Two-stage Approach for Unbalanced Classification with Time-varying Decision Boundary: Application to Marine Container Inspection

Richard Hoshino
Laboratory and Scientific Services Directorate, Canada Border Services Agency, 79 Bentley Avenue, Ottawa, ON, Canada K2E 6T7
richard.hoshino@gmail.com

R. Wayne Oldford
Centre for Computational Mathematics in Industry and Commerce, University of Waterloo, 200 University Avenue West, Waterloo, ON, Canada N2L 3G1
rwoldford@uwaterloo.ca

Mu Zhu*
Department of Statistics and Actuarial Science, University of Waterloo, 200 University Avenue West, Waterloo, ON, Canada N2L 3G1
m3zhu@uwaterloo.ca

ABSTRACT

Two million marine containers arrive each year at Canadian ports, representing a significant percentage of Canada's trade with its overseas partners. While the majority of these commercial shipments are perfectly legitimate, some marine containers are used by criminals to smuggle drugs and weapons. To address this risk, the Canada Border Services Agency (CBSA) employs a predictive model to identify high-risk containers. Recent data-mining initiatives at CBSA led us to study unbalanced classification problems in which the optimal decision boundary may change over time. In this paper, we propose a simple, two-stage approach to deal with such problems. While we focus on the marine container problem at CBSA, our proposed two-stage approach is general.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*; I.2.6 [**Artificial Intelligence**]: Learning; I.5.1 [**Pattern Recognition**]: Models—*Statistical*

General Terms

Algorithms, Experimentation, Performance

1. INTRODUCTION

Since its inception in December 2003, the Canada Border Services Agency (CBSA) has gradually moved towards a "smart border," where day-to-day operational decisions are increasingly guided by science and technology rather than the "gut feelings" of customs officers. Over the past few years, data mining has been used to improve the Agency's security efforts, an approach that is constantly evolving as the Agency receives new data and develops more sophisticated techniques to assess and manage risk.

Over 90 percent of all world trade is transported in marine cargo containers moving from port to port. Approximately two million containers arrive at Canadian ports each year, and represent a significant portion of the national economy. While the grand majority of commercial shipments are perfectly legitimate, some of these marine containers are used by criminals to transport drugs and weapons into Canada.

By Canadian law, shippers are required to send CBSA a customs document, known as a cargo manifest, which contains important transactional information. Among other data elements, the cargo manifest includes the importer name, vendor name, container weight, port of loading, and a description of what is inside the container.

Several years ago, CBSA developed an automated system so that shippers could send the cargo manifest electronically. This information is processed and risk-assessed, based on several dozen variables that predict risk. These indicators are used to assign a risk score, indicating the likelihood that the container holds contraband or other undesirable goods.

1.1 Data, History, and Motivation

About two years ago, a research team at CBSA procured 24 months of manifest data on marine containers that had been fully inspected by CBSA officers, and began to apply data-mining techniques in order to develop a better algorithm based on statistical evidence rather than ineffective profiling.

The data were obtained in two separate batches. The first batch (used as the training set) consists of 15,279 containers from the first few months, and the second batch (used as the test set) consists of 6,259 containers from the remaining months. There are 83 predictors. The exact nature of these predictors is classified information. Each observation is labeled either "clean" or "dirty," where "dirty" means CBSA officers found drugs, weapons, or other items either in direct violation of the Canada Customs Act or otherwise deemed to have posed a threat to the health, safety and security

*Corresponding author.

ISI-KDD 2010, 25 July 2010, Washington, D.C., USA

of the Canadian population. Only a small fraction of the containers — about 2% — were labeled "dirty."

The CBSA research team tested multiple predictive models, such as decision trees and neural networks. In 2008-2009, CBSA started to collaborate with the University of Waterloo (UW) on this project. To start, a team of UW students under the supervision of their professors tested a large number of learning algorithms. The UW team also found that the first principal component direction of the training set was dominated by 18 predictors, and that a simple logistic regression model fitted using just those 18 predictors (simply "Logit18" below) performed as well as other cutting-edge algorithms such as the support vector machine (SVM) [e.g. 4, 8] and random forest [2].

