融合 3σ 法则和假设检验的图书称重复核①

沈洪骥

(山东新华书店集团有限公司，济南 250000)

摘　要：针对图书物流中复核效率低下的情况提出一种称重复核方法。该方法利用重量作为度量，将图书复核看作分类问题。首先基于各类样本图书的均值和方差用 3s 法则和假设检验方法分别构造初始分类器，然后利用 Adaboost 算法从初始分类器出发构造最终分类器，最后用最终分类器对图书物流箱进行称重复核。该方法克服了初始分类器无法全面准确分类的缺点，具有较高的准确率，同时提高了图书复核的效率。

关键词：称重复核；3s 法则；假设检验；Adaboost 算法

Fusing 3σ Theory and Hypothesis Testing for Weight Double-check of Books

SHEN Hong-Ji

(Xinhua Bookstore of Shandong, Jinan 250002, China)

Abstract: Due to the low efficiency of book double-check in current book logistics, we propose a method of weight double-check. Using weight as a quantitative representation, this method considers book double-check as a classification problem. First, 3s principle and hypothesis testing are adopted to build the initial classifiers, based on the mean and variance of the weight of different classes. Then we construct the final classifier from the initial classifiers through Adaboost algorithm. This method overcomes the deficiency of initial classifiers, and improve the efficiency of current double-check approach.

Key words: weight double-check; 3s theory; hypothesis testing; Adaboost algorithm

1　前言

在现代图书物流系统中，根据客户订单要求将图书手工分装到物流箱中，然后对原始分装情况进行检查，筛选出分装时出现类别或数量错误的图书箱并纠正错误称为图书复核。人工开箱复核的方式效率极低，严重降低了整个图书物流系统的运转速度，使得图书复核成为制约整个系统效率提升的瓶颈环节，因此急需一种有效的手段进行快速图书复核。

在物流系统的正常运转下，原始的手工分装已能保持较高正确率，因此图书复核问题可利用称重的思路来解决，即将样本重量 i 与标准重量作比较，如果误差落在某一阈值之内则认为样本不存在错误，该阈值通常根据样本统计信息来计算并在应用过程中不断修正。称重复核作为一种科学、高效的复核方式，已经广泛应用于邮政[1]、航空货运[2]、医药等行业，但在图书物流系统中尚未推广应用。目前在发达国家，称重复核技术在图书物流领域已有较为成功的案例，例如：日本东京图书配送中心能够利用图书称重复核技术完成全东京每天数千种杂志配送到 25000 个客户的任务。在国内图书称重复核技术尚未得到广泛的应用，只有上海新华传媒物流中心[3]等极个别企业进行过尝试，效果未知。

目前的图书称重复核技术仅仅使用单一统计方法进行判定，导致准确率难以达到理想水平。在此情况下，我们将图书称重复核看作一个基于统计的二值分类问题，根据分类结果判断某图书物流箱的图书是否存在分装错误。该方法步骤为：1) 根据任意类别图书重量近似服从正态分布的特性计算样本图书重量的分

① 收稿时间:2011-05-10;收到修改稿时间:2011-06-13

布概率，采用 3s 法则[4]构造初始分类器；2) 基于样本图书重量分布采用假设检验方法[5]构造初始分类器；3) 利用 Adaboost 算法[6]将初始分类器有效整合为最终分类器进行称重复核。3s 法则能以高准确率识别错误的样本，而假设检验方法则能以高准确率识别正确样本。两种方法分别从正反两个角度进行识别，无法提供全面准确的结果，而 Adaboost 算法能够将较弱分类器整合为更强分类器，在人脸检测等很多应用中表现出异常优异的性能[6]，因此本文算法能够有效提高称重复核的准确率。

2 融合3σ法则和假设检验的方法

图书称重复核问题的目的是根据任意一个包含多种类别图书的图书物流箱的重量判断该物流箱内的图书是否存在填充上的错误，如误将其他种类的书放入，或者误将某种类错误数量的书放入等等。因此可将该问题归结为一个分类问题，简单地说就是判断某图书物流箱内的图书是否有问题。

3σ 法则和假设检验方法已在计量科学领域内获得广泛应用，但两种方法是从正反两个角度出发进行分类。具体说来，3σ法则目的是判断有明显错误的物流箱，且只能检测出和标准值差异较大的个体，如果差异较小则较难检测出内容分装的错误，而假设检验方法能以较大概率判断出正确的物流箱。可见，单独采用任何一种方法均无法全面准确的进行称重复核。本文给出一种新方法，利用 Adaboost 算法将 3σ 法则和假设检验方法融合在一起进行图书称重复核。

