

一种用于情报威胁评估的数据分类算法研究

王卓君

(解放军国际关系学院 南京 210039)

摘要 针对情报威胁评估的实际需要,提出了一种面向情报数据威胁评估的情报数据分类算法,突破了原有的单层分类算法框架,构造了双层数据分类框架,设计和实现了各层的算法,并验证了相关算法的正确性和有效性,扩展了现有侦察情报分析系统的战场态势监控和辅助决策功能。

关键词 军事情报 威胁评估 分类算法 双层

中图分类号 G350.7

文献标识码 A

文章编号 1002-1965(2011)10-0156-07

Research On a Data Classifying Algorithm in Threat Assessment

WANG Zhuojun

(International Studies University of PLA, Nanjing 210039)

Abstract Be aimed at the needs of threat assessment, this thesis provided a information data classifying algorithm which oriented information data threat assessment, it broke through the original single layer classifying algorithm framework and constructed double layer data classifying algorithm framework, it devised the and realized the algorithm of every layer, and tested the rightness and effectiveness of related algorithms, then extended the existing function of battlefield situation supervise and command policy in reconnaissance system.

Key words military information threat assessment classifying algorithm double layer

0 引言

威胁评估是对敌方杀伤能力及对我方威胁程度的评估,是在态势估计的基础上,依据敌我兵力和武器、电子设备性能、敌作战企图和敌我双方的作战策略,以定量形式对敌方威胁程度做出的分析和评估^[1]。情报分析作为一个知识发现的过程,是生成战场态势一个重要的启动条件和作业条件,在整个威胁评估过程中有着举足轻重的作用。例如:美国陆军目前使用的轻型全源分析系统(ASAS-1)能够提供自动情报分析、实现战场可视化、进行情报和电子战资源管理、生产并分发情报,可自动地对全部雷达情报、卫星情报、预警机情报、部队侦察情报、技术侦察情报、远方情报(含上级及友邻通报等)等多源情报资料进行综合处理,实现情报分析去伪存真的功能,并产生总态势图及总表格,进行目标识别和威胁判断。

目前对威胁评估的研究方法很多,主要有:层次分析法、支持向量机方法、神经网络方法、属性分析法和变权理论法等^[2-4]。其中,对目标威胁判断常用的层

次分析法^[5]主观性太强,而支持向量机(SVM)^[6]的方法其核函数选取困难且精度不高,神经网络方法以其对噪音数据的高承受能力和对未训练数据的分类能力显著优势被广泛地用于数据的分类、聚类、特征挖掘、预测和模式识别等方面。同时,神经网络最大的不足是需要较长的训练时间并且可解释性较差。

为了克服神经网络的这一弱点,本文尝试利用决策树算法中ID3算法^[7]和神经网络算法中的Boltzmann机^[8]构建出一种新的情报数据分类算法用于威胁评估模型的开发设计中。决策树算法^[9]是一种从训练样本集中推理出判定树表示形式的分类规则的方法。其优点在于它的直观性和易理解性,该算法不仅能做出分类和预测,而且它的生成过程、分类、预测以及从中所提取的分类规则都具有很强的可理解性。同时,本文针对Boltzmann机和ID3算法尚存的不足做出改进,改善了Boltzmann机在训练过程中易出现网络麻痹与温度训练过拟合的问题,同时降低了ID3的计算复杂度,加快了情报威胁评估整体的速度。

1 情报数据分类算法模型

1.1 情报数据分类算法的基本构成 本文研究的情报数据分类算法是由 Boltzmann 机和 ID3 算法组合形成的双层数据分类算法,可针对各类战场传感器采集的战场态势情报数据进行威胁评估,并将处理结果以规范化、结构化的形式提交后续的情报规则生成环节。

与普通数据分类不同,此处的情报数据威胁评估面向的数据是一个实时的、到达次序独立的、不受应用系统所控制的、规模宏大且不能预知其最大值并且一经处理不特意保留的实时战场态势情报数据。考虑到该种情报数据的特殊性以及对分类数据算法特性的分析,分别对 Boltzmann 机的 Sigmoid 函数和 ID3 中信息熵值进行了改进,从而改善了 Boltzmann 机在训练过程中易出现网络麻痹与温度训练过拟合的问题,并且降低了 ID3 的计算复杂度,加快了建树的速度。

情报数据分类算法的第一层是基于改进后的 Boltzmann 机的快速计算层,第二层是基于改进后的 ID3 算法的精确分析层。情报数据分类的目的是对战场地理环境、部队单位编成、武器装备和周边敌我双方兵力部署情况根据以往作战时产生的规则进行威胁等级评估分类,为部队作战提供决策。因此,第一层产生的结果是以概率形式出现的初步分类结果,其分类结果进入第二层时,数据开始重新根据属性集中各个属性出现的相对频率进行决策树节点分裂计算,选择属性出现频率最大的为决策树节点分裂值。对进入第二层的情报数据属性集进行反复计算,形成决策树,得出当前威胁等级的评估情况,并根据此情况为部队作战提供决策。双层情报数据分类算法原理示意图如图 1 所示。

1.2 情报数据分类算法的算法原理与改进

1.2.1 Boltzmann 机算法原理与改进。作为第一层数据快速计算层中的核心算法,依据情报数据自身的特点,这里选择 Boltzmann 机是较为合适的,Boltzmann 行动选择策略适于求解非精确状态信息下的顺序决策过程问题的行动选择策略。针对实时进行采集的情报数据,它采用随机接收准则选择分类类型,并根据当前状态下可选类型的估计价值决定选择类型的概