However, while exploring the data, the Waterloo team noticed the probability that a container is "dirty" seemed to fluctuate from time to time. This is related to the idea of concept drift [e.g., 9], and led us to study an interesting unbalanced classification problem where the optimal decision boundary is changing over time, which is what the current paper is about.

1.2 Problem

Mathematically, the problem can be described as follows: Given an observation $(\mathbf{x}, t)$, where $\mathbf{x} \in \mathbb{R}^d$ is a vector of predictors and $t \in \mathbb{R}$ is a time variable, we would like to determine the likelihood that it belongs to a rare class.

Two features distinguish our problem from regular classification problems:

F1. The problem is highly unbalanced. The class of interest is extremely rare; the majority of observations belong to the background class.

F2. There is an extra time variable, t.

Many solutions to the class-imbalance problem (F1) have been proposed, at both data and algorithmic levels. At the data level, these solutions include random oversampling with replacement and random undersampling. At the algorithmic level, solutions include adjusting the costs of the various classes to counter class imbalances and modifying the decision threshold [3].

As noted by many [e.g., 5, 7], that the optimal classification boundary may change over time (F2) is a particularly challenging problem and one that has not yet received much attention in the literature. Part of the difficulty, we believe, is that we don't have enough data to pinpoint how things change over time; we shall be more specific about this difficulty below (Section 2). For highly unbalanced problems, this difficulty is multiplied further, because information about the class of interest is especially limited.

2. METHODOLOGY

Let $y \in \{0, 1\}$ be the class label, where $y = 0$ represents the background class and $y = 1$ represents the rare class. The most natural way for dealing with F2 is to treat the time variable simply as another predictor and model the full conditional distribution, $p(y|\mathbf{x}, t)$, directly. Due to F1 and the points made in the preceding paragraphs, however, we shall take a different approach. In particular, we shall make an extra assumption and obtain a much simpler, and more practically useful, two-stage modeling strategy.

Although the problem that motivated our work has to do with screening marine cargo containers to support border management, the proposed two-stage strategy can be applied to any unbalanced classification problem with a time-varying decision boundary, e.g., statistical fraud detection [1].

We start by making the following assumption.

Assumption 2.1 *Let $y \in \{0, 1\}$ be the class label of an observation $(\mathbf{x}, t)$, where the class of interest is extremely rare. We assume that, given the class label, y, the (conditional) distribution of $\mathbf{x}$ does not change over time. In other words,*

$$p(\mathbf{x}|y, t) = p(\mathbf{x}|y). \tag{1}$$

Regardless of its actual validity, there is a practical reason for making this assumption. On the one hand, $p(\mathbf{x}|y, t)$ is often a high-dimensional probability distribution function and difficult to learn even under fairly generous conditions. On the other hand, any reasonable strategy must only allow samples collected near time t to substantially influence the learning of $p(\mathbf{x}|y, t)$. For example, one can create a time interval, $(t - \Delta t, t + \Delta t)$, and use only samples within that interval to learn $p(\mathbf{x}|y, t)$. It is possible to use all the samples but, to do so, one must weigh the ones near time t more heavily. Either way, only a fraction of the samples can effectively be used. Due to class imbalance, only a very small number of those samples would belong to the rare class. In situations like this, one simply does not have enough information about how $p(\mathbf{x}|y, t)$ changes over time and is effectively "forced" to operate as if Assumption 2.1 were true, even if one adopts an indirect strategy that does not learn the distribution $p(\mathbf{x}|y, t)$ explicitly.

If the distribution of $\mathbf{x}$ given y is assumed not to depend on t, then it is clear intuitively that either the distribution of y must depend on t — that is, $p(y|t) \neq p(y)$ — or the time variable t is entirely irrelevant. Theorem 2.1 formalizes this intuition; its proof is in the appendix.