2.1 基于 3σ 法则的图书称重复核

设 x 为包含多个类别图书的图书物流箱重量，且 x 符合均值为 μ、方差为 σ 的正态分布，其中 μ、σ 可通过各类样本图书重量的均值 μ_i、方差 σ_i(i 表示类别)，则 3σ 法则的内容是：当一个图书箱的重量超出区间 $\langle \mu-3\sigma, \mu+3\sigma \rangle$ 时，该图书箱出错的概率为 99.74%。

该法则可以推导如下：

$$\begin{aligned} &P(|\mathrm{x}-\mu|<3\sigma) \\ &=P(\mu-3\sigma<\mathrm{x}<\mu+3\sigma) \\ &=\Phi\left(\frac{\mu+3\sigma-\mu}{\sigma}\right)-\Phi\left(\frac{\mu-3\sigma-\mu}{\sigma}\right) \\ &=\Phi(3)-\Phi(-3) \\ &=2\Phi(3)-1=2\times 0.9987-1=0.9974 \end{aligned} \tag{1}$$

其中 Φ 为标准正态分布。上式说明，某次随机事件中 x 落入区间 $\langle \mu-3\sigma, \mu+3\sigma \rangle$ 的概率为 0.9974，超出此区间的可能性极小，这恰好说明当重量超出此区间时，出错的概率为 99.74%。3σ 法则可作为一种基本算法来判断有明显错误的物流箱，但它只能检测出和样本差异较大的个体错误，而在 $\langle \mu-3\sigma, \mu+3\sigma \rangle$ 区间内，既有正确样本（99.74%），也有 3σ 定理没有识别出的具有很小样本差异的错误样本。

2.2 通过假设检验（显著性差异）进行图书称重复核

与海量数据挖掘相比，图书称重复核属于小样本问题，因此根据假设检验理论，可假设显著性水平 α 为 0.05。设 μ 为某类图书箱应有的标准平均重量(根据其中各类图书重量的均值 μ_i 计算)，$\bar{x}$ 为某次测试中样本的平均重量，我们假设 $\bar{x}$ 与 μ 的偏差为 d 时，显著性水平达到 0.05，则有：

$$P(|\bar{x}-\mu|\geq d)=\alpha \tag{2}$$

设任意一箱图书重量 x 符合正态分布 $x\propto N(\mu,\sigma^2)$，n 为此次测试中的样本数量，则样本均值 $\bar{x}$ 符合均值为 μ，方差为 $\frac{\sigma}{n}$ 的正态分布，即 $\bar{x}\propto N(\mu,\sigma^2/n)$。

令 $y=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$，有：

$$P(|y|\geq \frac{d}{\sigma/\sqrt{n}})=\alpha \tag{3}$$

由于 $y\propto N(0,1)$，故有 $\frac{d}{\sigma/n}=y_{\frac{\alpha}{2}}$。统计量 y 在假设检验中称为检验统计量，把 $y_{\frac{\alpha}{2}}$ 称为临界值，由于 α 为 0.05，查表可得 $y_{\frac{\alpha}{2}}=1.96$，故有 $d=1.96\frac{\sigma}{\sqrt{n}}$，即置信区间为 $\left\langle \mu-1.96\frac{\sigma}{\sqrt{n}},\mu+1.96\frac{\sigma}{\sqrt{n}} \right\rangle$，也就是说当 α 为 0.05 时，95%正确的图书箱符合通过置信区间，能够顺利通过称重复核检验。可以通过调整 α 值获得不同的识别准确率。

2.3 利用 Adaboost 方法进行算法融合

3σ 法则和假设检验分别可以识别出 99.74%的错误和 95%的正确样本，但两种方法均无法进行准确的全面识别，因此本文采用 Adaboost 方法将两种方法的识别结果进行融合。Adaboost 在模式识别领域中已被广泛应用，但尚未用来进行图书称重复核，其基本思路是利用一些较弱的分类器加权组合生成一个强分类器，权重通过衡量分类器的性能得到。Adaboost 方法通常是个迭代的过程，在迭代的每一步考察每个样本

