【其他研究】

基于 Logistic 回归的识别算法适应性研究

黄彬彬[1],赵久奋[1],彭会釜[1],曹　锐[2]

(1. 第二炮兵工程学院,西安　710025; 2. 中国人民解放军 96167 部队, 福建 永安　366000)

摘要:对自动目标识别算法的适应性进行了研究。通过仿真实验观察识别算法在 2 种影响因素(俯仰角与信噪比)的不同取值下的性能差异,建立了 Logistic 回归模型来验证其准确性,从而实现对算法的适应性研究。实验证明,该方法在识别算法的适应性研究上具有一定的可行性,可以为目标识别算法的选择提供理论依据。

关键词:自动目标识别;适应性;Logistic 回归

中图分类号:TN957　　**文献标识码**:A　　**文章编号**:1006 - 0707(2011)10 - 0131 - 03

精确制导武器在当今的信息化战争中显得越来越关键,因而目标识别在精确制导中的地位就显得不可或缺。作为武器设计部门,选择一种目标识别算法应该要考虑到算法的实际战场应用情况,找到一种适应当时当地战场环境的算法,以提高作战效率,减少不必要的损失。一种目标识别算法的优劣,应该从它的实际应用情况来分析,有些算法能适应的场景变化范围宽一些,适应性好;有些算法则反之,对某些图像处理得很好,但情况一变,性能就下降得很快,适应性差。因而在复杂多变的应用场景中,其实际性能往往无法预测[1]。因此,多场景的自适应问题已成为 ATR 系统发展中的一个最关键问题,也是 ATR 技术中最为严重和最具挑战性的问题。ATR 的适应性不仅仅是一个期望的特性,它更是一个非常关键的功能要求[2]。事实上,ATR 的适应性问题早已为人们所认识,并将其理解为在任何情况下,ATR 系统都能良好运行,而且能取得与训练条件下相近的目标检测和识别性能[3]。这意味着识别算法必须具备足够的灵活性,能适应很大范围的场景条件,整个 ATR 系统必须具有非常广泛的适应性[4]。因此,研究 ATR 算法在各种影响因素下的不同性能就显得十分必要。

1　识别算法适应性研究

1.1　算法适应性研究现状及原理

识别算法的性能评估主要是建立在统计学基础上的,其主要方法有假设检验、回归分析、区间估计、ROC 曲线等方法,而针对已有实验数据的算法适应性,通常采取回归分析方法来进行研究,主要有线性回归、非线性回归、Logistic 回归、最小二乘回归、序回归、加权估计、概率单位回归等。

ATR 算法对某类目标进行识别,识别结果是一种二值离散变量:“正确识别”和“未正确识别”。一般情况下,用变量 $Y_i(i=1,2,\cdots,n)$ 表示第 i 次的识别结果,并规定 Y_i 的取值为

$$Y_i=\begin{cases}1, & \text{正确识别}\\ 0, & \text{未正确识别}\end{cases}$$

Y_i 是一种二分类变量或称为二值离散变量。在分析分类变量时,通常采用对数线性模型。当对数线性模型中的某个二分类变量被作为应变量并被定义为一系列自变量的函数时,对数线性模型就变成了 Logistic 回归模型。

Logistic 回归模型适合分析 ATR 算法的识别结果这种二值随机变量和影响因素 X 之间的相互作用关系,因此可用 Logistic 模型预测所评估的 ATR 算法在影响因素取值不同情况下的目标识别概率。相同的 ATR 算法在影响因素作用下对同一类目标也会有着不同的识别结果,对不同算法的识别结果进行 Logistic 回归将得到不同的 Logistic 模型及模型参数。模型间的差异反映出不同 ATR 算法的性能差别。利用 Logistic 回归分析得到 ATR 算法模型,结合模型所代表的实际意义,就可以实现对 ATR 算法的性能评估[2]。

1.2　Logistic 回归模型的方法原理

首先假定在相同条件下,识别算法对同类型的目标有同样的识别概率。在此前提下,每进行一次识别,就相当于进行了一次 Bernoulli 试验。设 p_i 为第 i 次识别所处的测试条件下 $Y_i=1$ 的概率(即 $P\{Y_i=1\}=p_i$)。显然,p_i 就是算法在此条件下对这类目标的识别率。Y_i 实质上是一个服从二项分布的随机变量,其数学期望就等于 p_i[2]

$$E(Y_i)=p_i \tag{1}$$

设 X 为自变量,Y 为应变量,第 i 次识别测试时影响因素 $X=X_i$,则对应的应变量 $Y_i=1$ 发生的条件概率 p_i 可表示为 $P\{Y_i=1\mid X_i\}=p_i$。在只考虑一个自变量的情况下,Logistic 回归模型的基本形式为

