依据给定标准选择一个属性进行测试，然后按照所选属性的一切可能值向下建立分枝，由此将训练样本进行划分，直到一个节点上的所有样本数据都被划分到同一个类，或者该节点中的样本数据个数低于给定值时为止，从而构造出一个二叉树或多叉树，就称其为决策树。本阶段结束后，把训练样本集划分为若干互斥的区域，然后对每个区域赋予一个标志来表示该区域内数据的特色属性。这一阶段称为树生成（或树构造）。前面的树构造过程所得的并不是最简单、紧凑的决策树，因为其中部分分枝反映的并不是训练样本数据的固有特性，而可能是训练样本数据中的噪声或孤立点，因此需要进行下一阶段工作，即树剪枝过程。树剪枝过程的意图就是检测并去除这些噪声或孤立点对应的分枝，以提高对未来数据进行分类的准确性。树剪枝有先剪枝方法、后剪枝方法或两者相互结合的方法。树剪枝必须先确定一种剪枝标准，常用的有期望错误率最小原则和最小描述长度原则（MDL）。

采用期望错误率最小原则来构造决策树时必须选择一个误差指标 $E(t)$，作为量化节点 t（表示某一属性）从不同区域中分叉数据（或事件）的性能指标，该指标表示了节点 t 为噪声或孤立点的程度。把决策树中各节点的误差指标称为杂质函数，对某一节点，如果给定数据集的数据都属于同一分类，杂质函数取得最小值 0；如果给定数据集的数据均匀分布于所有可能的分类时，杂质函数就达到最大值 1。

通常选取熵函数或 Gini 指标函数作为杂质函数。

熵函数公式为

$$\Phi_e(p_1,p_2,\cdots,p_J)=-\sum_{j=1}^{J}p_j\ln p_j。$$

Gini 指标函数为

$$\Phi_g(p_1,p_2,\cdots,p_J)=-\sum_{i\neq j}p_ip_j=1-\sum_{i=1}^{J}p_i^2。$$

树生成得到的未经过剪枝的决策树规模通常很大，而且同训练样本集有较大偏差。因此，对于测试样本集，用这种树往往不能得到期望输出，而对于训练样本集即使能得到期望输出，但在精度上也并不可靠。在树剪枝过程中，为得到合适规模的决策树，使用的方法有多种。其中，最有效的 2 种方法是基于最小代价复杂性方法和基于最弱子树收缩原理方法。

3.2　计算结果与分析

高校公共突发事件智能预警系统的核心就是利用先进的信息技术手段，采集、存储、挖掘、分析高校公共突发事件的有关信息，最终实现对高校公共突发事件的预警决策。机器学习、知识发现为从繁杂的历史数据中挖掘出事件发生与各种起因的关联关系提供了方法支持。机器学习和知识发现可以应用于各类高校公共突发事件之中，下面以校园火灾事件为例，说明利用决策树方法进行决策的过程。

根据近年来校园火灾事件发生的网上调查数据（见表 1），共计 8 个样本，考虑了电器使用、燃气炉使用、

电路老化和偶然因素等指标。对火灾事件的严重程度主要考虑人员伤亡(RYSW)和财产损失(CCSS)2个方面,用RYSW=2人和CCSS=0.5万元作为分界值,将严重程度分为较高和较低2类。利用上述数据,应用决策树方法进行计算,得到分类结果如表2所示。

表1 校园火灾事件网上调查数据

Tab.1 Data of institution fire scenes on web

样本	DQ 电器使用	RQ 燃气炉使用	DL 电路老化	OR 偶然因素
1	√			
2	√			
3	√			
4		√		
5		√		
6				√
7			√	
8	√			

表2 应用决策树算法对数据进行计算的结果

Tab.2 Result after decision tree algorithm processing

	IF	THEN	置信度/%	支持度/%
1	DQ or RQ	RYSW>2人 OR CCSS>0.5万元	100	62.5
2	DL	CCSS>0.5万元	100	12.5
3	OR	CCSS<0.5万元	50	12.5

从表2可以看出,电器使用、燃气炉使用对发生重大火灾事件的影响非常大,电路老化的影响次之,而与偶然因素基本没有关系。根据决策树方法的计算结果,在智能预警系统中,确定影响校园火灾事件发生的各因素权重,得出事件发生可能性的量化模型,利用一定的阈值触发预警系统发出预警。

由于本例采样数据的容量较小,只重点考虑了人员伤亡及财产损失与各影响因素间的关系,关于事故严重程度的描述比较简单。当然,对于大容量的训练样本集,应该事先对数据集进行相应的预处理,比如利用统计方法分析人员伤亡及财产损失与各参数变化的相关关系模型,使对事故严重程度的描述进一步科学细化。智能预警系统的目的,就是根据历史数据,应用各种数据处理方法使系统获取知识,然后根据当前的某些因素指标值,对突发事件的发生作出预见和警示,进而为应急管理提供决策支持。

4 结 语

随着高等学校内部管理体制和监测体系的进一步完善,其智能预警系统所积累的数据也会更加庞大,从系统数据库中利用机器学习方法获得相应的知识,从而建立高校公共突发事件生成演化规律的知识库,可以为高校公共突发事件的预防、预警以及应急管理提供技术和方法支撑。

参考文献:

[1] 杨炳儒. 知识工程与知识发现[M]. 北京:冶金工业出版社, 2000.

[2] RECKNAGEL F. Applications of machine learning to ecological modeling[J]. Ecological Modelling, 2001, 146(1-3):303-310.

[3] BOBBIN J, RECKNAGEL F. Knowledge discovery for prediction and explanation of blue-green algal dynamics in lakes by evolutionary algorithms[J]. Ecological Modelling, 2001, 146(1-3): 253-262.

[4] BOBBIN J, RECKNAGEL F. Inducing explanatory rules for the prediction of algal blooms by genetic algorithms[J]. Environment International, 2001, 27(2-3): 237-242.

[5] 李 玮,沙占友,于健骐,等.数据触合技术在火灾报警系统中的应用研究[J].河北科技大学学报(Journal of Hebei University of Science and Technology),2010,31(2):112-116.

[6] 原建伟.基于内容分析的数据挖掘研究[J].河北工业科技(Hebei Journal of Industrial Science and Technology),2011,28(5):299-302.

[7] 沈 菲,王洪礼,冯剑丰,等. 知识发现在赤潮预测预警系统研究中的应用[J]. 海洋技术(Ocean Technology),2003, 22(2):19-22.

[8] 蔡如钰. 基于人工神经网络的夜光藻密度预测模型[J]. 中国环境监测(Environmental Monitoring in China), 2001, 17(3):52-55.

[9] 侯旭东, 张 兢. 基于模糊神经网络融合技术的智能火灾预警系统[J]. 重庆工学院学报(自然科学)(Journal of Chongqing Institute of Technology(Natural Science)), 2008, 22(9):141-144.