Combining feature ranking methods for high dimensional data analysis

Romaric Gaudel[1,2] and Antoine Cornuéjols[3]

[1] Équipe TAO Laboratoire de Recherche en Informatique, CNRS UMR 8623
Bâtiment 490, Université Paris-Sud, 91405 - Orsay Cedex
[2] ENS Cachan
[3] INAPG - OMIP, 16, rue Claude Bernard, 75005 - Paris

1 Introduction

Many data analysis tasks, such as Quantitative Structure-Activity Relationship (QSAR) or microarray analysis require to deal with a great number of features, among which few are relevant for the considered classification problem. Both for easing the classification task and because features (e.g. gene activity) might be costly to measure, it is of paramount importance to be able to select as early and best as possible the relevant features.

Numerous methods have been proposed in recent years for feature selection (see [1, 3, 5, 7, 8]). They usually rely on measures of the correlation between the feature value and the class of the objects at hand (e.g. class of a microarray) to assess the relevance of each feature, and on educated guesses to set a threshold separating the good candidate features from the others. However, in general, the poor quality of the data together with their scarcity render the whole process quite unreliable.

In this paper, following [2], we study how to combine several such unreliable feature evaluation methods. Indeed, because most feature estimation methods measure the same kind of regularities in the data to determine the relevant features, the rankings they return are more correlated for the relevant features than for the "random" ones. We propose to translate this property through a generative (parametric) model for the correlation of two rankings in order to determine the most likely relevant features.

2 A model of the correlation

In the following, d denotes the number of features, of which p, the unknown to be determined, are supposed to be relevant. Applying two classical feature evaluation methods (e.g. ANOVA [4] and RELIEF [6]) on the available data produces two rankings of the d features from the most promising ones to the less promising. If the two methods were perfectly informed, both would put the p relevant features on top of their ranking followed by the $d - p$ irrelevant ones in no special order. In this case, examining the intersection of the two rankings would immediately give away the relevant features. Let indeed $\text{top}_n(M)$ denote the n features top-ranked by method M, and M_1 and M_2 be the two methods used to rank the features. We would then get $|top_p(M_1) \bigcap top_p(M_2)| = p$ since the top_p-ranked features from the two ranking would contain all p relevant features.

Meanwhile, were the two methods M_1 and M_2 be uncorrelated (except from their perfect information on the relevance of the feature), the size of the intersection of their top_n ranked features when $n \leq p$ would follow an hypergeometric law characterizing the size of the intersection of two subsets randomly drawn from p features[4], whereas, for $n > p$, the intersection size would be equal to p plus the value of an hypergeometrical law describing the intersection size of two subsets of size $n-p$ randomly taken from a set of $d - p$ elements.

In practice, however, the methods are too crude to rank all p relevant features at the top, and, in addition, they may be correlated in the way they rank the features. The generative model predicting the (expected) size k of $|top_n(M_1) \bigcap top_n(M_2)|$ must therefore take into account both the estimated quality of the methods and their correlation.

Ignoring for an instant the correlation between the methods, and assuming that both methods are equally good at selecting the relevant features (that is they both find $m \leq n$ relevant features among their top_n), the expected intersection size k would follow the following law:

$$p(\cap = k | d, p, n, m, \mu_{\mathcal{H}_0}) \;=\; \frac{\binom{p}{m}\binom{d-p}{n-m}\sum_{k^+=2m-p}^{m}\binom{m}{k^+}\binom{p-m}{m-k^+}\binom{n-m}{k-k^+}\binom{d-n-(p-m)}{n-m-(k-k^+)}}{\binom{d}{n}\cdot\binom{d}{n}}$$

[4] The hypergeometric law gives the probability that the intersection size of two randomly drawn subsets of size n in a set of size d be k:

$$\mathcal{H}(d, n, k) \;=\; \frac{\binom{n}{k}\cdot\binom{d-n}{n-k}}{\binom{d}{n}}$$

This expression computes the number of ways one can get an intersection size of k given d, p, n, m divided by the total number of ways one can get two drawings of n features among d. k^+ stands for the part of the intersection size k that correspond to relevant features. (See Figure 1).

