Looking at Similarities through Relevance Vector Machine

Daniela Marconi[12], Valter Gattei[2], Renato Campanini[1]

1. Physics Department, University of Bologna, 2. Clinical and Experimental Hematology Unit, C.R.O., I.R.C.C.S., Aviano, Italy

Proposed analytical objective

In microarray data analysis most of the literature concerns the identification of significant differences in expression profiles between two or more classes. Both supervised and unsupervised algorithms are designed to distinguish, with a certain level of confidence, one class from another.

However, looking at microarray data in a prognostic and diagnostic clinical framework, not only differences could have a crucial role. In some cases similarities can give useful and, sometimes even more, important information.

The introduction of prognosticators based on classification algorithms, applied to disease microarray dataset, could be the next future (Vant'veer's group is one of the pioneers in this field [1]). But one of the limitation is that at moment the response of the classificator is just a "hard" binary decision. Instead it could be useful to obtain an estimate of the posterior probability of new sample's membership to a class , i.e. to say how much the new sample is "similar" to the selected class samples.

Another way to interpret a measure of similarities is a three class problem. The goal, given three classes, could be to establish, with a certain level of confidence, if the third one is similar to the first or the second one.

In this work we show that Relevance Vector Machine (RVM) [2] could be a possible solutions to the limitation of standard supervised classification. In fact, RVM offers many advantages compared, for example, with his well-known precursor (Support Vector Machine - SVM [3]). Among these advantages, the estimate of posterior probability of class membership represents a key feature to address the similarity issue. This is a highly important, but often overlooked, option of any practical pattern recognition system.

A brief summary of the analytical effort

RVM [2] is part of a general Bayesian framework for obtaining sparse solutions to regression and classification tasks utilising models linear in the parameters. Is a model of identical functional form to the popular state-of-the-art Support Vector Machine (SVM).

$$y(x;w) = \sum_{n=1}^{N} w_n K(x, x_n) + w_0$$

Consider a two-class problem with training points $X = \{x1,...,xN\}$ and corresponding class labels t={t_1,...,t_N} t with $t_i \in \{0,1\}$. Based on the Bernoulli distribution, the likelihood (the target conditional distribution) is expressed as:

$$p(t|w) = \prod_{i=1}^{N} \sigma\{(y(x_i))\}^{t_i} [1 - \sigma\{(y(x_i))\}]^{1-t_i}$$

Where σ(y) is the logistic sigmoid function.

The function to mximize is $J(w_1,..., w_n) = \sum_{i=1}^{N} \log p(t_i|w_i) + \sum_{i=1}^{N} \log p(w_i|\alpha_i^*)$

Compared to SVM, RVM is found to be advantageous on several aspects including: 1) The RVM decision function can be much sparser than the SVM classifier, i.e., the number of relevance vectors can be much smaller than that of support vectors; 2) RVM does not need the tuning of a regularization parameter (C) as in SVM during the training phase; 3) RVM have a posterior probability output. As a drawback, however, the training phase of RVM typically involves a highly nonlinear optimization process.

We, however, omit to compare the performance of RVM with other algorithms in the gene expression classification problems, because the topic was faced in the work of Li et al.[4]

Data and preprocessing

To show the potentiality of RVM in estimating similarities we use a subset of META-Analysis dataset proposed at CAMDA 2007 (GEO Accession GSE3494) of breast cancer, proposed in a three-class problem setup. The used subset is a clinical-characterized gene expression dataset. In GEO is possible to obtain the clinical characterization and the .CEL files. We focused on Tumor-Grade-three-class problem, so we have 67 samples of grade I (G1), 54 samples of grade 3 (G3) and 100 samples of grade 2 (G2). The goal is to find a model able to separate G1 from G3, then evaluate the third class G2 as test-set to obtain the probability for samples of G2 to be member of class G1 or class G3.We pre-processed the data using gcrma algorithm [3].After we filtered the data trough a regularized t-test with a significance level for fdr adjusted p-value of 0.01.Then we standardize the data with a Z-score normalization, before applying RVM with a distance kernel.

Results from a biological and clinical point of view

For this dataset we obtain a double output for each sample of class G2: the binary classification (class-membership) and the probability for the class-membership.

The interesting results from a clinical and biological point of view is that the 90% of G2 samples are classified as G1 with a probability greater then 50%, and 66% of these samples have a probability greater then 90%. The 10% that are classified as G3 have a probability lower then 40%. So we can conclude that breast cancer samples of grade II have a molecular profiling more similar to breast cancer samples of grade I, rather then to breast cancer samples of grade III. Also looking at this new information we can make some biological and clinical consideration based on the other clinical features (disease-specific events, lymph node status for example) that characterize patients with low probability class membership.

Conclusion and future work

The use of the RVM from a probabilistic point of view can solve many open prognostic and diagnostic problems in medicine and can give to specialists some more insight, then a hard binary classification.

However an intensive work has to be done on the introduction in the RVM algorithm of prior information (such as annotation information) in order to take full advantage of this Bayesian approach.

References

1. Buyse, M., et al. (2006).Validation and Clinical Utility of a 70-Gene Prognostic Signature for Women With Node-Negative Breast Cancer JNCI *Journal of the National Cancer Institute* **98**(17):1183-1192
2. Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag
3. Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine**.** *Journal of Machine Learning Research* **1**, 211–244
4. Li, Y., C. Campbell, and M. E. Tipping (2002). Bayesian automatic relevance determination algorithms for classifying gene expression data. *Bioinformatics* **18**(10), 1332–1339.
5. Zhijin,W., Irizarry, R.A., Gentleman, R., Martinez-Murillo, F., Spencer, F. (2004). A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *Journal of the American Statistical Association,* **99,** 909.