Unsupervised Object Ranking Using Not Even Weak experts

Antoine Cornuéjols and Christine Martin

AgroParisTech, UMR-MIA-518 AgroParisTech - INRA,
16, rue Claude Bernard, F-75231 Paris Cedex 05, France
{antoine.cornuejols,christine.martin}@agroparistech.fr

Abstract. Many problems, like feature selection, involve evaluating objects while ignoring the relevant underlying properties that determine their true value. Generally, an heuristic evaluating device (e.g. filter, wrapper, etc) is then used with no guarantee on the result. We show in this paper how a set of experts (evaluation function of the objects), not even necessarily weakly positively correlated with the unknown ideal expert, can be used to dramatically improve the accuracy of the selection of positive objects, or of the resulting ranking. Experimental results obtained on both synthetic and real data confirm the validity of the approach. General lessons and possible extensions are discussed.

Key words: Ensemble methods, unsupervised learning, feature selection, ranking.

1 Introduction

Imagine you are asked to identify students who have taken a course in some subject. All you have is a set of handouts from a collection of students, some of whom have followed the course of interest. Because you do not know the specifics of the target course, you do not know how to evaluate the handouts in order to single out the students you are looking for. However, you can use the services of a set of colleagues who do not know either, *a priori*, how to recognize the "positive" students, but have the ability to grade the copies using their own set of evaluation criteria. These can be quite diverse. For instance, a grader might be sensitive to the color of the paper, the margins, the thickness of the paper and the number of pages. Another could count the number of underlinings and use the length of the name and the color of the ink to evaluate the student's handouts. And a third one could measure the number of dates cited in the copy, the average length of the sentences, and the average space between lines.

Would that help you in identifying the students who have taken a course, unknown to you? And by combining in some way the various evaluations and rankings of your "experts", could you somehow increase your confidence in detecting the right students?

While grading students using a bunch of unknowledgeable and weak experts might certainly interest many of us, the problem outlined above is just one illustration of a much larger set of applications.

Actually our research on this subject started when, some years ago, a biologist came to us for asking our help in discovering the genes that were involved in the response of cells to weak radioactivity level. The organism studied was the yeast *Saccharomyces cerevisiae* and its 6,135 genes. Thanks to microarray technology, their activity level in conditions where low radiation doses were present and conditions where no such radioactivity existed was measured. All together, 23 microarrays were obtained. In a way, our students here were the genes and their handouts were their profile of responses to the 23 conditions. We did not know what kind of profile was relevant to identify the genes that interested the biologist, but we had some techniques, like SAM [1], ANOVA [2, 3], RELIEF [4, 5] and others, that could be used to rank the genes according to some statistical patterns or regularities found in the profiles of the genes.

This paper shows how, in the absence of supervised information, can we still attain a high precision level in finding the "positive" instances in a large collection, given that we can make use of a varied set of "weak" experts and generate samples according to some null hypothesis.

Our paper is organized as follows. Section 2 provides a more formal definition of the problem and discusses why existing approaches using ensemble techniques, like Rankboost [6] and other similar methods, are inadequate. The bases of our new approach and the algorithm for a single iteration are presented in Section 3. Experimental results, on synthetic data as well as real data, are then reported in Section 4, while Section 5 describes the extension of the method to multiple iterations. Finally, lessons and perspectives are discussed in Section 6.

2 The problem and related works

2.1 Definition of the problem

We can define the problem we are interested in as follows.

- A sample $\mathcal{S}$ of *objects* (e.g. students, genes, features) of which it is strongly suggested that some are "positive" objects, while the others are to be considered as "negative" ones (e.g. students who did not take the target course, genes not sensitive to radioactivity, irrelevant attributes).
- A set $\mathcal{E}$ of "*experts*", also called graders, who, given an object, return a value according to their own set of criteria.