的分类情况，在下一步对那些分错的样本给予更多的注意力，从而实现分类器的增强[6]。

我们利用一组已知的图书箱作为样本来训练分类器。设 $S=\{\langle x_1,y_1\rangle,...,\langle x_n,y_n\rangle\}$ 为训练，其中 x_i 为样本数据(图书箱)，y_i 为样本的已知分装情况(y_i=1 为正确，y_i=-1 为错误)，$H=\{h_1,h_2\}$ 为初始阶段的弱分类器集合，其中 h_1 是采用 3σ 法则的分类器，h_2 是基于假设检验的分类器，$h_j(x)$为针对 x 的分类结果。我们的目的是训练 T 个弱分类器 h_t，并用这 T 个若分类器构造强分类器 h。利用函数 g()表示分类结果正确与否：

$$g(x_i)=\begin{cases}1 & y_i\neq h_j(x_i)\\ 0 & y_i=h_j(x_i)\end{cases}$$

其中 $h_j\in H$。用 w^j 表示为每个样本分配的权重，用 β_t 表示每个分类器的权重，ε 表示分类误差。在算法初始化时 $w^j=\frac{1}{n}$，本文的 Adaboost 方法过程如下：

$$\begin{aligned}&for\ \ t=1,...,T\\&\quad for\ \ h_j\in H\\&\qquad \varepsilon_j=\sum_{i=1}^{n}w_t^j\cdot g(x_i);\\&\quad end\\&\quad h_t=\arg\min_{h_j\in H}\varepsilon_j;\\&\quad \beta_t=\frac{1}{2}\ln\frac{1-\varepsilon_t}{\varepsilon_t};\\&\quad w_{t+1}^j=\begin{cases}w_t^j\exp(-\beta_t)/Z_t & y_i=h_t(x_i)\\ w_t^j\exp(\beta_t)/Z_t & y_i\neq h_t(x_i)\end{cases};\\&end\end{aligned}$$

其中 $Z_t=\sum_{i=1}^{n}w_t^j$。迭代结束后得到的最终分类器为：$h(x)=sign(\sum_{i=1}^{T}\beta_t h_t(x))$，当取值为 1 时，该箱识别为正确，当取值为-1 时识别为错误。

3　实验

为检验本文算法设计了 3 组实验。首先，本文算法依赖于各类样本图书重量的均值 μ_i 与方差 σ_i，因此需要测试算法对样本数量的依赖性。其次，需要测试待称重图书箱中图书的册数和类别数对本文算法的影响。最后将本文提出的方法与人工复核方法进行对比。

从各类图书中分别取 5、10、20、50、100 册计算该类图书的均值与方差，并基于此用本文提出的方法进行称重复核，识别准确率如图 1 所示。图 1 表明各类样本册数越多，计算出的 σ_i 和 μ_i 就越有效，用在称重复核算法中准确率越高，同时将 3σ 方法和假设检验方法组合在一起的能进一步提高准确率，因此在余下试验中，我们取各类样本 100 册计算均值与方差，并用组合算法代表本文算法。

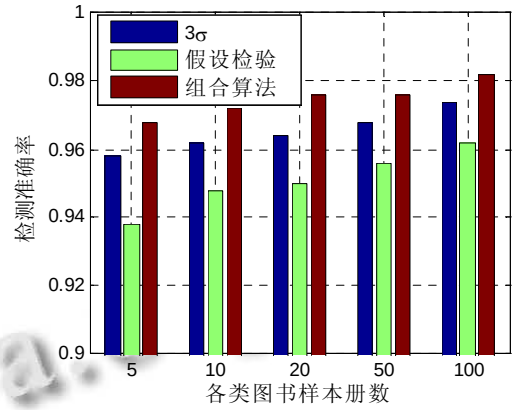

图 1　检测准确率对各类图书样本册数的依赖性

为检验图书箱中图书的数量对本文称重复核算法的影响，取包含册数分别为 10、20、30、50、100 册的 500 箱图书进行 5 组实验，结果如表 1 所示。这说明称重复核算法的准确性与箱内包含图书的册数无直接联系。

表 1　包含不同数量图书的图书箱称重复核准确率

	10 册	20 册	30 册	50 册	100 册
本文方法	97.2%	96.6%	96.6%	97.4%	97.0%

称重复核算法面临的更加复杂和实际的情况是品种的随机变化。通过对本单位最近一年度物流数据的统计得知，所有发运图书箱所包含品种数最小为 1，最大为 103，平均品种数为 10.8。根据历史数据，我们取箱内图书品种分别为 1、5、10、20、50 种的 500 箱图书进行 5 组实验，每箱图书册数为 50 册，结果如表 2 所示。表 2 表明一个图书箱内图书品种的增加，会导致本文称重复核算法准确率的降低。

表 2　包含不同品种图书的图书箱称重复核准确率

	单品种	5 品种	10 品种	20 品种	50 品种
本文方法	97.6%	96.8%	95.6%	93.2%	90.8%

图 2 展示了本文方法与人工复核方法的准确率和时间效率的比较情况。实验的全过程流水线速度控制在 40 箱/分钟，也就是说称重复核的速度也是 40 箱/分钟，人工复核的平均速度为 1 箱/分钟，而整条流水线设计了的 16 个包装口，如果将 16 个包

