Expert Systems with Applications 38 (2011) 5110–5118

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Regularized logistic regression without a penalty term: An application to cancer classification with microarray data

Concha Bielza [a,*], Víctor Robles [b], Pedro Larrañaga [a]

[a] Department of Artificial Intelligence, Technical University of Madrid, Madrid, Spain
[b] Department of Computer Architecture and Technology, Technical University of Madrid, Madrid, Spain

ARTICLE INFO

Keywords:
Logistic regression
Regularization
Estimation of distribution algorithms
Cancer classification
Microarray data

ABSTRACT

Regularized logistic regression is a useful classification method for problems with few samples and a huge number of variables. This regression needs to determine the regularization term, which amounts to searching for the optimal penalty parameter and the norm of the regression coefficient vector. This paper presents a new regularized logistic regression method based on the evolution of the regression coefficients using estimation of distribution algorithms. The main novelty is that it avoids the determination of the regularization term. The chosen simulation method of new coefficients at each step of the evolutionary process guarantees their shrinkage as an intrinsic regularization. Experimental results comparing the behavior of the proposed method with Lasso and ridge logistic regression in three cancer classification problems with microarray data are shown.

1. Introduction

Logistic regression (Hosmer & Lemeshow, 2000) is a simple and efficient supervised classification method that provides explicit probabilities of class membership and an easy interpretation of the regression coefficients of predictor variables. The class variable is binary while the explanatory variables are of any type, not even requiring strong assumptions, like gaussianity of the predictor variables given the class or assumptions about the correlation structure. This lends great flexibility to this approach having shown a very good performance in a variety of fields (Baumgartner et al., 2004; Kiang, 2003).

Many of the most challenging current classification problems involve extremely high dimensionality k (thousands of variables) and small sample sizes N (less than one hundred cases). This is the so-called "large k, small N" problem, since it hinders proper parameter estimation when trying to build a classification model. Microarray data classification falls into this category.

In logistic regression we identify four problems in the "large k, small N" case. First, a large number of parameters – regression coefficients – have to be estimated using a very small number of samples. Therefore, an infinite number of solutions is possible as the problem is undetermined. Second, multicollinearity is largely present. As the dimensionality of the model increases, the chance grows that a variable can be constructed as a linear combination of other predictor variables, thereby supplying no new information. Third, over-fitting may occur, i.e. the model may fit the training data well but perform badly on new samples. These problems yield unstable parameter estimates. Fourth, there are also computational problems due to the large number of predictor variables. Traditional algorithms for finding the estimates numerically, like Newton–Raphson's method (Thisted, 1988), require prohibitive computations to invert a huge, sometimes singular matrix, at each iteration.

Within the context of logistic regression, the "large k, small N" problem has been tackled from three fronts: dimensionality reduction, feature (or variable) selection and regularization, or sometimes a combination of them.

As regards *dimensionality reduction*, principal components analysis is one of the most widespread methods (Aguilera, Escabias, & Valderrama, 2006). This preprocessing of high-dimensional variables outputs transformed variables, of which only a reduced set is used. These transformed variables are the classifier inputs. The main drawback is that principal components tend to need all the original variables in their expressions. As a result, the information requirements of model application are not reduced and there is also a loss of interpretability of the variables. Furthermore, there is not guarantee of class separability coinciding with the selected principal components (Weber, Vinterbo, & Ohno-Machado, 2004). Other methods, such as partial least squares (Antoniadis, Lambert-Lacroix, & Leblanc, 2003) or an adaptive dimension reduction through regression (Nguyen & Rocke, 2002) have also been used.

* Corresponding author. Tel.: +34 913366596; fax: +34 913524819.
E-mail addresses: mcbielza@fi.upm.es (C. Bielza), pedro.larranaga@fi.upm.es (V. Robles), vrobles@fi.upm.es (P. Larrañaga).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.09.140

Feature selection methods yield parsimonious models which reduce information costs, are easier to explain and understand, and increase model applicability and robustness. The selected features are good for discriminating between the different classes and may be sought via different heuristic search approaches (Liu & Motoda, 2008). The goodness of a proposed feature subset may be assessed via an initial screening process using a scoring metric. The metric is based on intrinsic characteristics of the data computed from simple statistics on the empirical distribution, totally ignoring the effects of the selected features on classifier performance. This is the so-called *filter* approach to feature selection in machine learning, or *screening* in statistics (West et al., 2001). By contrast, the *wrapper* approach searches good subsets using the classifier itself as part of their function evaluation (Kohavi & John, 1997). A performance estimate of the classifier trained with each subset assesses the merit of this subset. Some recent studies combine filter and wrapper approaches (Uncu & Türksen, 2007). In the context of logistic regression and $k \gg N$, Lee, Lee, Park, and Song (2005) propose different filter metrics to select a fixed number of features, the top-ranked ones, such that they are always fewer than the sample size.Avoiding the curse of dimensionality in a similar way, Weber et al. (2004) perform a preliminary feature selection by choosing the $N-1$ variables maximally correlated with the class variable. In a second phase, a logistic regression model is constructed with the selected features, and it is further simplified via a backwards variable selection.

The third front to tackle the "large k, small N" problem is using *regularization* methods. These methods impose a penalty on the size of logistic regression coefficients, trying to shrink them towards zero. Therefore, regularized estimators are restricted maximum likelihood estimators (MLE), since they maximize the likelihood function subject to restrictions on the logistic regression parameters. The little bias allowed provides more stable estimates with smaller variance. Regularization methods are more continuous than usual discrete processes of retaining-or-discarding features thereby not suffering as much from high variability (Hastie, Tibshirani, & Friedman, 2001). This shrinkage of coefficients was initially introduced in the ordinary linear regression scenario by Hoerl and Kennard (1970), where restrictions were spherical. This is the so-called ridge or quadratic (penalized) regression. Lee and Silvapulle (1988), LeCessie and vanHouwelingen (1992) extended the framework to logistic regression. Ridge estimators are expected to be on average closer to the real value of the parameters than the ordinary unrestricted MLEs, i.e. with smaller mean-squared error. See Fan and Li (2006), Bickel and Li (2006) for recent developments and a unified conceptual framework of the regularization theory.

