Partially Supervised Feature Selection with Regularized Linear Models

Daniel Cerdán, Fernando Freire

April 3, 2019

Contents

1 Partially Supervised Feature Selection with Regularized Linear Models

1.1 Summary

1.1.1 Feature selection methods overview

This item is based con the first paper.

Goals of feature selection

Scenarios related to few tens of samples but thousands dimensions: microarray data,

1. To avoid overfitting and improve model performance, prediction performance in the case of supervised classification and better cluster detection in unsupervised scenarios.
2. To provide more efficient models
3. To gain a deeper insight into the underlying processes that generated the data. The excess of dimensionality difficult the understanding.

The problem is related to find the optimal model parameters for the optimal feature subset. So, the model parameters becomes dependent of the features selected and need to be computed more or less coupled with the guessing of model parameters.

From less (zero) to more coupled computation, we have three strategies:

1. Filter techniques. Two step process, firs the filtering, then the training of the model. Take into account only the properties of the data and in some cases a certain amount of prior knowledge. Therefore it's independent of the classification method. In its most simplest form ignores dependences on the data (univariate).

   Examples: Euclidean distance, i-test Information gain, Markov blanket filter

2. Wrapper methods. Once selected a candidate subset of features, the classification model is evaluated by training and testing the model. This is iterated over a ensemble of candidate subsets, and the model (with his feature subsets) selected is the model with the best accuracy.

   It's very important to construct a good searching algorithm of subsets, in order to reduce the number of sets to model with. This methods are dependent of the classifier, model feature dependencies and have the risk to be bind to a local optima. With randomizing techniques this problem is bypassed to some extent.

   Examples: Sequential forward selection (SFS) , Sequential backward elimination, Simulated annealing, Randomized hill climbing, Genetic algorithms.

3. Embedded methods. The search of the optimal subset of features is built into the classifier. Have the advantage that they include the interaction with the classification model, while at the same time being far less computationally intensive than wrapper methods.

   Examples: Decision trees Weighted naive Bayes, Feature selection using the weight vector of SVM, AROM

1.1.2 AROM methods

The acronym derives from *Approximation of Minimization zeRO-norm*

The problem is obtain a linear predictor h, minimizing the number of independent variables (features) without loss of accuracy:

$$h(\mathbf{x}) = sign(\mathbf{w} \cdot \mathbf{x} + b)$$

for n samples $x_i \in \mathbb{R}^n$ and m labels $y_i \in \{\pm 1\}$.
The accuracy constraint requires correspondence of sign
$sign(y_i) \cdot sign(h_i) > 0$ or in other form $y_i \cdot h_i = 1$
or less restrictive, enabling $\mathbf{w}$ to scale freely $y_i \cdot h_i \geq 1$
so

$$y_i(\mathbf{w} \cdot \mathbf{x} + b) \geq 1$$

The minimization is done with a norm defined over the vectorial space of $\mathbf{w}$. One approach is to minimize the zero-norm, that is, the number of components of the vector (number of non null w_i). But it's know to be a NP-Hard problem.

It's more adequate compute over a 1-norm or a 2-norm. In the second paper, the author deduce a suitable form for the function that could be minimized, taken into account the former constraint:

$$\sum_{j=1}^{n} ln(|w_j| + \epsilon)$$

The term ϵ is included to protect from zero values inside logarithm.

AROM methods are therefore feature selection embedded methods.

l1-AROM and **l2-AROM** (in this case by means of a 2-norm minimization) algorithms optimize this algorithm by iterative rescaling of inputs and doing a smooth feature selection since the weight coefficients along some dimensions progressively drop below the machine precision while other dimensions become more significant.

1.1.3 AROM semi-supervised

Third and Fourth papers explore a improvement of these previous described methods.

Goal

Classification of microarray data: few tens of samples against several thousand dimensions (genes).

Key differential strategy

Extend AROM methods by means of partial supervision on the dimensions of a feature selection procedure. The technique proposes to use of prior knowledge to guide feature selection, but flexible enough to let the final selection depart from it if necessary to optimize the classification objective.

The preferential features are previously selected from similar datasets in large microarray databases because it's known that different sub-samples of patients lead to very similar sets of biomarkers, as expected if we are aware that the biological process explaining the outcome is common among different patients.

This datasets are called source datasets and we expect that the prediction for a similar feature vector is the same than the prediction for this vector in our dataset (the target).

In third paper prior knowledge is incorporated by biological information

So, if we have some knowledge on the relative importance of each feature (either from actual prior knowledge or from a related dataset), the supervised AROM objective can be modified by adding a prior relevance vector $\beta = [\beta_1, ..., \beta_n]$ defined over the n dimensions and where $\beta_j > 0$ is the prior relevance of the j feature.

So in this case, the function to minimize in the case of 1-norm is:

$$\sum_{j=1}^{n} \frac{1}{\beta_j} ln(|w_j| + \epsilon)$$

1.2 Project planification