Computational Statistics and Data Analysis 55 (2011) 168–183

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Robust weighted kernel logistic regression in imbalanced and rare events data

Maher Maalouf*, Theodore B. Trafalis

School of Industrial Engineering, The University of Oklahoma, 212 West Boyd Street, Room 124, Norman, OK, 73019, United States

ARTICLE INFO

Article history:
Received 26 December 2009
Received in revised form 12 June 2010
Accepted 12 June 2010
Available online 25 June 2010

Keywords:
Classification
Endogenous sampling
Logistic regression
Kernel methods
Truncated Newton

ABSTRACT

Recent developments in computing and technology, along with the availability of large amounts of raw data, have contributed to the creation of many effective techniques and algorithms in the fields of pattern recognition and machine learning. The main objectives for developing these algorithms include identifying patterns within the available data or making predictions, or both. Great success has been achieved with many classification techniques in real-life applications. With regard to binary data classification in particular, analysis of data containing rare events or disproportionate class distributions poses a great challenge to industry and to the machine learning community. This study examines rare events (REs) with binary dependent variables containing many more non-events (zeros) than events (ones). These variables are difficult to predict and to explain as has been evidenced in the literature. This research combines rare events corrections to Logistic Regression (LR) with truncated Newton methods and applies these techniques to Kernel Logistic Regression (KLR). The resulting model, Rare Event Weighted Kernel Logistic Regression (RE-WKLR), is a combination of weighting, regularization, approximate numerical methods, kernelization, bias correction, and efficient implementation, all of which are critical to enabling RE-WKLR to be an effective and powerful method for predicting rare events. Comparing RE-WKLR to SVM and TR-KLR, using non-linearly separable, small and large binary rare event datasets, we find that RE-WKLR is as fast as TR-KLR and much faster than SVM. In addition, according to the statistical significance test, RE-WKLR is more accurate than both SVM and TR-KLR.

1. Introduction

Rare events (REs), class imbalance, and rare classes are critical to prediction and hence human response in the field of data mining and particularly data classification. Examples of rare events include fraudulent credit card transactions (Chan and Stolfo, 1998), word mispronunciation (Busser and Daelemans, 1999), tornadoes (Trafalis et al., 2003), telecommunication equipment failures (Weiss and Hirsh, 2000), oil spills (Kubat et al., 1998), international conflicts (King and Zeng, 2001a), state failure (King and Zeng, 2001b), landslides (Eeckhaut et al., 2006; Bai et al., 2008), train derailments (Quigley et al., 2007), rare events in a series of queues (Tsoucas, 1992) and other rare events.

By definition, rare events are occurrences that take place with a significantly lower frequency compared to more common events. Given their infrequency, rare events have an even greater value when correctly classified. However, the imbalanced distribution of classes calls for correct classification. The rare class presents several problems and challenges to existing classification algorithms (Weiss, 2004; King and Zeng, 2001c).

* Corresponding author. Tel.: +1 405 410 6365.
E-mail addresses: mcm@ou.edu (M. Maalouf), ttrafalis@ou.edu (T.B. Trafalis).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.06.014

Sampling is undoubtedly one of the most important techniques in dealing with REs. The underlying objective of sampling is minimizing the effects of rareness by changing the distribution of the training instances. Sampling techniques can be either basic (random) or advanced (intelligent). Van-Hulse et al. (2007) provide a comprehensive survey on both random and intelligent data sampling techniques and their impact on various classification algorithms. Seiffert et al. (2007) observed that data sampling is very effective in alleviating the problems presented by rare events.

Basic sampling methods consist of under-sampling and over-sampling. The former eliminates examples from the majority class, while the latter adds more training examples on behalf of the minority class. Over-sampling can thus increase processing time. In addition, over-sampling risks over-fitting, since it involves making identical copies of the minority class. Drummond and Holte (2003) found that under-sampling using C4.5 (a decision tree algorithm) is most effective for imbalanced data. Maloof (2003) showed, however, that under-sampling and over-sampling are almost equal in effect using Naive Bayes and C5.0 (a commercial successor to C4.5). Japkowicz (2000) came to similar conclusion but found that under-sampling the majority class works better on large domains. Prati et al. (2004), without providing conclusive evidence, proposed over-sampling combined with data cleaning methods as a possible remedy for classifying REs. The basic sampling strategy is known in econometrics and transportation studies as *choice-based*, *state-based* or *endogenous* sampling. In medical research it is known as *case control*. King and Zeng (2001c) advocate under-sampling of the majority class when statistical methods such as logistic regression are employed. They clearly demonstrated that such designs are only consistent and efficient with the appropriate corrections. Unfortunately, few researchers are aware of the fact that any kind of under-sampling is a form of choice-based sampling which leads to biased estimates. Thus, they proceed to solve likelihoods that are only appropriate for random sampling.