Note that nothing is known beforehand about the alignment of our "experts" with the ideal grader. As far as we know, some experts may tend to rank objects similarly as the target expert, but other may well rank them in a somewhat reverse order, while still others may be completely orthogonal to the target regularities (e.g. it might be expected that the number of letters of the name of the students do not provide any information about the courses they have taken).

In this truly unsupervised setting, is it then possible to use such a set of experts, or a subset thereof, in order to rank the objects in the sample $\mathcal{S}$ with some guarantee about the proximity with the target ranking (one that would put all the positive objects before the negative ones)?

2.2 Related works

In the ranking problem, the goal is to learn an ordering or ranking over a set of objects. One typical application arises in Information Retrieval where one is to prioritize answers or documents to a given query. This can be approached in several ways. The first one is to learn a *utility* or *scoring* function that evaluates objects relative to the query, and can therefore be used to induce an ordering. This can be seen as a kind of ordinal regression task (see [9]). A second approach consists in learning a *preference* function defined over pairs of objects. Based on this preference function, or partial ordering, one can then try to infer a complete ordering that verify all the known local constraints between pairs of objects.

Both of these approaches need training data which, generally, takes the form of pairs of instances labeled by the relation $\preceq$ (*must precede or be at the same rank*) or $\succ$ (*must follow*). In the case of the bipartite ranking problem, there are only two classes of objects, labeled as positive or negative, and the goal is to learn a scoring function that ranks positive instances higher than negative ones [8]. Ensemble methods developed for supervised classification have thus been adapted to this learning problem using a training sample of ordered pairs of instances. Rankboost [6] is a prominent example of these methods.

Another perspective on the ranking problem assumes that orderings on the set of objects, albeit imperfect and/or incomplete, are available. The goal is then to complete, merge or reconcile these rankings in the hope of getting a better combined one. This is in particular at the core of Collaborative Filtering where training data is composed of partial orderings on the set of objects provided by users. Using similarity measures between objects and between users, the learning task amounts then to completing the matrix under some optimization criterion.

In [7], another approach was taken where, starting from rankings supposedly independent and identically corrupted from an ideal ranking, the latter could be approximated with increasing precision by taking the average rank of each object in an increasing number of rankings. However, the underlying assumptions of independence and of corruption centered on the true target ranking were disputable and the experimental results somewhat disappointing.

Above all, all these methods rely on supervised training data, either in the form of labelled pairs of instances, or in the form of partial rankings. In the latter case, these rankings are supposed to be mild alterations of the target ranking.

In this paper, the problem we tackle does not presuppose any learning data. Furthermore, the evaluation functions or "experts" that are considered are not supposed to be positively correlated with the target evaluation function.

3 A new method

3.1 The principle

Let us return to the situation outlined in the introduction where the task is to distinguish the students in a university who have taken a course in some discipline of which you do not know the characteristics. Given our ignorance on

both the expert's expertise for the question at hand, and on the target concept, the situation might seem hopeless. However, the key observation is that the sample at hand is not a random sample: the sample of students very likely includes positive instances (students) to some significant extent (e.g. more than could be due to accidental fluctuations in the background of students). This provides us with a lever.

We consider now *pairs* of experts, and measure their correlation both *a priori*, by averaging on all possible ranking problems, and on the target problem. If the two experts are sensitive to the "signal" in the sample, then their correlation on this sample will differ from their *a priori* one.

This observation is at the basis of the ensemble method we propose. Instead of relying on a combination of separate experts, it uses pairs of experts both to detect experts that are relevant and to assign them a weight. The utterly unsupervised, and seemingly hopeless, task is tamed thanks to the measurements of correlations of higher orders between experts.

3.2 Formal development

Let d be the number of objects in the target sample $\mathcal{S}$ (e.g. the students) and suppose two graders or experts rank the elements of $\mathcal{S}$.