率,这使得分类算法有可能跳出分类空间中局部最优子空间的陷阱,寻找到最优的分类策略,并且在从高温到低温的退火中,能量并不会停留在局部极小值上,而是以最大的概率到达全局最小值。

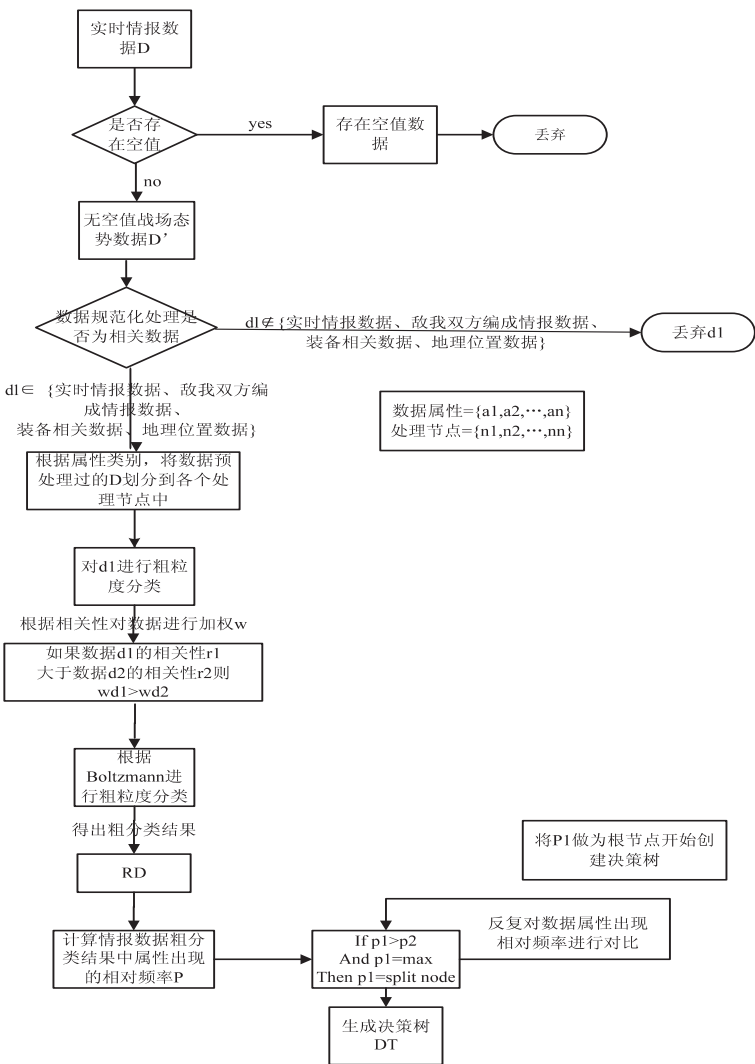


图 1 情报数据分类算法原理示意图

Boltzmann 机由能量函数 E 所表征,能量函数的值由机器的个体神经元占据的特定状态所决定:

$$E(x)=-\frac{1}{2}\sum_i\sum_{j_i}w_{ji}x_ix_j\tag{1}$$

根据公式(1),求 $L(w)$ 对 w_{ji} 的微分:

$$\frac{\partial L(w)}{\partial w_{ij}}=\frac{1}{T}\big(\sum_{X\in\tau}\big(\sum_{x_\beta}P(X_\beta=x_\beta\mid X_\alpha=x_\alpha)x_jx_i-\sum_xP(X=x)x_jx_i\big)\big)\tag{2}$$

$$\frac{\partial L(w)}{\partial w_{ij}}=\frac{1}{T}\big(\sum_{x_\alpha}\frac{1}{P(X_\alpha=x_\alpha)}\frac{\partial P(X_\alpha=x_\alpha)}{\partial w_{ij}}\big)\tag{3}$$

根据边缘概率推导,这里引入单个事件 x_α 及联合事件 x_β 和 x 。

$$x_\alpha:X_j=x_j,x_\beta:\{X_i=x_i\}_{i=1}^K,i\neq j,X:\{X_i=x_i\}_{i=1}^K\tag{4}$$

实际上, x_β 联合事件排斥 x_α ,联合事件 x 包括 x_α

和 $x\beta$ 。 $x\beta$ 的概率是 x 关于的边缘概率。

$$P(X_a = x_a) = \sum_{x_a \in \tau} P(x_a, x_\beta) = \sum_{x_\beta} P(X = x) \quad (5)$$

$$P(X = x) = P(X = x | X_a = x_a) P(X_a = x_a) \quad (6)$$

这个关系定义联合事件 $X = x = P(x_a, x_\beta)$ 的概率。这时,偏导数

$$\frac{\partial L(w)}{\partial w_{ji}} = \sum_{X \in \tau} \sum_{x_\beta} \frac{P(X = x | X_a = x_a)}{P(X = x)} \frac{\partial P(X = x)}{\partial w_{ji}} \quad (7)$$

可以写成:

$$P(X = x) = \prod_j \varphi\left(\frac{x_j}{T} \left(\sum_{i < j} w_{ji} x_i\right)\right) \quad (8)$$

$\varphi(\cdot)$ 为其变元的 sigmoid 函数。

具体算法流程图 2 如下:

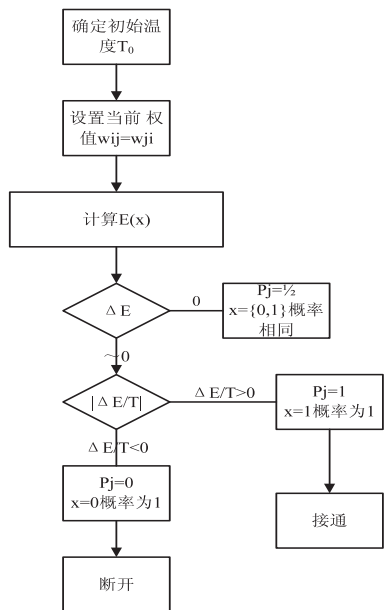


图2 Boltzmann 机算法流程图

虽然 Boltzmann 机可用作最邻近模式匹配器和存储器,也能够用来求解组合优化等问题,但仍存在着训练时间长和对统计错误敏感的问题。在实际应用中,如果不采取这些启发式的方法,将导致较慢的收敛速度、较差的推广能力等不理想的结果。对网络模型改进的主要目标有两个:防止网络训练过程中麻痹现象的出现,提高网络的训练速度;提高网络的泛化能力,避免过拟合现象。

由于 Boltzmann 机是基于梯度下降法进行训练的,所以网络的激活函数要求连续可微,参数导数的存在性对学习至关重要,因此 Boltzmann 机网络一般不采用阈值函数和符号函数作为激活函数^[10]。Boltzmann 机的激活函数一般要求非线性,否则多层网络将不提供高于两层网络之上的任何计算能力。有界性也是激活函数的一个条件,这可以限定权值和单元输出的上下边界,使训练次数也有限,如果输出是代表一个概率时,有界性尤其重要。单调性也是激活函数的一

个期望的性质,因为如果激活函数在定义域中不是单调的,存在一个或多个极值,则会延长训练时间并对错误敏感。

当 sigmoid 函数的输出接近饱和值时,其梯度很小,相应的权值调节量也很小,学习速度很慢,这就是麻痹现象产生的原因。一旦产生网络麻痹,则会不断地对采集过程中产生的过程数据中的某些属性值不断进行退火降温,拉长数据学习时间,从而降低整个数据分类效率。为了防止这种现象产生,本文提出了对 sigmoid 函数的输出进行了限制的方法,限制其最大输出值小于饱和值,改进的 sigmoid 函数为:

$$\begin{aligned} |\text{net}| &< \frac{x_i}{T} - \varepsilon \\ \text{net}| &> \frac{x_i}{T} - \varepsilon \quad \varphi|\text{net}| = \begin{cases} \frac{x_i}{T} \sum_{i < j} w_{ij} x_i \\ \rho \end{cases} \end{aligned} \quad (9)$$

其中: $\varphi(\frac{x_i}{T} - \varepsilon) = \rho$, ρ 为 $0 \sim \frac{x_i}{T}$ 之间的数, ρ 可

取比 $\frac{x_i}{T}$ 稍小的数,例如 $\frac{x_i}{T}$ 取 2.3 时, ρ 可以取为 2.15 - 2.25。sigmoid 函数的梯度的最小值 $\varepsilon(\partial L(w)/\partial w_{ij}) \min = \eta(p_{ij}^2 - \rho^2)$, 可对 sigmoid 函数的梯度设置最小值限制,这样权值的误差调节就可以得到有效控制,并减低网络麻痹现象出现的频率。

1.2.2 ID3 算法原理与改进。ID3 的基本原理是基于二叉分类问题,但很容易将其扩展到多叉分类上。假设训练集中共有 m 个样本,样本分别属于 c 个不同的类,每个类的预设训练实例集为 X ,学习的目的是将训练实例分为 n 类,记为 $C = \{X_1, X_2, \dots, X_n\}$ 。设第 i 类的训练实例个数是 $|X_i| = C_i$, X 中总的训练实例个数为 $|X|$,记一个实例属于第 i 类的概率为 $P(X_i)$,则有:

$$P(X_i) = \frac{C_i}{|X|} \quad (10)$$

此时决策树对划分 C 的不确定程度为 $I(X, C)$, 简记为 $I(X)$:

$$I(X) = - \sum P(X_i) \log_2 P(X_i) \quad (11)$$

对熵压缩的度量过程就是缩小对数据划分不确定程度的过程。若选择测试属性 A 进行测试,设属性 A 具有性质 $a_1, a_2, a_3, \dots, a_l$, 在 $A = a_j$ 的情况下属于第 i 类的实例个数为 C_{ij} , 即为测试属性 A 的取值为 a_j 时,它属于第 i 类的概率。记为 $A = a_j$ 时的实例集,此时决策树对分类的不确定程度就是训练实例集对属性 A 的条件熵:

$$I(Y_j) = - \sum P(X_i) \log_2 P(X_i | A = a_j) \quad (12)$$

叶结点 X_j 对于分类信息的信息熵为:

$$I(X|A) = \sum_j P(A = a_j) I(Y_j) = - \sum_i \sum_j P(X_i|A = a_j) \log_2(X_i|A = a_j) \tag{13}$$

即属性 A 的熵压缩为:

$$\text{Gain}(A) = I(X) - I(X|A) \tag{14}$$

其中, $I(X|A)$ 越小, $\text{Gain}(A)$ 的值越大。说明选择测试属性 A 对于分类提供的信息越大, 选择 A 之后对分类的不确定程度越小。

ID3 算法是把信息熵作为选择测试属性的标准, 即树结点的选择策略。但在计算基于属性的信息熵时, 公式比较复杂, 计算量较大, 相应的复杂度也高, 当数据量很大的时候很耗费硬件资源, 计算花费的时间较长。

