Knowledge-Based Systems 59 (2014) 142–148

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Weighted logistic regression for large-scale imbalanced and rare events data

Maher Maalouf [a,*], Mohammad Siddiqi [b]

[a] Industrial & Systems Engineering, Khalifa University, P.O. Box 127788, Abu Dhabi, United Arab Emirates
[b] Aerospace & Mechanical Engineering, Khalifa University, P.O. Box 127788, Abu Dhabi, United Arab Emirates

ARTICLE INFO

Article history:
Received 22 September 2013
Received in revised form 14 January 2014
Accepted 14 January 2014
Available online 27 January 2014

Keywords:
Classification
Endogenous sampling
Logistic regression
Kernel methods
Truncated Newton

ABSTRACT

Latest developments in computing and technology, along with the availability of large amounts of raw data, have led to the development of many computational techniques and algorithms. Concerning 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. Logistic Regression (LR) is a powerful classifier. The combination of LR and the truncated-regularized iteratively re-weighted least squares (TR-IRLS) algorithm, has provided a powerful classification method for large data sets. This study examines imbalanced data with binary response variables containing many more non-events (zeros) than events (ones). It has been established in the literature that these variables are difficult to predict and explain. This research combines rare events corrections to LR with truncated Newton methods. The proposed method, Rare Event Weighted Logistic Regression (RE-WLR), is capable of processing large imbalanced data sets at relatively the same processing speed as the TR-IRLS, however, with higher accuracy.

1. Introduction

In recent years, much attention in the machine learning community has been drawn to the problem of imbalanced or rare-events data. There are two main reasons for this. The first is that most of the traditional models and algorithms are based on the assumption that the classes in the data are balanced or evenly distributed. However, in many real-life applications the data is imbalanced, and when the imbalance is extreme, this problem is termed the *rare events* problem or the *imbalanced data* problem. Hence, the rare class presents several problems and challenges to existing classification algorithms [1,2].

The second reason for concern is the importance of rare events in real-life applications. By definition, rare events are occurrences that take place with a substantially lower frequency than commonly occurring events. Applications such as internet security [3], bankruptcy early warning systems and predictions [4,5] are gaining more importance in recent years. Other examples of rare events include fraudulent credit card transactions [6], word mispronunciation [7], tornadoes [8], telecommunication equipment failures [9], oil spills [10], international conflicts [11], state failure [12], landslides [13,14], train derailments [15] and rare events in a series of queues [16] among others.

King and Zeng [2] state that the problems associated with REs stem from two sources. First, when probabilistic statistical methods, such as Logistic Regression (LR), 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 dilemma exists between gathering more observations (instances) and including more informational, useful variables in the data set. 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 much data as possible for the rare class. This in turn could significantly increase the cost of data collection without boosting the underestimated probability of detecting the rare class or the rare event. King and Zeng [2] advocate under-sampling of the majority class when statistical methods such as LR are employed. They clearly demonstrated, however, that such designs are only consistent and efficient with the appropriate corrections.

Linear classification is an extremely important machine-learning and data-mining tool. Compared to other classification techniques, such as the kernel methods, which transform data into higher dimensional space, linear classifiers are implemented directly on data in their original space. The main advantage of linear classifiers is their efficient training and testing procedures, especially when

* Corresponding author. Tel.: +971 24018000.
E-mail addresses: maher.maalouf@kustar.ac.ae (M. Maalouf), mohammad.siddiqi@kustar.ac.ae (M. Siddiqi).

http://dx.doi.org/10.1016/j.knosys.2014.01.012
0950-7051/

implemented on large and high-dimensional data sets [17]. Logistic regression [18,19], which is a linear classifier, has been proven to be a powerful classifier by providing probabilities and by extending to multi-class classification problems [20,21]. The advantages of using LR are that it has been extensively studied [22], and recently it has been improved through the use of truncated Newton's methods [23–27]. Furthermore, LR does not make assumptions about the distribution of the independent variables and it includes the probabilities of occurrences as a natural extension. Moreover, LR requires solving only unconstrained optimization problems. Hence, with the right algorithms, the computation time can be much less than that of other methods, such as Support Vector Machines (SVM) [28], which require solving a constrained quadratic optimization problem. Komarek [29] were the first to implement the truncated-regularized iteratively re-weighted least squares (TR-IRLS) on LR to classify large data sets, and they demonstrated that it can outperform the Support Vector Machines (SVM) algorithm. Later on, trust region Newton method [24], which is a type of truncated Newton, and truncated Newton interior-point methods [30] were applied for large scale LR problems.

The objective of this study is to provide a basis for solving problems with data that are at once large and imbalanced or rare-event data. This paper is an extension of the work proposed by Maalouf and Saleh [31], which introduces the implementation of LR rare-event corrections to the TR-IRLS algorithm. The algorithm proposed is termed Rare Event-Weighted Logistic Regression (RE-WLR), and is based on the RE-WKLR algorithm, developed by Maalouf and Trafalis [32]. The RE-WKLR is appropriate for small-to-medium size data sets in terms of both computational speed and accuracy. The ultimate objective is to gain significantly more accuracy in predictive REs with diminished bias and variance. Weighting, regularization, approximate numerical methods, bias correction, and efficient implementation are critical to enabling RE-WLR to be an effective and powerful method for predicting rare events in large data sets. Our analysis involves the standard multivariate cases in *finite* dimensional spaces.

In Section 2 we derive the LR model for the rare events and imbalanced data problems. Section 3 describes the Rare-Event Weighted Logistic Regression (RE-WLR) algorithm. Numerical results are presented in Section 4, and Section 5 addresses the conclusions and future work.

