Classification of Imbalanced Data with Random Sets and Mean-Variance Filtering

Vladimir Nikulin, Suncorp, Australia

ABSTRACT

Imbalanced data represent a significant problem because the corresponding classifier has a tendency to ignore patterns which have smaller representation in the training set. We propose to consider a large number of balanced training subsets where representatives from the larger pattern are selected randomly. As an outcome, the system will produce a matrix of linear regression coefficients where rows represent random subsets and columns represent features. Based on the above matrix we make an assessment of the stability of the influence of the particular features. It is proposed to keep in the model only features with stable influence. The final model represents an average of the single models, which are not necessarily a linear regression. The above model had proven to be efficient and competitive during the PAKDD-2007 Data Mining Competition.

Keywords: boosting; classification; feature selection; random forest

INTRODUCTION

We consider a standard binary classification applied to strongly imbalanced training data where we assume without loss of generality that a “positive” class represents a minority.

Ideally, our target is to find a transformation from a multidimensional space of features to one-dimensional Euclidean space in order to maximize the difference between different classes and minimize volatility inside classes (Huber, 1985).

Our approach was motivated by Breiman (1996, 2001) and represents a compromise between two major tendencies. On the one hand, we would like to deal with balanced data. On the other hand, we are interested in exploiting all available information. Respectively, we consider a large number n of balanced subsets of available data where any single subset includes two parts 1) all “positive” instances and 2) randomly selected “negative” instances.

Linear regression represents the simplest example of decision function. Combined with quadratic loss function, it has an essential advantage: using gradient-based search procedures, we can optimise the value of the step size.
Consequently, we will observe rapid decline of the target function (Hastie, Tibshirani, & Friedman, 2001). However, squared loss function excessively penalizes large values of the decision functions. In order to overcome this problem we can apply a generalised linear model with exponential loss function. Boosting algorithms (Friedman, Hastie, & Tibshirani, 2000) may be efficient in order to facilitate optimisation procedure. Furthermore, it appears to be natural to link any boosting iteration with randomly selected balanced subset.

By definition, regression coefficients may be regarded as natural measurements of influence of the corresponding features. In our case, we have \( n \) vectors of regression coefficients, and can investigate stability of the particular coefficients.

Feature selection problem is very important (Guyon, Weston, Barnhill, & Vapnik, 2002) in order to reduce overfitting. We remove features with unstable coefficients, and recompute classifiers. Note that stability of the coefficients may be measured using different methods. For example, we can apply \( t \)-statistic as a ratio of mean to standard deviation.

Clearly, selection of a linear regression as a base model is not necessary. Based on cross-validation experiments, we can apply another model such as decision trees, random forests or neural networks. The list of options is not limited and may be continued further (Guyon, Alamdari, Dror, & Buhmann, 2006). By definition, the final decision function represents a sample average of single decision functions.

The proposed approach is general and flexible. We cannot expect that a single algorithm will work optimally on all conceivable applications and, therefore, an opportunity of tuning and tailoring is very essential.

Experiments were conducted during the time of the PAKDD-2007 Data Mining Competition using real-world data, which were provided by a consumer finance company, with the aim of finding better solutions for a cross-selling problem. The data are strongly imbalanced, with a significantly smaller proportion of positive cases, which have the following practical interpretation: a customer opened a home loan with the company within 12 months after opening the credit card.

**MODELING TECHNIQUE**

Let \( X = (x_t, y_t), t = 1 \ldots m \), be a training sample of observations where \( x_t \in \mathbb{R}^l \) is a vector of features, and \( y_t \) is binary label: \( y_t \in \{0, 1\} \).

In a practical situation, the label \( y_t \) may be hidden, and the task is to estimate it using a vector of features. Let us consider the most simple linear decision function

\[
    u_t = u(x_t) = \sum_{j=0}^{l} w_j \cdot x_{tj},
\]

where \( x_{t0} \) is a constant term.

We can define the decision rule as a function of decision function and threshold parameter

\[
    f_t = f(u_t, \Delta) = \begin{cases} 
        1 & \text{if} \quad u_t \geq \Delta; \\
        -1, & \text{otherwise.}
    \end{cases}
\]

We used the area under Receiver Operating Curve (ROC) as an evaluation criterion. By definition, Receiver Operating Curve (ROC) is a graphical plot of True Positive Rates (TPR) against False Positive Rates (FPR) (see Figure 2(a, b)).

According to the proposed method, we consider large numbers of classifiers where any particular classifier is based on a relatively balanced subset with all “positive” and randomly selected (without replacement) “negative” data. The final decision function \( (d.f.) \) has a form of logistic average of the single decision functions.

**Definition 1.** We call the above subsets as random sets \( RS(\alpha, \beta, n) \) where \( \alpha \) is a number of positive cases, \( \beta \) is a number of negative cases and \( n \) is the total number of random sets.
Cooperative Learning and Virtual Reality-Based Visualization for Data Mining
[www.igi-global.com/chapter/cooperative-learning-virtual-reality-based/7596?camid=4v1a](www.igi-global.com/chapter/cooperative-learning-virtual-reality-based/7596?camid=4v1a)