Theorem 2.1 *Under Assumption 2.1, the following relationship holds:*

$$p(y|\mathbf{x}, t) = p(y|\mathbf{x}) \times \frac{p(y|t)}{p(y)} \times C(\mathbf{x}, t) \tag{2}$$

where $C(\mathbf{x}, t)$ is a quantity that does not depend on y.

2.1 Time Adjustment Factor (TAF)

We refer to the ratio, $p(y|t)/p(y)$, as the "time adjustment factor." If $p(y|t) = p(y)$, then this factor is equal to one and, by Theorem 2.1, all that matters for classification is the usual conditional distribution $p(y|\mathbf{x})$ — the time variable t becomes irrelevant. Notice that the quantity, $C(\mathbf{x}, t)$, does not affect classification because it does not involve y; it is simply a normalizing constant to ensure that

$$\sum_y p(y|\mathbf{x}, t) = 1, \quad \text{for any given } \mathbf{x} \text{ and } t.$$

Theorem 2.1, therefore, suggests a two-stage approach. To model $p(y|\mathbf{x}, t)$, first build a baseline classifier, $p(y|\mathbf{x})$, and then modify it by the "time adjustment factor." This modular strategy could be especially attractive to managers running real-world operations because the tasks could be easily streamlined; the analysis could be handled by two

separate teams: one with expertise in time-series analysis and forecasting, and another with expertise in supervised learning and classification.

Even though it may appear at first to be the most natural approach, it is generally not possible to model the full conditional distribution, $p(y|\mathbf{x}, t)$, by treating t simply as another predictor. For example, the best baseline classifier, $p(y|\mathbf{x})$, may be an SVM with the radial basis kernel function

$$K_h(\mathbf{x}_i; \mathbf{x}_j) = e^{-h\|\mathbf{x}_i - \mathbf{x}_j\|^2},$$

where h is a tuning parameter. Clearly, we cannot simply treat the time variable, t, as another predictor and use

$$K_h((\mathbf{x}_i, t_i); (\mathbf{x}_j, t_j)) = e^{-h\|(\mathbf{x}_i, t_i) - (\mathbf{x}_j, t_j)\|^2}.$$

In addition, when predicting the future, the time variable will always be outside the range of the training set, and treating t simply as another predictor in an ordinary classifier would often cause a disaster. This is precisely why our two-stage approach is especially valuable. It allows us to deal with the time variable separately so that we can easily exploit existing methods for analyzing time-series data.

2.2 Estimation of TAF

Any existing classifier can serve as the baseline classifier, $p(y|\mathbf{x})$. To obtain the "time adjustment factor," two quantities are needed, $p(y|t)$ and $p(y)$. As usual, $p(y)$ is easily estimated from sample class proportions. This leaves $p(y|t)$, which can be estimated in a variety of ways. Here, we briefly describe two options, but there is no reason why other methods cannot be applied.

2.2.1 Smoothing estimate

Recall that the class label y is coded as 0 for the background class and 1 for the rare class. One simple option is to estimate $p(y = 1|t)$ by the sample proportion of the rare class within a certain time interval immediately before t, e.g.,

$$\hat{p}(y = 1|t) = \frac{\sum_i y_i \delta(t - \Delta t \le t_i < t)}{\sum_i \delta(t - \Delta t \le t_i < t)}, \quad (3)$$

where

$$\delta(A) = \begin{cases} 1, & \text{if } A \text{ is true;} \\ 0, & \text{if } A \text{ is false.} \end{cases}$$

This can be viewed as a *smoothing* or a *moving-average* model, where Δt is the size of the moving window and a smoothing parameter.