2. Logistic regression and sampling on the dependent variable

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 $i = 1 \dots N$, 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 treated as a Bernoulli trial with an expected value $E(y_i)$ or probability p_i. The logistic function commonly used to model each instance $\mathbf{x}_i$ with its expected outcome is given by the following formula [22]:

$$E[y_i|\mathbf{x}_i,\beta] = p_i = \frac{e^{\mathbf{x}_i\beta}}{1+e^{\mathbf{x}_i\beta}}, \tag{1}$$

where β is the vector of parameters with the assumption that $x_{i0} = 1$ so that the intercept β_0 is a constant term. From then on, the assumption is that the intercept is included in the vector β.

The logistic (logit) transformation is the logarithm of the odds of the positive response and is defined as

$$\eta_i = \ln\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i\beta. \tag{2}$$

In matrix form, the logit function is expressed as

$$\eta = \mathbf{X}\beta. \tag{3}$$

Now, with the assumption that the observations are independent, the likelihood function is

$$\mathbb{L}(\beta) = \prod_{i=1}^{\ell}(p_i)^{y_i}(1-p_i)^{1-y_i} = \prod_{i=1}^{\ell}\left(\frac{e^{\mathbf{x}_i\beta}}{1+e^{\mathbf{x}_i\beta}}\right)^{y_i}\left(\frac{1}{1+e^{\mathbf{x}_i\beta}}\right)^{1-y_i}. \tag{4}$$

The regularized log likelihood [22] is defined as

$$\log\mathbb{L}(\beta) = \sum_{i=1}^{\ell}\left(y_i\log\left(\frac{e^{\mathbf{x}_i\beta}}{1+e^{\mathbf{x}_i\beta}}\right) + (1-y_i)\log\left(\frac{1}{1+e^{\mathbf{x}_i\beta}}\right)\right) - \frac{\lambda}{2}\|\beta\|^2, \tag{5}$$

$$= -\sum_{i=1}^{\ell}\log\left(e^{-y_i\mathbf{x}_i\beta}(1+e^{\mathbf{x}_i\beta})\right) - \frac{\lambda}{2}\|\beta\|^2, \tag{6}$$

where the regularization (penalty) term $\frac{\lambda}{2}\|\beta\|^2$ was added to obtain better generalization. Since the log likelihood function is strictly concave, the objective is then to find the Maximum Likelihood Estimate (MLE), $\hat{\beta}$, which maximizes the log likelihood. For binary outputs, the loss function or the deviance $\mathbb{DEV}$ is the negative log likelihood and is given by the formula [29,22]

$$\mathbb{DEV}(\hat{\beta}) = -2\ln\mathbb{L}(\beta). \tag{7}$$

Minimizing the deviance $\mathbb{DEV}(\hat{\beta})$ given in (7) is equivalent to maximizing the log-likelihood given in (2) [22]. The deviance function (above) is nonlinear in β. Minimizing it requires numerical methods in order to find the Maximum Likelihood Estimate (MLE) of β, which is $\hat{\beta}$. Recent studies have shown that the CG method provides better results to estimate β than any other numerical method [33,34].

When one of the $\mathbf{y}$ classes is rare in the population, then random selection within values of $\mathbf{y}$ would save significant resources in data collection [2,35]. Several advantages are associated with the selection on the response variable. First, in conducting surveys, cost reduction and time saving can be achieved by using stratified samples instead of collecting random samples, especially when the event of interest is rare in the population. Second, greater computational efficiency can be reached, because there is no need to analyze massive data sets. Finally, the explanatory power of the Logistic model can be enriched by making the proportions of events and non-events more balanced [2]. However, since the objective is to derive inferences about the population from the sample, the estimates obtained by the common likelihood using pure endogenous sampling are inconsistent. To see why this is so, under pure endogenous sampling, the conditioning is on $\mathbf{X}$ rather than $\mathbf{y}$ [36,37], and the joint distribution of $\mathbf{y}$ and $\mathbf{X}$ in the sample is

$$f_s(\mathbf{y},\mathbf{X}|\beta) = P_s(\mathbf{X}|\mathbf{y},\beta)P_s(\mathbf{y}), \tag{8}$$

where β is the unknown parameter vector to be estimated. Yet, since $\mathbf{X}$ is a matrix of exogenous variables, then the conditional probability of $\mathbf{X}$ in the sample is equal to that in the population, or $P_s(\mathbf{X}|\mathbf{y},\beta) = P(\mathbf{X}|\mathbf{y},\beta)$. However, the conditional probability in the population is

$$P(\mathbf{X}|\mathbf{y},\beta) = \frac{f(\mathbf{y},\mathbf{X}|\beta)}{P(\mathbf{yF})}, \tag{9}$$

but

$$f(\mathbf{y},\mathbf{X}|\beta) = P(\mathbf{y}|\mathbf{X},\beta)P(\mathbf{X}), \tag{10}$$

and hence, substituting and rearranging yields

$$f_s(\mathbf{y},\mathbf{X}|\beta) = \frac{P_s(\mathbf{y})}{P(\mathbf{y})}P(\mathbf{y}|\mathbf{X},\beta)P(\mathbf{X}), \tag{11}$$

$$= \frac{H}{Q}P(\mathbf{y}|\mathbf{X},\beta)P(\mathbf{X}), \tag{12}$$

where $\frac{H}{Q} = \frac{P_s(\mathbf{y})}{P(\mathbf{y})}$, with H representing the proportions in the sample and Q the proportions in the population. The likelihood is then