Unsupervised Discovery from Gene Tracking with RFE Classification Systems

Cesare Purlanello, Stefano Mezler, Grisver P. Futerman, Maria Stratani
ITC-irst, via Sommarive 14, 38050 Povo (Trento), Italy

Introduction

Recurrent Feature Elimination (RFE) [6] and its variants [9] [are dialing methods typically used in Support Vector Machine (SVM) ]s, allowing to determine the subset of relevant genes in experiments with microarrays.

The use of complex validation experiment (e.g.) requires that portions of the available data remain unused for training and are considered only for model accuracy evaluation, typically obtained by averaging cross-validation experiments (e.g., by cross-validation).

The cross-validation experiments allow to monitor the behavior of any single-sample during the RFE process. i.e., for different numbers of genes involved in the model building.

With the single-sample response during the RFE process allows to derive an indicator for the detection of outliers and of missed data. A smooth solution can be therefore obtained by avoiding to use of outliers during the training phase (see for example [7]). Moreover, it is a promising strategy for outliers discovery.

Objectives

The goal is to:

- outlier detection and visualization of the solution by data reduction;
- subspace discovery (to groups of patients are correctly classified by different grouped genes).

Background

3

Recursive Feature Elimination [6]

Given a binary classification data set D = (x_i, y_i), with where y_i = {−1, 1}.

For j = 1, 2, ..., J:

- train a SVM /; J(2j+1) + 1 = 2j+1, J = 1, 2, ..., j
- for each feature compute an importance index I(2j+1) = I(2j+1)
- depending on the weight vector in 
- discard the feature j such that (x_i)min(−min_j (2j+1)), i = 1, ..., N (eliminate the least important feature).

Dynamic Time Warping [8]

DTW allows to find an optimal match between the sequence of patterns, and the sequence of the segments. The problem is represented as finding the minimum distance between a template stream and a query stream. The algorithm searches to minimize distance between the segment of the input stream to that of the template stream, so that the total cost is minimized.

Methods

Outlier detection

Given a binary classification data set D = (x_i, y_i), with where y_i = {−1, 1}.

1. Compute R partitions of the output D into training T_r and validation V_r, r = 1, 2, ..., R.
2. For r = 1, 2, ..., R:

- train the RFE in its standard fashion on training T_r, obtaining a ranking list L_r of feature importance.
- Train classification models f(x) based on an increasing number of features n_r, selected according to L_r.
- Test the models f(x) on the validation V_r.

For each data point compute the outlying score as:

where R is the number of sample features and m is the number of the current features.

The error curves show a different shape according to the category of data they are associated to.

4. Use DTW to compute the distance between the two RFE models associated to each data point.

5. Choose the minimum distance using the DTW distance in order to identify different categories of data points.

Subspace discovery

The same procedure can be used to identify different groups of patterns, characterized by a different number of discriminant genes involved in the RFE process.

Results

Synthetic data

The data consist of 200 randomly extracted samples described by 30 features (of them 10 discriminants). Outliers were generated by inverting the label.

The procedure was able to correctly identify the two outliers. Two groups of hard points were also detected, characterized by generalization error 0.25 when all the variables were used and generalization error 0.80 when only the 5 (discriminant) variables were used.

Mouse data

The C57BL/6J Mice Model of Myocardial Infarction is a data set designed to study the molecular bases of heart failure associated with left coronary artery blockage. Samples are described by 2284 genes spotted on Affymetrix chips M 8.0 (details in [8]). We considered 20 selected groups from the left coronary artery (20 samples) and 20 samples from the corresponding region in the extensively angioplasty group (200 samples). Fisher’s exact tests revealed 4490 (100×) connected variables.

Finally, 2 groups can be distinguished within the heart failure cases, Group 1 (5 cases) is characterized by 3 genes coming from controls, while a second group 2 (4 cases) is characterized by at least 20 genes.

Conclusions

In this work an effective procedure for outlier detection has been introduced based on the sample-linking techniques. Repeatability for subspace discovery is still under investigation.

References


After outliers removal, it is possible to distinguish different patterns in the error curves of the patients.