Here we introduce *estimation of distribution algorithms* (EDAs) as intrinsic regularizers within the logistic regression context. EDAs are optimization heuristics included in the class of stochastic population-based search methods (Larrañaga & Lozano, 2002; Lozano, Larrañaga, Inza, & Bengoetxea, 2006; Pelikan, 2005). EDAs work by constructing an explicit probability model from a set of selected solutions, which is then conveniently used to generate new promising solutions in the next iteration of the evolutionary process. In our proposal, an EDA obtains the regularized estimates in a direct way in the sense that the objective function to be optimized is still the likelihood, not including any regularization term. It is a specifically chosen simulation process during the evolution which accounts intrinsically for the regularization. EDAs receive the unrestricted likelihood equations as inputs and generate the restricted MLEs as outputs.

The paper is organized as follows. Section 2 reviews both the classical and regularized versions of the logistic regression model. Section 3 describes EDAs and how we propose to use them to solve the regularized case. Experimental studies on several microarray data sets, a great exponent of the "large k, small N" problem, are presented in Section 4. Finally, Section 5 includes some conclusions and future work.

2. Regularized logistic regression

2.1. The need for regularizing logistic regression

Assume we have a (training) data set $\mathcal{D}_N$ of N independent samples from some experiment. $\mathcal{D}_N = \{(c_j, x_{j1}, \ldots, x_{jk}),\ j = 1, \ldots, N\}$, where $\mathbf{x}_j = (x_{j1}, \ldots, x_{jk})^t \in \mathbb{R}^k$ is the value of the jth sample, x_{ji} indicates the ith variable outcome of the jth sample and c_j is the known class label of the jth sample, 0 or 1, for the binary case considered in this paper.

Logistic regression uses the $\mathbf{x}$ values to determine the probability π of a sample belonging to one of the two classes. Thus, we have $k+1$ variables: the class or response dichotomous variable C and its predictor variables or covariates $X_1, \ldots, X_k$. The logistic model should be able to classify any new sample that comes along, characterized by just its covariate values.

Let π_j denote $P(C = 1|\mathbf{x}_j)$, $j = 1, \ldots, N$. Then the logistic regression model is defined as

$$\log \frac{\pi_j}{1-\pi_j} = \beta_0 + \sum_{i=1}^{k} \beta_i x_{ji} = \eta_j \Longleftrightarrow \pi_j = \frac{1}{1+e^{-\eta_j}} \tag{1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^t$ denotes the vector of regression coefficients including a constant or intercept β_0. These are usually estimated from data by the maximum likelihood estimation method. From $\mathcal{D}_N$, the log-likelihood function is built as

$$l(\boldsymbol{\beta}) = \sum_{j=1}^{N} (c_j \log \pi_j + (1-c_j)\log(1-\pi_j)), \tag{2}$$

where π_j is given by expression (1). Maximum likelihood estimators, $\widehat{\beta}_i$, are obtained by maximizing l with respect to $\boldsymbol{\beta}$. Let $\mathbf{c}$ denote the vector of response values c_j ($j = 1, \ldots, N$), $\boldsymbol{\pi}$ be the vector of π_j values, $\mathbf{X}$ be an $N \times k$ matrix with each row given by $\mathbf{x}_j^t$, and $\mathbf{u}$ an N-vector of ones. Thus, the following system of $k+1$ equations and $k+1$ unknowns – called the likelihood equations – has to be solved:

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \mathbf{Z}^{\mathbf{t}}(\mathbf{c} - \boldsymbol{\pi}) = \mathbf{0},$$

where $\mathbf{Z}$ is the matrix $[\mathbf{u}|\mathbf{X}]$.

Newton–Raphson's algorithm is traditionally used to solve the resulting *nonlinear* equations for $\widehat{\beta}_i$ numerically. Each iteration provides an updating formula given by

$$\hat{\boldsymbol{\beta}}^{\text{new}} = \hat{\boldsymbol{\beta}}^{\text{old}} + (\mathbf{Z}^{\mathbf{t}}\mathbf{W}^{\mathbf{old}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathbf{t}}(\mathbf{c} - \hat{\boldsymbol{\pi}}^{\text{old}}),$$

where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)^t$, and $\hat{\boldsymbol{\pi}}$ denotes the vector of estimated values at that iteration, i.e. its jth-component is

$$\hat{\pi}_j^{\text{old}} = \left[1 + e^{-(\hat{\beta}_0^{\text{old}} + \hat{\beta}_1^{\text{old}} x_{j1} + \cdots + \hat{\beta}_{kx}^{\text{old}} x_{jk})}\right]^{-1}, \quad j = 1, \ldots, N$$

and $\mathbf{W}^{\text{old}}$ denotes a diagonal matrix with elements $\hat{\pi}_j^{\text{old}}\left(1 - \hat{\pi}_j^{\text{old}}\right)$.

In the context of data involving high dimensionality (k) and small sample sizes (N), the logistic regression approach has a number of problems, explained in the introduction section: undetermined problem to be solved, multicollinearity, over-fitting and computational difficulties. Regularization emerges as one of the most promising solutions for these problems. In this section we review the state-of-the-art in the case of regularized logistic regression.

Regularized logistic regression maximizes the penalized log-likelihood given by