装口全部配备人员　采用人工复核的方式，按照上述人工复核的平均速度，整条流水线人工复核的速度也仅仅为 16 箱/分钟，高速度的称重复核提升了整条流水线的速度，提升幅度达到至少 275%。同时，称重复核算法能够保证 95%以上的准确性，而由于人生理因素的影响，人工复核的方法复核的准确率，也仅仅达到 97%，并没有明显的提高。综合复核准确率和复核效率考虑，本文算法在实际应用中能取得更好效果。

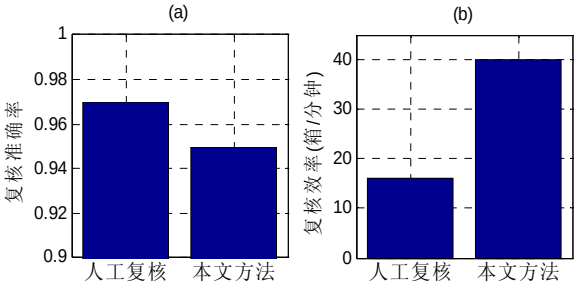

图 2　本文方法与人工复核方法的比较

4　结论

本文提出一种图书称重复核算法，根据样本图书类重量的均值和方差用 3σ 法则和假设检验方法分别构造初始分类器，然后利用 Adaboost 算法基于初始分类器构造最终分类器，用最终分类器对实现图书箱的称重复核。该方法克服了初始分类器无法全面准确分类的缺点，与人工复核相比大大提高了复核效率，同时保持了很高的称重复核准确率。在今后的工作中，拟将不同种类的纸张可能导致的误差、季节、湿度等环境因素可能导致的误差纳入概率统计范围，进一步提高称重复核准确率。

参考文献

1 徐永强.邮用秤的流转管理.中国计量,2001,64(3):1–3.
2 冯建忠,张仁颐.航空货运重量复核系统的改造.仪器仪表用户,2006,13(1):54–56.
3 侯凌燕,尹军琪.图书重量复核技术的应用.物流技术与应用,2008,13(1):92–93.
4 何晓群.关于 6 Sigma 与 3 Sigma 的比较.数理统计与管理,2006,25(2):175–177.
5 Lehmann EL, Joseph PR. Testing Statistical Hypotheses. New York: Springer, 2005 ISBN.0387988645.
6 熊盛武,宗欣露,朱国锋.改进的基于 Adaboost 算法的人脸检测方法.计算机应用研究,2007,24(11):298–300.

(上接第 52 页)

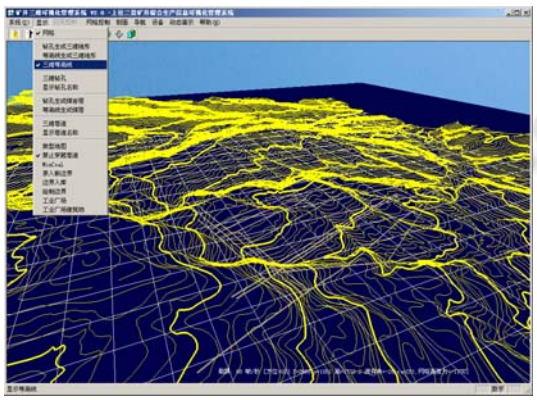

图 9　井田三维地形等高线

5　结论

山西阳泉上社二景矿地质信息可视化管理系统采用先进的计算机网络技术、数据库技术、计算机图形处理技术，采用支持面向对象技术的 Microsoft Visual VC++优秀软件开发工具，基于客户机/服务器应用架构开发的矿山三维可视化管理平台，通过集成矿井地质信息三维可视化管理系统，实现了矿井地质信息可视化管理，具有强大的二维和三维图形处理功能，实时、高效的地质信息获取、查询，为矿井的安全、经济开采提供了科学、高效的现代化地质信息三围可视技术，提高了企业的信息化水平和管理水平。测试和应用结果表明，系统结构合理，功能完善，性能卓越，运行高效，交互界面良好，操作简捷，能够完全满足上山西阳泉上社二景矿的实际生产，具有很高的实用性和可靠性。

参考文献

1 孙波.OpneGL 编程实例学习教程.北京:北京大学出版社,2000.42–46.
2 李培军.层状地质体的三维模拟和可视化.地学前缘,2000,7(8):32–36.
3 邓寅生,曲鹏举,庞玉娟.基于 OpenGL 的地质体三维可视化系统开发.微计算机信息,2007,(3):18–20.