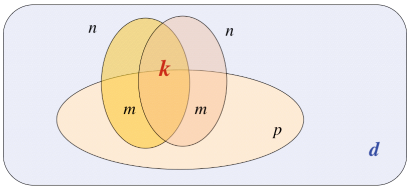

Figure 1. The sets involved in the generative model of the intersection size k between two uncorrelated methods.

It remains to account for the *a priori* correlation between the methods. This can certainly be done in many ways. We have chosen a method where the *a priori* correlation is expressed by the fact that both methods must select their features in a set of size called $d_{corr}(n)$, with $n \leq d_{corr}(n) \leq d$. The lowest is $d_{corr}(n)$, the more tighten is the *a priori* correlation between the methods.

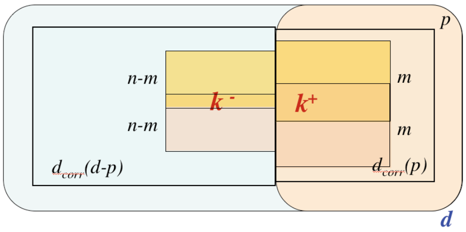

Figure 2. The model for correlated methods.

Taking into account the *a priori* correlation between the methods, the probability of having an intersection of size k for their Top_n is :

$$P(k|d,n,p,m) = \sum_{k_+=0}^{k} \mathcal{H}(d_{corr}(p,m), m, k_+) \cdot \mathcal{H}(d_{corr}(d-p, n-m), n-m, k-k_+)$$

3 Experiments

In order to test the model, artificial data were generated as follows. The learning set contained 20 examples (10 + and 10 -). For each example, the each feature (in a set of $d = 1000$ features) follows a gaussian law of variance σ and of mean 0 if it is a non-relevant feature, or of mean $\pm\delta$ (depending of the classof the example) if it is one of the p relevant features.

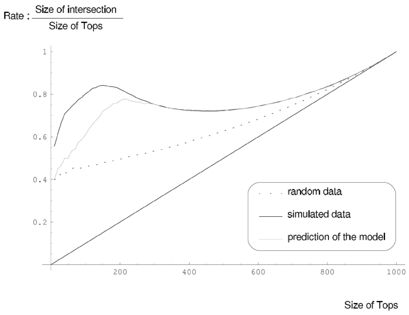

Figure 3. comparaison between mesured and predicted intersection sizes for 1000 artificial data with $p = 200$, $\delta = 1.5$, $\sigma = 1$

Figure 3 shows that even if the model is not good for small values of n, it can predict the intersection size for other values. In fact, the model predicts well when the value of m is closed to p. As in this case the intersection between relevant features is fixed (we take two time $\approx p$ elements in a set of p elements so the intersection size will be $\approx p$) this means that the model correctly predicts the interactions between non-relevant features but not as well those between relevant features. Corrections to the model are under way to put this right.

References

[1] A. Blum and P. Langley, Selection of relevant features and examples in machine learning, *Artificial Intelligence journal*,97:245-271, 1997.

[2] A. Cornuéjols and C. Froidevaux and J. Mary, Comparing and combining feature estimation methods for the analysis of microarray data, *JOBIM-05 : Journées Ouvertes Biologie Informatique Mathématiques*, Lyon, France, 2005.

[3] I. Guyon and A. Elisseeff, An introduction to variable and feature selection, *Journal of Machine Learning Research*,3:1157-1182, 2003.

[4] K. Kerr and M. Martin and G. Churchill, Analysis of variance for gene expression microarray data, *J. of Comp. Biol.*, 7(6):818-837, 2000.

[5] W. Li and I. Grosse, Gene selection criterion for discriminant microarray data analysis based on extreme value distributions, *RECOMB'03*, ACM Press, 2003.

[6] M. Robnik-Sikonja and I. Kononenko, Theoretical and empirical analysis of ReliefF and RReliefF, *Machine Learning Journal*, 53(23), 2003.

[7] V. G. Tusher and R. Tibshirani and G. Chu, Significance analysis of microarrays applied to the ionizong radiation response, *Proceedings of the National Academy of Sciences, USA*, 98(9):5116-5121, 2001.

[8] Y. H. Yang and Y. Xiao and M. Segal, Identifying differentially expressed genes from microarray experiments via statistic synthesis, *Bioinformatics*, 21(7):1084-1093, 2005.