In case the experts were random rankers, the top n elements of each ranking would be equivalent to a random drawing of n elements. Therefore the size k of their intersection would obey the *hypergeometric law*:

$$H(d,n,k) \;=\; \frac{\binom{n}{k}\cdot\binom{d-n}{n-k}}{\binom{d}{n}} \tag{1}$$

where $H(d,n,k)$ denotes the probability of observing an intersection of size k when drawing at random two subsets of n elements from a set of size d.

For instance, in the case of the intersection of two subsets of 500 elements randomly drawn from a set of 6,135 elements, the most likely intersection size is 41. It can be noticed that $k/n = n/d$ (e.g. $41/500 \approx 500/6{,}135$).

In other words, if two totally uncorrelated graders were asked to grade 6,135 copies, and if one looked at the intersection of their 500 top ranked students, one would likely find an intersection of 41. However, two graders using exactly the same evaluation criteria would completely agree on their ranking no matter what. Then the intersection size k would be equal to n for all values of n. The opposite case of two anti-correlated graders would yield two opposite rankings for any sample. The intersection size would therefore be zero up to $n = d/2$ and then grow up as $2(n - d/2)/n$.

There is therefore a whole spectrum of possible correlation degrees between pairs of experts, from 'totally correlated', to 'maximally anti-correlated', going through 'uncorrelated' (case of the hypergeometric law) as shown on Figure 1.

As an illustration, Figure 2 shows the curve obtained for the pair of "experts" ANOVA and RELIEF when they ranked samples of possible genes. It appears that the two methods are positively correlated. The curve stands above

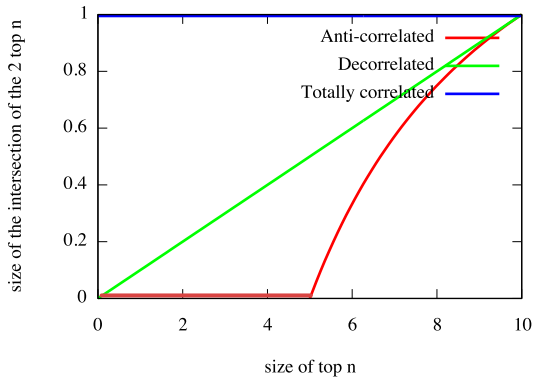

Fig. 1. Illustration of the classes of possible correlation curves between pairs of experts. The x-axis stands for n, the size of the top n ranked elements considered in each ranking. The y-axis represents the size of the intersection $k(n)$. Any curve between the "totally correlated" one and the "anti-correlated" one are possible.

the diagonal) at approximatively two standard deviations above, except for the tail of the curves. This "over-correlation" starts sharply when the intersection is computed on the top few hundreds genes in both rankings, and then it levels off, running approximatively parallel to the *a priori* correlation curve. This suggests that this is in the top of both rankings that the genes present patterns that significantly differ from the patterns that can be observed in the general population of instances. Actually, the fact that the relative difference between the curves is maximal for $n \approx 180$ and $n \approx 540$ would imply that it is best to consider the top_{180} or the top_{540} ranked genes by ANOVA on one hand and by RELIEF on the other hand because they should contain the largest number of genes corresponding to information that is specific to the data.

3.3 Estimating the number of positive instances

Following the computation of the hypergeometric law, one can compute the number of combinations to obtain an intersection of size k when one compares the top n elements of both rankings.

Let us call k_{corr} the number of elements that the two experts tend to take in common in their top n ranked elements, and k^+_{corr} the number of positive elements within this set. Then the probability of observing a combination as described on Figure 3 is:

$$H"(k, k^-_{corr}, k^+_{corr}, p_1, p_2, n, d) = \frac{\overbrace{\binom{n-p_1}{k^-_{corr}}}^{a}\overbrace{\binom{d-p-(n-p_1)}{n-p_2-k^-_{corr}}}^{b}\overbrace{\binom{p_1}{k^+-k^+_{corr}}}^{c}\overbrace{\binom{p-p_1}{p_2-k^+-k^+_{corr}}}^{d}}{\binom{d}{n}} \quad (2)$$