2.2.2 Regression estimate

Alternatively, we can also build explicit *regression* models using the time variable, t. A quick way to do so is to consider equally spaced time intervals, e.g., weekly or monthly intervals. Suppose the intervals are $I(1), I(2), ..., I(T)$ — see Figure 1 for an illustration. For each $t = 1, 2, \ldots, T$, let

$$\pi(t) = \frac{1}{|I(t)|} \sum_{t_i \in I(t)} y_i$$

be the proportion of the rare class within the time interval $I(t)$. A crude estimate of $p(y|t)$ can be obtained simply by regressing $\pi(t)$ onto t. Different regression models can be considered, depending on the amount of information in the data, e.g., a simple linear model,

$$\mathrm{E}(\pi(t)) = \beta_0 + \beta_1 t, \quad t = 1, 2, ..., T, \quad (4)$$

which captures just an overall trend, or a more complicated model that includes a cyclic or a seasonal component, such as

$$\mathrm{E}(\pi(t)) = \beta_0 + \beta_1 t + \alpha \sin(\gamma_0 + \gamma_1 t), \quad t = 1, 2, ..., T. \quad (5)$$

An advantage of working with the fractions $\pi(1), \pi(2), ..., \pi(T)$ explicitly is easy exploration of different regression models by the method of least squares.

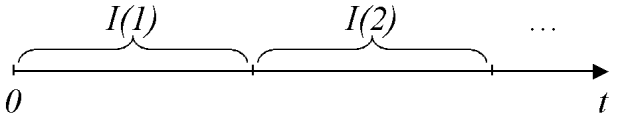

Figure 1: *Schematic illustration. Dividing the time axis into equally spaced intervals.*

3. RESULTS

In this section, we apply our two-stage approach to the CBSA marine container data (Section 1.1), and show that the "time adjustment factor" — $p(y|t)/p(y)$ — improves the performance of the baseline classifier, $p(y|\mathbf{x})$.

3.1 Performance Measure

The receiver-operating characteristic (ROC) curve [6] is a simple two-dimensional graph measuring a model's false positive rate versus its true positive rate over all possible decision thresholds. The area under the ROC curve, or simply "area under the curve" (AUC), is a common metric for performance evaluation, and is especially useful in this context, where we wish to evaluate how well a model assigns high risk scores to "dirty" containers and low risk scores to "clean" ones.

3.2 Baseline Classifier

Any classifier can be used as the baseline classifier, $p(y|\mathbf{x})$. Since the main focus of this paper is not on the baseline classifier and the method we developed in Section 2 is independent of which baseline classifier is used, we simply use our "Logit18" model as the baseline classifier, because it performed as well as other, more sophisticated classifiers such as SVM and random forest (see Section 1.1).

3.3 Details

Only surrogate time labels are available for estimating the TAF; the actual time labels (e.g. arrival on June 17, 2007) are considered classified information and cannot be used in this analysis.

The smoothing model (3) was fitted with different values of Δt (see Figure 2). It is evident from Figure 2 that, if Δt is too small, e.g., $\Delta t = 1$ month, the resulting model becomes very noisy, whereas, if Δt is too large, e.g., $\Delta t = 12$ months, the resulting model becomes almost flat. Therefore, only

"reasonable" values of Δt are investigated. In Table 1 below, we report results for four values: $\Delta t = 2, 3, 5$ and 6.