收稿日期:2011 - 08 - 11

作者简介:黄彬彬(1985—),男,硕士研究生,主要从事精确制导理论与技术研究。

$$p_i = \frac{e^{\alpha+\beta X_i}}{1+e^{\alpha+\beta X_i}} \tag{2}$$

式中，α 和 β 分别为回归截距和回归系数。对式(2)中的 Logistic 回归模型进行推导后可以得到

$$\ln\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta X_i \tag{3}$$

式中 $\ln(p_i/(1-p_i))$ 被称为对数发生比。对具体的一组观测值$(X_i, Y_i)(i=1,2,\cdots,n)$一般采用最大似然估计法进行 Logistic 回归来获得参数 α 和 β，一些专业统计软件如 SAS、SPSS、EXCEL 等都提供这种 Logistic 回归分析功能[5-6]。

可以看出，Logistic 回归模型假设了对数发生比与自变量之间为线性关系。如果式(3)中所假定的线性关系得不到满足，那么采用最大似然估计法所进行的参数估计将发生偏差[2]。

由式(1)和式(2)可得

$$p_i = \frac{e^{(\alpha+\beta X_i)}}{1+e^{(\alpha+\beta X_i)}} \tag{4}$$

对一组观测值(X_i, Y_i)进行 Logistic 回归得到模型参数 α 和 β 后，就能够描述出应变量 Y 的数学期望 $E(Y)$ 和自变量 X 之间的关系

$$E(Y) = \frac{e^{(\alpha+\beta X)}}{1+e^{(\alpha+\beta X)}} \tag{5}$$

式(5)通常被称为 Logistic 函数。

对一种目标识别算法作评估时，决策者真正关心的是不同影响因素值情况下的算法识别率，Logistic 函数恰恰给出了识别率 $E(Y)$ 随影响因素 X 而变化的这种函数关系，因此可以用 Logistic 函数预测所评估的算法在不同影响因素值情况下的识别率，从而判断该算法的适应性[3]。当然，对不同的识别算法进行 Logistic 回归分析也将得到不同的 Logistic 函数，它们之间的差异也能反映出不同算法之间识别性能的不同[6]。

2 实验过程及结果分析

2.1 实验过程

根据实验设计，本文选用一种识别算法进行测试，测试数据集由 800 个目标仿真数据组成，其中俯仰角取值范围为 60°~90°，信噪比取值为 0~25 dB，实验目标为观察检验该算法分别受俯仰角及信噪比的影响程度。

首先，用算法分别对 800 个数据进行识别，记录每次测试识别结果，如表 1 所示。

表 1 测试记录结果

	俯仰角 pitchangle /(°)				信噪比 *SNR* /dB			
取值	90	75	70	60	10	15	20	25
识别率	0.99	0.95	0.87	0.76	0.99	0.98	0.95	0.96

其次，根据表 1 所示，分别计算算法在两种因素影响下的对数发生比，由此得到对数发生比与信噪比、俯仰角的关系分别如图 1、图 2 所示。由图 1、图 2 可以看出，该算法对数发生比与俯仰角、信噪比之间基本均为线性关系，满足进行 Logistic 回归分析的基本条件。

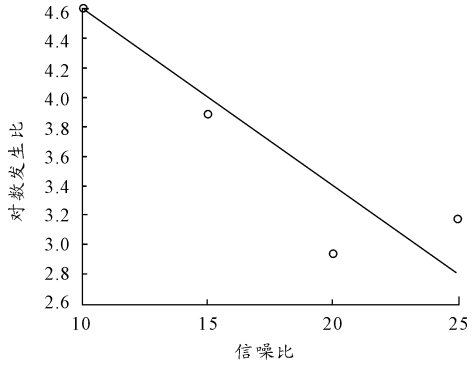

图 1 对数发生比与信噪比的关系

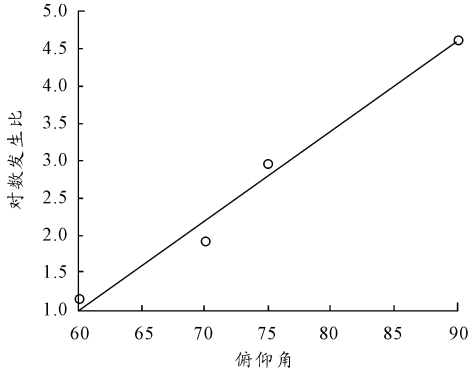

图 2 对数发生比与俯仰角的关系

最后，根据第一步的记录结果进行 Logistic 回归分析，计算后可得出相应的回归系数，卡方值及显著性水平等，可用 SPSS 统计软件进行建模[4]，结果分别如表 2、表 3。