改进后的 ID3 算法结合洛伦茨曲线思想, 设属性划分绝对平等曲线和实际属性划分曲线之间的面积为 A, 实际属性划分曲线右下方的面积是 B。并以 A 除以 A+B 的商表示不平等程度。如果 A 为零, 系数为零, 表示属性划分完全平等; 如果 B 为零则系数为 1, 属性划分绝对不平等。曲线的弧度越大, 那么系数也越大。具体曲线关系图 3 如下:

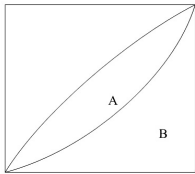


图 3 绝对平等曲线与实际属性划分曲线图

此算法区别于传统决策树计算期望信息的方法, 以往在计算不同类的信息概率后:

首先, 将计算后的所有值进行相减, 得出分类期望信息 - $\sum_{i=1}^m p_i \log_2(p_i)$

随后, 再分别计算对类中不同的属性的熵, 对这些熵进行相加, 得出子集的熵 - $\sum_{i=1}^m \frac{s_1 + s_2 + \cdots + s_m}{s} I(s_1 + s_2 + \cdots + s_m)$

最后, 在将期望信息与子集的熵相减得出这个分支上的编码信息:

$$\text{Gain}(A) = - \sum_{i=1}^m p_i \log_2(p_i) - \sum_{i=1}^m \frac{s_1 + s_2 + \cdots + s_m}{s} I(s_1 + s_2 + \cdots + s_m) \tag{15}$$

这样的计算步骤繁杂, 在计算机语言中难以表达, 故此, 本文提出一种反向熵压缩度量法, 算法中对 $I(s_1, s_2, \cdots, s_m)$ 的定义进行改进, 缩小分析的粒度, 立足点设立在每个集合中的属性分布情况, 从而降低测试复杂度, 减少计算时间, 下面对反向熵度量做出介绍:

在这里, $I(s_1 + s_2 + \cdots + s_m)$ 是一个计算根节点分

裂的关键要素, 是类中每个属性的信息值。当对根节点进行分裂时, 直接计算数据集中每个类中不同属性的熵值, 根据每个类中属性值总合的大小对整个数据集进行分裂。熵值越小, 子集划分纯度越高。

$$I(s) = 1 - (- \sum_{i=1}^m p_i \log_2(p_i)) \tag{16}$$

其中 P_i 是属性 i 在属性集中出现的相对频率。

如果类 I 按照某个划分点分成 I1 和 I2, 则划分后的属性信息和为:

$$RE = \frac{n_1}{n} * I(s_1) + \frac{n_2}{n} * I(s_2) \tag{17}$$

其中 n, n_1, n_2 分别为 I, I_1, I_2 的记录数。RE 值越小, 表明划分规则越好。最小的 RE 值就被划分为最优的节点分裂标准。

2 情报威胁评估模型实例分析

2.1 情报数据快速分类层 这里采用一组伊拉克战争的模拟数据做为分析处理的数据源, 当情报数据进入数据快速计算层时, 采用抽取主表进行分析的策略。这里选择陆战编成情报模拟数据和陆战当前状态情报模拟数据这两类情报数据作为数据快速计算层分析的主要内容。

数据有规定的范围, 数据规范化后给定一个阈值 $\text{Threshold}_{\text{xtk}} > 0 \parallel \text{Threshold}_{\text{xhp}} > 0 \parallel \text{Threshold}_{\text{xdd}} > 0 \parallel \text{Threshold}_{\text{pbwql}} > 0$ 为真时, Threshold 相应的位取值为 1, 否则, 相应的位取值为 0。Threshold 取值确定后, 再根据样例取值中该类的状态总数决定布尔量化结果。量化结果如表 1 所示:

表 1 布尔量化结果表

Threshold _{xtk, xhp, xdd, pbwql}	布尔量化结果
0000	9
0001	0
0010	0
0011	0
0100	4
0101	1
0110	1
0111	0
1000	3
1001	0
1010	0
1011	0
1100	1
1101	1
1110	0
1111	0

任取初始权重 $w_1 = 0.5, w_2 = 0.4, w_3 = 0.2, w_4 = 0.3$ 。训练集由状态 {0000, 0100, 0101, 0110, 1000,