King and Zeng (2001c) state that the problems associated with REs stem from two sources. First, when probabilistic statistical methods, such as logistic regression, are used, they underestimate the probability of rare events, because they tend to be biased towards the majority class, which is the less important class. Second, commonly used data collection strategies are inefficient for rare events data. A trade-off exists between gathering more observations (instances) and including more informational, useful variables in the dataset. When one of the classes represents a rare event, researchers tend to collect very large numbers of observations with very few explanatory variables in order to include as many data as possible for the rare class. This in turn could significantly increase the data collection cost and not help much with the underestimated probability of detecting the rare class or the rare event.

Kernel Logistic Regression (KLR) (Canu and Smola, 2005; Jaakkola and Haussler, 1999), which is a kernel version of Logistic Regression (LR), has been proven to be a powerful classifier. Just like LR, KLR can naturally provide probabilities and extend to multi-class classification problems (Hastie et al., 2001; Karsmakers et al., 2007). The advantages of using LR are that it has been extensively studied (Hosmer and Lemeshow, 2000), and recently it has been improved through the use of truncated Newton methods (Komarek and Moore, 2005; Lin et al., 2007). This has also been shown recently for KLR by Maalouf and Trafalis (2008). Furthermore, LR and KLR do not make assumptions about the distribution of the independent variables. LR and KLR include the probabilities of occurrences as a natural extension. Moreover, LR and KLR can be extended to handle multi-class classification problems and they require solving only unconstrained optimization problems. Hence, with the right algorithms, the computation time can be much less than that for other methods, such as using Support Vector Machines (SVM) (Vapnik, 1995), which require solving a constrained quadratic optimization problem. In sum, King and Zeng (2001c) applied LR to REs data with the appropriate bias and probabilities corrections. Komarek and Moore (2005) implemented the TRuncated Newton method in LR (TR-IRLS). Maalouf and Trafalis (2008) implemented the TRuncated Newton method in KLR (TR-KLR).

The focus of this study is the implementation of fast and robust adaptations of KLR in imbalanced and rare events data. The algorithm is termed Rare Event Weighted Kernel Logistic Regression (RE-WKLR). The ultimate objective is to gain significantly more accuracy in predictive REs with diminished bias and variance. Weighting, regularization, approximate numerical methods, kernelization, bias correction, and efficient implementation are critical to enabling RE-KLR to be an effective and powerful method for predicting rare events. Our analysis involves the standard multivariate cases in *finite* dimensional spaces. Recent advances in *Functional Data Analysis* (FDA) (Ramsay and Silverman, 2005) and their extension to non-parametric functional data analysis (Ferraty and Vieu, 2006) allow for consideration of cases in which random variables take on *infinite* dimensional spaces (functional spaces).

In Section 2, we provide a brief description of sample selection bias. In Section 3, we give an overview of LR for rare events. Section 4 derives the KLR model for the rare events and imbalanced data problems. Section 5 describes the Rare Event Weighted Kernel Logistic Regression (RE-WKLR) algorithm. Numerical results are presented in Section 6, and Section 7 addresses the conclusions and future work.

2. Sample selection bias, endogenous sampling, and biased estimates

Following Zadrozny (2004), let s be a binary random variable, which takes the value of 1 if a sample is selected and 0 otherwise. Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be a data matrix where n is the number of instances (examples) and d is the number of features (parameters or attributes), and $\mathbf{y}$ be a binary outcomes vector. For every instance $\mathbf{x}_i \in \mathbb{R}^d$ (a row vector in $\mathbf{X}$), where $i = 1, \ldots, n$, the outcome is either $y_i = 1$ or $y_i = 0$. Let the instances with outcomes of $y_i = 1$ belong to the positive class, and the instances with outcomes $y_i = 0$ belong to the negative class. The goal is to classify the instance $\mathbf{x}_i$ as positive or negative. An instance can be thought of as a Bernoulli trial with an expected value $E[y_i]$ or probability p_i. In addition,