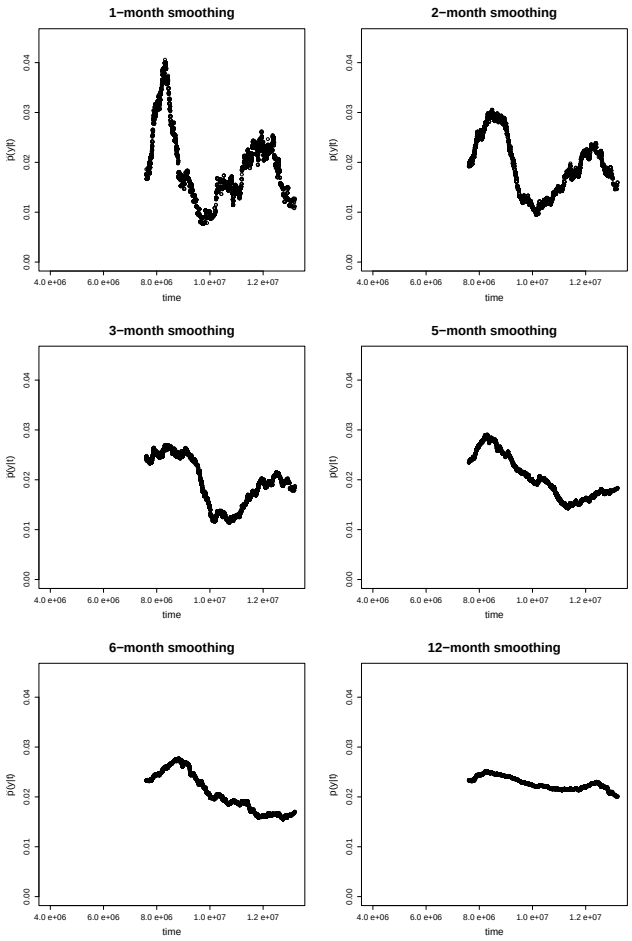

Figure 2: *CBSA Data. Smoothing model (3) fitted with different Δt's on the second half of the training set.*

To fit regression models (4) and (5), the time axis was first divided into 15 intervals, that is, $T = 15$. The training set spanned the first 10 intervals, and the test set spanned the remaining 5. If actual time labels had been available, we would have used more natural time intervals, such as monthly intervals. Here, we chose $T = 15$ so that all time intervals spanned by the training set contained at least one observation labeled "dirty," meaning that $\pi(t)$ is strictly positive and non-zero for every time interval in the training set. The method of least squares was then used to fit both regression models (4) and (5). Figure 3 shows the fitted regression functions.

3.4 Discussion

Table 1 shows the performance of these different models on the test set. In general, the two-stage strategy can be seen to improve the baseline model. It also makes it very easy for us to adopt more sophisticated time-series models, such as regression model (5). Even though (5) is still a very crude model, the improvement is quite substantial, especially in view of the fact that many sophisticated baseline classifiers such as SVM and random forest failed to offer any performance improvement over the simple Logit18 model (see Section 1.1).

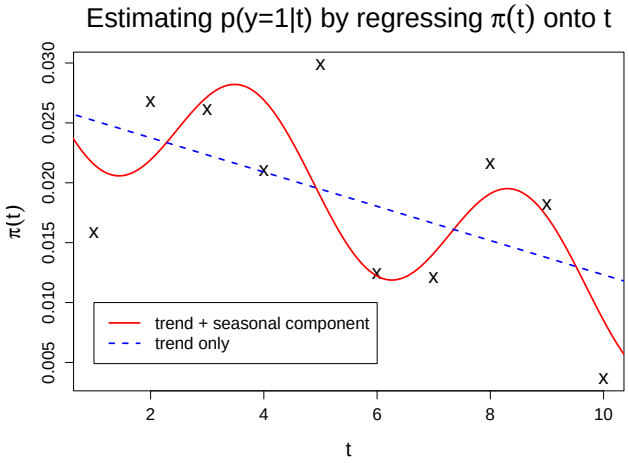

Figure 3: *CBSA Data. Regression models (4) and (5) fitted on the training set.*

Judging from Figures 2 and 3, there is overwhelming statistical evidence that $p(y)$ does depend on t. That the "time adjustment factor" (estimated using the training set alone) can improve the performance of the baseline classifier on the test set also strongly indicates that the dependence is real.