1100,1101}组成的,它们的阈值分别为-0.9,-0.2,-0.3,0.7,初始温度为0.25,0.5,1。

计算热平衡概率分布,首先计算某一状态下各节点单元激活函数值。

状态 $v_1v_2v_3v_4 = (0000)$:
 $A_1 = w_1v_1 - \theta_1 = 0.9A_2 = w_2v_2 - \theta_2 = 0.2A_3 = w_3v_3 - \theta_3 = 0.3A_4 = w_4v_4 - \theta_4 = -0.7$
状态 $v_1v_2v_3v_4 = (0100)$:
 $A_1 = w_1v_1 - \theta_1 = 0.9A_2 = w_2v_2 - \theta_2 = 0.6A_3 = w_3v_3 - \theta_3 = 0.3A_4 = w_4v_4 - \theta_4 = -0.7$
状态 $v_1v_2v_3v_4 = (0101)$:
 $A_1 = w_1v_1 - \theta_1 = 0.9A_2 = w_2v_2 - \theta_2 = 0.6A_3 = w_3v_3 - \theta_3 = 0.3A_4 = w_4v_4 - \theta_4 = -0.4$
状态 $v_1v_2v_3v_4 = (0110)$:
 $A_1 = w_1v_1 - \theta_1 = 0.9A_2 = w_2v_2 - \theta_2 = 0.2A_3 = w_3v_3 - \theta_3 = 0.5A_4 = w_4v_4 - \theta_4 = -0.7$
状态 $v_1v_2v_3v_4 = (1000)$:
 $A_1 = w_1v_1 - \theta_1 = 1.4A_2 = w_2v_2 - \theta_2 = 0.2A_3 = w_3v_3 - \theta_3 = 0.3A_4 = w_4v_4 - \theta_4 = -0.7$
状态 $v_1v_2v_3v_4 = (1100)$:
 $A_1 = w_1v_1 - \theta_1 = 1.4A_2 = w_2v_2 - \theta_2 = 0.6A_3 = w_3v_3 - \theta_3 = 0.3A_4 = w_4v_4 - \theta_4 = -0.7$
状态 $v_1v_2v_3v_4 = (1101)$:
 $A_1 = w_1v_1 - \theta_1 = 1.4A_2 = w_2v_2 - \theta_2 = 0.6A_3 = w_3v_3 - \theta_3 = 0.3A_4 = w_4v_4 - \theta_4 = -0.4$
最终状态 $v_1v_2v_3v_4 = (1111)$:
 $A_1 = w_1v_1 - \theta_1 = 1.4A_2 = w_2v_2 - \theta_2 = 0.6A_3 = w_3v_3 - \theta_3 = 0.5A_4 = w_4v_4 - \theta_4 = -0.4$
当属性状态从0转移至1时,

$$\eta = \frac{\varepsilon}{T}$$

(18)

$$\varepsilon = T\eta = \frac{T}{t} = \frac{0.25}{32} = 0.008$$

在初始设定温度为0.25时,当 $\varphi(s) < 3.992$,取

$\frac{v_j}{T} \sum_{i < j} w_{ij}v_i$;当 $\varphi(s) > 3.992$,取 ρ 值。 ρ 取{0,3.992}之间的一个值,这里定为3.9;当在初始设定温度为0.5时,则 ρ 取{0,3.994}之间的一个值,这里也定为3.9;当在初始设定温度为1时,则 ρ 取{0,3.968}之间的一个值,这里同样定为3.9。

当神经单元状态(0000)转移到(1111)的概率为 $\frac{\rho_1(1)}{4}$,计算 $\rho(A_1)$:

$$\rho_A(1111) = \prod_j \varphi(\frac{v_2(1111)}{T}) \sum_{i < j} w_2v_2(0100) = 0.576$$

$$\frac{\rho_1(1)}{4} = \frac{\rho_{A1}(1111)}{4} = 0.144$$

因为 $0.144 < 3.992$,所以还是采用 $\frac{1}{1 + \exp(-s)}$ 的值。此后计算和上面雷同,不再重复计算,直接得出表2:

表2 第一层 Boltzmann 机网络训练结果表

网络状态	节点	T=0.25		T=0.5		T=1	
		P(1)	P(0)	P(1)	P(0)	P(1)	P(0)
S(0000)	1	0.63	0.37	3.315	0.685	0.1575	0.8425
	2	0.048	0.952	0.028	0.972	0.012	0.988
	3	0.03	0.97	0.015	0.985	0.0075	0.9925
	4	0.084	0.916	0.042	0.958	0.021	0.979
S(0100)	1	0.63	0.37	0.315	0.685	0.1575	0.8425
	2	0.144	0.856	0.072	0.928	0.036	0.974
	3	0.03	0.97	0.015	0.985	0.0075	0.9925
	4	0.084	0.916	0.042	0.958	0.021	0.979
S(0101)	1	0.63	0.37	0.315	0.685	0.1575	0.8425
	2	0.144	0.856	0.072	0.928	0.036	0.974
	3	0.03	0.97	0.015	0.985	0.0075	0.9925
	4	0.048	0.952	0.024	0.976	0.012	0.988
S(0110)	1	0.63	0.37	0.315	0.685	0.1575	0.8425
	2	0.144	0.856	0.072	0.928	0.036	0.974
	3	0.05	0.95	0.025	0.975	0.0125	0.9875
	4	0.084	0.916	0.042	0.958	0.021	0.979
S(1000)	1	0.98	0.02	0.225	0.775	0.245	0.755
	2	0.048	0.952	0.028	0.972	0.012	0.988
	3	0.03	0.97	0.015	0.985	0.0075	0.9925
	4	0.084	0.916	0.042	0.958	0.021	0.979
S(1100)	1	0.98	0.02	0.225	0.775	0.245	0.755
	2	0.144	0.856	0.072	0.928	0.036	0.974
	3	0.03	0.97	0.015	0.985	0.0075	0.9925
	4	0.084	0.916	0.042	0.958	0.021	0.979
S(1101)	1	0.98	0.02	0.225	0.775	0.245	0.755
	2	0.144	0.856	0.072	0.928	0.036	0.974
	3	0.03	0.97	0.015	0.985	0.0075	0.9925
	4	0.048	0.952	0.024	0.976	0.012	0.988

这里对陆战情报模拟数据中的火力打击能力进行情报分析,根据P(1)的综合指标指数初步判定火力打击水平PSD,P(1)综合值越高,打击水平越强。根据P(0)的综合指标指数初步判定危险级别DL,P(0)综合值越小危险级别越高,结果见表3:

2.2 情报数据精确分类层 目前尚未分析处理的数据还有:美伊双方编成情报模拟数据、美伊双方部队编成准备情报模拟数据,只有对这些数据进行充分的挖掘处理,才能得到最终想要的结果。已分析过的数据中包含{zbx,zby,xtk,xhp,xdd,pbwql,bcxh,f}属性,下面展开对这些属性反向熵值的计算。