表 2 信噪比建模结果

方程中的变量	*B*	*S.E*,	*Wals*	d*f*	*Sig.*	Exp (*B*)
SNR	−0.056	0.120	0.217	1	0.642	0.946
常量	4.281	2.461	3.027	1	0.082	72.317

表 3 俯仰角建模结果

方程中的变量	*B*	*S. E*,	*Wals*	d*f*	*Sig.*	Exp (*B*)
pitchangle	0.109	0.053	4.274	1	0.039	1.115
常量	-5.479	3.507	2.442	1	0.118	0.004

由表 2、表 3 可知,该算法的 Logistic 回归函数为:

$$p_d = \frac{e^{4.281-0.056X_1}}{1+e^{4.281-0.056X_1}} \tag{6}$$

$$p_d = \frac{e^{-5.479-0.109X_2}}{1+e^{-5.479-0.109X_2}} \tag{7}$$

其中,X_1、X_2 分别代表信噪比与俯仰角两类影响因素。

2.2 实验结果分析

表 2、表 3 仅仅列出结果中对于建立 Logistic 回归模型有用的一部分。其中:*B* 项是回归方程的回归系数,它分别反映影响因素 X_1、X_2 对识别率的影响程度;*S. E.* 项是标准误差,它不是测量值的实际误差,也不是误差范围,它只是对一组测量数据可靠性的估计。标准误差小,测量的可靠性大一些,反之,测量就不大可靠;*Wals* 项是卡方值,它是卡方检验的统计量,主要用来检验数据的相关性,卡方值越大说明实际频数与理论频数的差别越明显;*df* 项是卡方检验的自由度;*Sig.* 项是卡方值的显著性,主要用来说明两个变量之间的相关性;Exp(*B*)是相对危险度 *OR* 值(即 odd ratio),也称优势比、比值比。若 *OR* 值大于 1,表示该因素是一个危险因素,反之则为保护因素,若 *OR* 值等于 1,表示该因素不起作用。

由于本文主要研究识别算法的适应性,因此主要考虑识别率与两种影响因素之间的相关性。由表 2 可以看出,*SNR* 的显著性指标 *Sig.* 项的值是 0.642 > 0.05(0.05 为设置的显著性水平),说明信噪比对该算法的识别结果影响不大,而表 3 中 pitchangle 的显著性指标 *Sig.* 项的值是 0.039 < 0.05,说明俯仰角对该算法的识别结果有较大的影响。比较表 1 所记录的结果也可以看出俯仰角对算法的影响显然比信噪比的影响要大得多,因此可以证明该方法对于评估识别算法的适应性具有一定的可行性。

3 结束语

本文在对目标识别算法研究的基础上,针对识别算法在俯仰角和信噪比这两种影响因素的不同取值下的性能进行了适应性实验,并根据实验结果建立了这两类影响因素和识别率之间的 Logistic 回归模型。通过该模型,验证了识别算法在这两类影响因素的不同测试水平的识别性能,从而评估了识别算法的适应性,也证明了 Logistic 回归分析方法在识别算法适应性研究上的可行性和实用性。

参考文献:

[1] 衡燕,郭桂蓉. 广义操作条件下的 ATR 算法性能评估[D]. 长沙:国防科学技术大学,2006.

[2] 何峻,卢再奇,付强. 一种基于 Logistic 回归模型的 ATR 算法性能评估方法[J]. 雷达科学与技术,2005(3):152-155.

[3] Yanxing SONG, Feng YUAN. The Research of A New Auto Target Recognition Directed Image Compression[C]// 3rd International Congress on Image and Signal Processing. [S. l.]:[s. n.],2010:224-228.

[4] Scott K. Ralph, John Irvine, Magnús Snorrason, et al. An Image Metric-Based ATR Performance Prediction Testbe [J]. IEEE Trans. on Computer Scoiety,2006.

[5] 王万东,陈燕平,钟晟. 应用 CF 和 Logistic 回归模型编制滑坡危险性区划图[J]. 中南大学学报,2009,40(4):2231-2236.

[6] 陈永胜. 基于 MATLAB 和 SPSS 的非线性回归分析[J]. 牡丹江大学学报,2009(5):372-376.

(责任编辑 周江川)