Table 1: *CBSA data. Performance of different modeling strategies on the test set.*

Model	AUC
Baseline - $p(y\|\mathbf{x})$	
$p(y\|\mathbf{x})$ - Logit18	0.728
Proposed - $p(y\|\mathbf{x}) \times p(y\|t)/p(y)$	
$p(y\|\mathbf{x})$ - Logit18; $p(y\|t)$ - smoothing	
model (3), $\Delta t = 2$ months	0.727
model (3), $\Delta t = 3$ months	0.740
model (3), $\Delta t = 5$ months	0.733
model (3), $\Delta t = 6$ months	0.732
$p(y\|\mathbf{x})$ - Logit18; $p(y\|t)$ - regression	
model (4)	0.745
model (5)	0.766

On the other hand, we are cautious about drawing any definite conclusions regarding the exact nature of the dependence. For example, regression model (5) has the best performance on the test set, but it is still too early to conclude that the probability of receiving a "dirty" container has such a regular cyclic pattern. Further research would need to be conducted to determine whether there are seasonal variations in contraband smuggling, or whether customs officers are more successful at identifying "dirty" containers in certain months of the year compared to others.

It is also worth pointing out that, while regression models (4) and (5) may appear to hold on the entire real line, periodic updating and re-calibration are absolutely necessary as new data become available over time. It is seldom the case that such simple stationary models can hold over a long pe-

riod of time. They are useful for predicting the immediate future but cannot be extrapolated into the long run. This is widely understood and, in practice, periodically retraining one's model is a common way to deal with concept drift [9]. With the right understanding of this kind, the apparent "problem" from Figure 3 that $\mathrm{E}(\pi(t))$ may eventually become negative for large t is of no real concern.

Overall, we find this two-stage approach to be valuable and are greatly encouraged by the positive results reported above.

4. SUMMARY

In our view, this work is significant in two ways. First, it made a tangible difference in the CBSA marine container problem. Accounting explicitly for the fact that the probability of receiving a "dirty" container may be changing over time constitutes a major conceptual step forward in our data-mining practice. Second, we have developed a method that is *not* limited to the particular CBSA application that motivated our work. Our two-stage approach of augmenting a baseline classifier by a "time adjustment factor" is general and easy to implement. It can be applied to many other data-mining and predictive-analytic problems.

Acknowledgments

This work was partially supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada, and by the Mathematics of Information Technology and Complex Systems (MITACS) network. Lu Xin and Nan Zhang experimented with a variety of baseline classifiers for the CBSA marine container data, the details of which were not directly reported in this paper (see Section 1.1). We also thank David Matthews of UW, Marc Gaudreau and Sylvain Coulombe of CBSA for their support of this project.

References

[1] R. J. Bolton and D. J. Hand. Statistical fraud detection: A review. *Statistical Science*, 17(3):235–255, 2002.

[2] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[3] N. Chawla, N. Japkowicz, and A. Kolcz. Editorial, special issue on learning from imbalanced data sets. *SIGKDD Explorations*, 6(1):1–6, 2004.

[4] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

[5] D. J. Hand. Classifier technology and the illusion of progress. *Statistical Science*, 21(1):1–14, 2006.

[6] M. S. Pepe. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, 2003.

[7] F. Provost. Comment on "statistical fraud detection: A review". *Statistical Science*, 17(3):249–251, 2002.

[8] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

[9] G. Widmer and M. Kubat. Learning in the presence of concept drift and hidden contexts. *Machine Learning*, 24:69–101, 1996.

APPENDIX

To prove Theorem 2.1, we first apply Bayes' theorem to $p(y|\mathbf{x},t)$ and get

$$p(y|\mathbf{x},t) = \frac{p(\mathbf{x}|y,t)p(y|t)}{p(\mathbf{x}|t)}.$$

Assumption 2.1 then implies

$$p(y|\mathbf{x},t) = p(\mathbf{x}|y) \times \frac{p(y|t)}{p(\mathbf{x}|t)}.$$

Now apply Bayes' theorem again, this time to $p(\mathbf{x}|y)$, and we get

$$p(y|\mathbf{x},t) = \frac{p(y|\mathbf{x})p(\mathbf{x})}{p(y)} \times \frac{p(y|t)}{p(\mathbf{x}|t)}. \quad (6)$$

Finally, by defining $C(\mathbf{x},t)$ to be $p(\mathbf{x})/p(\mathbf{x}|t)$, a quantity that does not depend on y, equation (6) can be easily re-arranged to give the desired result.