类标号f有两个不同的值(即{1,2},1代表美军,2代表伊军),因此有两个不同类C=2。设C1对应1,C2对应2。类1中有12个样本,类2中有8个样本,随后基于C中不同的类别开始计算属性的反向熵值。

首先从陆战编成情报模拟数据和陆战当前状态情报模拟数据中的属性开始,如表4:

表 3 评估结果表

Xtk	xhp	xdd	pbwql	P(1)	PSD	P(0)	DL
0	4	0	0	0.888	中等	3.112	高
0	9	0	0	0.888	中等	3.112	高
0	6	0	0	0.888	中等	3.112	高
0	0	0	0	0.792	弱	3.208	高
0	0	0	0	0.792	弱	3.208	高
0	0	0	0	0.792	弱	3.208	高
10	10	0	0	1.238	强	2.762	低
10	0	0	0	1.142	强	2.585	低
0	0	0	0	0.792	弱	3.208	高
10	0	0	0	1.142	强	2.585	低
0	10	0	50014	0.852	中等	3.148	低
0	9	12	0	0.89	中等	3.11	中等
10	10	0	50014	1.202	强	2.798	低
0	0	0	0	0.792	弱	3.208	高
0	0	0	0	0.792	弱	3.208	高
0	0	0	0	0.792	弱	3.208	高
0	0	0	0	0.792	弱	3.208	高
0	3	0	0	0.888	中等	3.112	高

表 4 陆战属性计算表

zbx<=1000;	A ₁₁ =5 A ₂₁ =6	I(A ₁₁ ,A ₂₁)=2.995
zbx>1000;	A ₁₂ =7 A ₂₂ =2	I(A ₁₂ ,A ₂₂)=2.863
zby<=1000;	A ₁₃ =1 A ₂₃ =0	I(A ₁₃ ,A ₂₃)=2
zby>1000;	A ₁₄ =11 A ₂₄ =8	I(A ₁₄ ,A ₂₄)=2.981
xtk=0;	A ₁₅ =8 A ₂₅ =7	I(A ₁₅ ,A ₂₅)=3.421
xtk<=10;	A ₁₆ =4 A ₂₆ =1	I(A ₁₆ ,A ₂₆)=2.786
xhp=0;	A ₁₇ =6 A ₂₇ =6	I(A ₁₇ ,A ₂₇)=3
xhp<=10;	A ₁₈ =6 A ₂₈ =2	I(A ₁₈ ,A ₂₈)=2.822
xdd=0;	A ₁₉ =11 A ₂₉ =8	I(A ₁₉ ,A ₂₉)=2.981
xdd<=10;	A ₁₁₀ =1 A ₂₁₀ =0	I(A ₁₁₀ ,A ₂₁₀)=2
bctx<001001001002000	A ₁₁₁ =7 A ₂₁₁ =2	I(A ₁₁₁ ,A ₂₁₁)=2.863
bctx>001001001002000	A ₁₁₂ =5 A ₂₁₂ =6	I(A ₁₁₂ ,A ₂₁₂)=2.995

根据公式(16)、公式(17)对每一个属性的反向熵值进行计算：

$$RE(zbx)=\frac{11}{20} * 2.995 + \frac{9}{20} * 2.863 = 2.935$$

$$RE(zby)=\frac{1}{20} * 2 + \frac{19}{20} * 2.981 = 2.932$$

$$RE(xtk)=\frac{15}{20} * 3.421 + \frac{5}{20} * 2.786 = 3.263$$

$$RE(xhp)=\frac{12}{20} * 3 + \frac{8}{20} * 2.822 = 2.929$$

$$RE(xdd)=\frac{19}{20} * 2.981 + \frac{1}{20} * 2 = 2.932$$

$$RE(bctx)=\frac{9}{20} * 2.863 + \frac{11}{20} * 2.995 = 2.935$$

目前最小的反向熵值属性是 xhp,所以在决策树的根节点处选择 xhp 作为其分裂节点。

随后对美军编成情报模拟数据、伊军编成情报模拟数据中的数据实行进行属性熵值分析,这时需要重新划分类 C。美军编成情报模拟数据、伊军编成情报

模拟数据中的类标号属性 bctx 有两类不同的值,分别是 { 001001001000001, …, 001001001001000 } 与 {001001001001001, …, 001001001002000},因此这时也存在两个不同的类 C = 2。设 C1 对应 {001001001000001, …, 001001001001000}, C2 对应 {001001001001001, …, 001001001002000}。

计算美军编成情报模拟数据中的属性熵值,如表 5:

表 5 美军编成属性计算表

部队编号=038	A ₁₁₃ =1 A ₁₁₄ =0	I(A ₁₁₃ ,A ₁₁₄)=2
部队编号=038001…038100	A ₁₁₅ =8 A ₁₁₆ =2	I(A ₁₁₅ ,A ₁₁₆)=2.722
部队编号=038001001…038009001	A ₁₁₇ =7 A ₁₁₈ =2	I(A ₁₁₇ ,A ₁₁₈)=2.863
编制人数≤500	A ₁₁₉ =7 A ₁₂₀ =9	I(A ₁₁₉ ,A ₁₂₀)=3.047
编制人数≥500	A ₁₂₁ =2 A ₁₂₂ =2	I(A ₁₂₁ ,A ₁₂₂)=3

$$RE(\text{部队编号})=\frac{1}{20} * 2 + \frac{10}{20} * 2.722 + \frac{9}{20} * 2.863$$

$$= 2.794$$

$$RE(\text{编制人数})=\frac{16}{20} * 3.047 + \frac{4}{20} * 3 = 3.038$$

计算伊军编成情报模拟数据中的属性熵值,如表 6:

表 6 伊军编成属性计算表

部队编号=001001001001001001001030000000…001001001001001001001999039999999	A ₁₂₃ =6 A ₁₂₄ =3	I(A ₁₂₃ ,A ₁₂₄)=2.918
部队编号=0010010010010010010020010300000000…0010010010010010029990399999999	A ₁₂₅ =3 A ₁₂₆ =8	I(A ₁₂₅ ,A ₁₂₆)=2.846
编制人数≤500	A ₁₂₇ =7 A ₁₂₈ =9	I(A ₁₂₇ ,A ₁₂₈)=3.044
编制人数≥500	A ₁₂₉ =2 A ₁₃₀ =2	I(A ₁₂₉ ,A ₁₃₀)=3

$$RE(\text{部队编号})=\frac{11}{20} * 2.918 + \frac{9}{20} * 2.846 = 2.886$$

$$RE(\text{编制人数})=\frac{16}{20} * 3.044 + \frac{4}{20} * 3 = 3.035$$

故此部队编号属性为决策树的子节点的属性。所有采集的数据都依次推算,最终得到一棵完整的决策树。

2.3 生成情报威胁评估结果 部分得到的情报威胁评估结果如表 7 所示,表中的 DL 分别取值“High”,“Normal”,“Low”,代表了指定单位实时威胁评估的风险等级。

3 测试结果分析

3.1 分类速度 首先采用改进 sigmoid 函数的 Boltzmann 机对问题进行训练,sigmoid 函数采用公式(1)中的函数,ρ 取 1.5。训练目标采用区间[−1.7,−1.0]

或[1.0,1.7],训练方法采用随机训练法。随后采用改进熵函数的 ID3 算法对经过改进后的 Boltzmann 机训练后的问题进行分类。同时,也采用没有改进的 Boltzmann 机和 ID3 算法对问题进行训练和分类。为了直观地比较两种情况的效果,我们分别做出了它们的训练曲线,实验结果如下:

表 7 第二层得到的部分情报威胁评估结果

Rule1: if (xtk xhp xdd pbwql = = 0000) (部队编号 = { 038001 ... 038100 } (装备编码 = { 20000,29999 } then (PSD=Low Dl=High)
Rule2: if (xtk xhp xdd pbwql = = 0000) (部队编号 = { 038001001 ... 038009001 } (装备编码 = { 20000,29999 } then (PSD=Low Dl=High)
Rule3: if (xtk xhp xdd pbwql = = 0000) (部队编号 = { 038001001 ... 038009001 } (装备编码 = { 40000,49999 } then (PSD=Normal Dl=High)
Rule4: if (xtk xhp xdd pbwql = = 0100) (部队编号 = { 038001 ... 038100 } (装备编码 = { 20000,29999 } then (PSD= Normal Dl=High)
Rule5: if (xtk xhp xdd pbwql = = 0100) (部队编号 = { 038001001 ... 038009001 } (装备编码 = { 40000,49999 } then (PSD= High Dl=High)
Rule6: if (xtk xhp xdd pbwql = = 1100) (部队编号 = { 038001001 ... 038009001 } (装备编码 = { 20000,29999 } then (PSD= High Dl=Low)
Rule7: if (xtk xhp xdd pbwql = = 1100) (部队编号 = { 038001001 ... 038009001 } (装备编码 = { 40000,49999 } then (PSD= High Dl=Low)
Rule8: if (xtk xhp xdd pbwql = = 0100) (部队编号 = { 038001001 ... 038009001 } (装备编码 = { 40000,49999 } then (PSD= Normal Dl=Low)
Rule9: if (xtk xhp xdd pbwql = = 0101) (部队编号 = { 038001001 ... 038009001 } (装备编码 = { 20000,29999 } then (PSD= Normal Dl= Normal)
Rule10: if (xtk xhp xdd pbwql = = 0101) (部队编号 = { 038001001 ... 038009001 } (装备编码 = { 40000,49999 } then (PSD= High Dl= Normal)
Rule11: if (xtk xhp xdd pbwql = = 0110) (部队编号 = { 038001001 ... 038009001 } (装备编码 = { 20000,29999 } then (PSD= Normal Dl= Normal)
Rule12: if (xtk xhp xdd pbwql = = 1101) (部队编号 = { 038001001 ... 038009001 } (装备编码 = { 20000,29999 } then (PSD= High Dl=Low)

实验分别采用了 83-53-13 结构的 Boltzmann 机网络进行训练,随后利用 ID3 算法对训练结果进行分类,例如:图 4 是利用改进后的 Boltzmann 机与未改进 Boltzmann 机进行问题训练曲线图。图 5 是利用改进后的 ID3 算法和未改进的 ID3 算法进行问题分类曲线图。图 6 是采用改进后的 Boltzmann 机和改进后的 ID3 算法和未做改进的 Boltzmann 机和 ID3 算法进行问题训练分类曲线图。如图中所示,在三种情况下,用改进后的 Boltzmann 机和 ID3 算法明显比未做改进的 Boltzmann 机和 ID3 算法训练分类速度快。

3.2 分类精度 精度分析是基于混淆矩阵的评价方法,包括错分误差和漏分误差。错分误差是指被用

户划分为某一类,而实际上是属于另一类的像元;漏分误差是本属于某一类,但是没有被分类器分到相应类别中的数据。

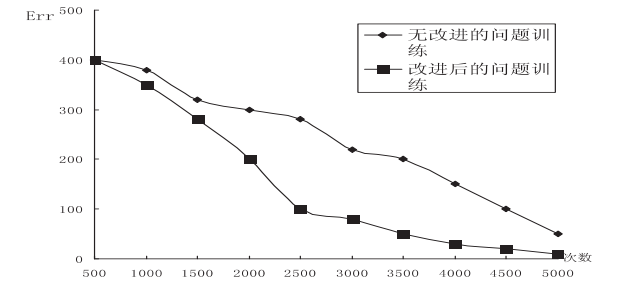


图 4 改进后的 Boltzmann 机与未改进 Boltzmann 机问题训练曲线图

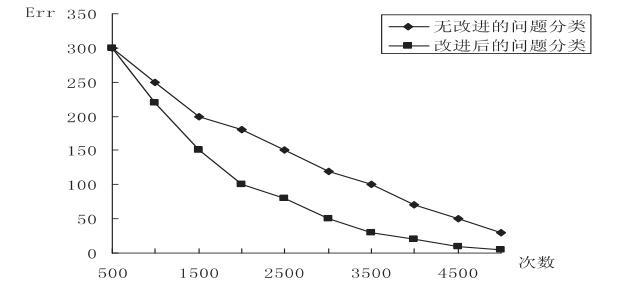


图 5 改进后的 ID3 算法与未改进 ID3 算法问题分类曲线图

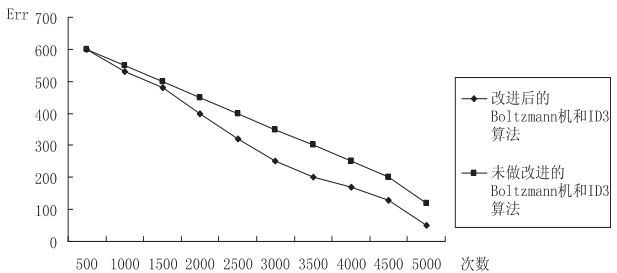


图 6 改进后的 Boltzmann 机和 ID3 算法和未做改进的 Boltzmann 机和 ID3 算法问题训练分类曲线图

根据精度分析的概念,对陆战当前状态情报数据中的 9000 条数据进行威胁评估分类。为更好地对改进前后的算法分类精度进行对比,事先已多次对该数据进行人工情报研判,并得出相应的威胁评估结果,其中,符合低风险等级(DL=LOW)的情报目标单位为 2 589 个,符合中等风险等级(DL=NORMAL)的情报目标单位为 4 578 个,符合高风险等级(DL=HIGH)的情报目标单位为 1 833 个。从分类结果上可以看出,改进后的 Boltzmann 机和 ID3 算法的分类提取的结果最接近实际情况,如表 8:

表 8 改进后的 Boltzmann 机和 ID3 算法和未做改进的 Boltzmann 机和 ID3 算法分类精度分析

类别	Boltzmann 机			ID3 算法			改进后的 Boltzmann 机和 ID3 算法		
	LOW	NORMAL	HIGH	LOW	NORMAL	HIGH	LOW	NORMAL	HIGH
LOW(2589)	1803	578	208	1764	421	404	2098	278	213
NORMAL(4578)	430	3409	739	386	3098	1094	287	3976	315
HIGH(1833)	129	399	1305	153	409	1271	63	216	1554
总体精度(%)		71		68		83			

(上接第 162 页)

4 结 论

本文利用两种情报数据分类算法构建出一种新的情报数据分类算法用于威胁评估模型的开发设计之中,针对原有算法的不足做出改进,并借助情报模拟数据对整个系统进行了实例分析与验证。从分析结果中发现,新的情报数据分类算法在数据训练、分类的速度上要优于未做改进的算法。

参 考 文 献

[1] 欧爱辉,朱自谦. 基于多属性决策和态势估计结果的空战威胁评估方法[J]. 火控雷达技术,2006,35(2):64-67

[2] 王 峰,潘 泉,高泉学. 一种基于神经网络的目标优先级确定方法[J]. 电光与控制,2003,10(4):38-41

[3] 王 猛,章新华,夏志军. 基于属性分析的威胁评估技术研究[J]. 系统工程与电子技术,2005,27(5):848-851

[4] 曹可劲,江 汉,赵宗贵. 一种基于变权理论的空中目标威胁估计方法[J]. 解放军理工大学学报,2006,7(1):32-35

[5] 宋国春,刘 忠,黄金才. AHP 方法的敌地域通信网通信链路威胁评估[J]. 火力与指挥控制,2008,33(2):16-20

[6] 黄现江,余思明. 基于修正核函数的支持向量机空袭目标威胁评估[J]. 指挥控制与仿真,2009,31(6):33-35

[7] 贾世楼. 信息理论基础[M]. 哈尔滨:哈尔滨工业大学出版社,1986:12

[8] D E Culler, R Karp, D Patterson, A Sahay K. E. Schauser, E. Santos, R. Subramonian T. Voneicken LogP: Towards a Realistic Model of Parallel Computation[A]. Proc. ACM Symp. on Principles and Practice of Parallel Programming,1993:1-12

[9] 王 旅,彭 宏,胡勤松. 基于判定树归纳分类的土质分类定名方[J]. 计算机工程与设计,2006,27(11):1929-1931

[10] 赵玉贵. 多层前向网络泛化能力的研究与应用[D]. 郑州:解放军信息工程大学,2005

(责编:白燕琼)