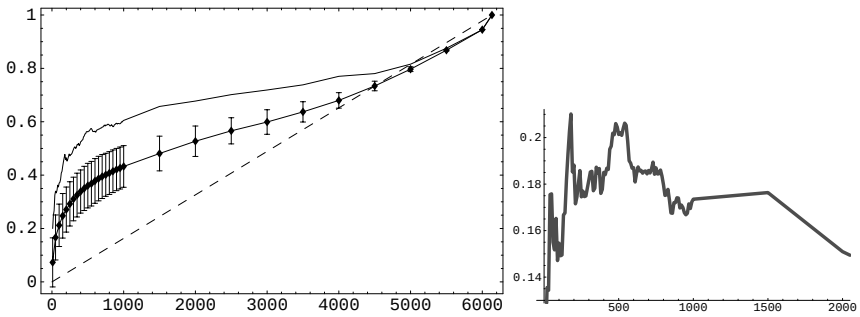

Fig. 2. The x-axis stands for the number n of top-ranked features (e.g. genes) considered. The y-axis stands for the ratio of the intersection size to n. **Left**: (Top curve) the intersection size for the true data. (Center curve): the mean intersection size due to the a priori correlation between ANOVA and RELIEF (with some standard deviation bars). (Lower curve): the intersection size explainable by randomness alone. **Right**: Curve of the relative difference, with respect to n, of the observed intersection size k and the intersection size k_{corr} due to a priori correlation between ANOVA and RELIEF. The curve focuses on the beginning of the curve, for $n < 2,000$, since this is the more interesting part.

where the overbraces refer to the sets indicated in Figure 3.

In this equation, k, k_{corr}, n and d are known. And since k_{corr} is independent on the classes of the objects, one can estimate that $k^+_{corr}/k_{corr} = p/d$. The unknown are therefore: k^+, p_1 and p_2. When there is no *a priori* reason to believe otherwise, one can suppose that the two experts are equally correlated to the ideal expert, which translates into $p_1 = p_2$, and there remains two unknowns only, p_1 and k^+. Using any optimization software, one can then look for the values of p_1 and k^+ which yield the maximum of Equation 2.

In the case of the low radiation doses data, the maximum likelihood principle, applied with $d = 6,135, n = 500, k_{corr} = 180$ and $k = 280$ yields $p = 420 \pm 20$ and $p_1 = 340 \pm 20$ as the most likely numbers of total relevant genes and of the relevant genes among the top_{500} ranked by both methods. From these estimations, a biological interpretation of the tissues of the cell affected by low radioactivity was proposed [10].

3.4 Experimental results on synthetic data

In these experiments, $d = 1,000$ genes, or features, were considered, whose value were probabilistically function of the condition. For each feature, 10 values were measured in condition 1 and 10 values in condition 2. The relevant features were such that condition 1 and condition 2 were associated with two distinct gaussian distributions. The difference δ between the means μ_1 and μ_2 could be varied, as well as the variance. The values of the irrelevant features were drawn from a unique gaussian distribution with a given variance. In the experiments reported here, the number p of relevant features was varied in the range [50, 400], the

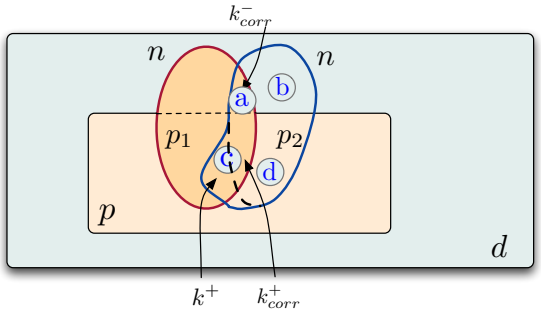

Fig. 3. The sets involved in the generative model of the intersection size k when the correlation *a priori* is taken into account (see Equation 2).

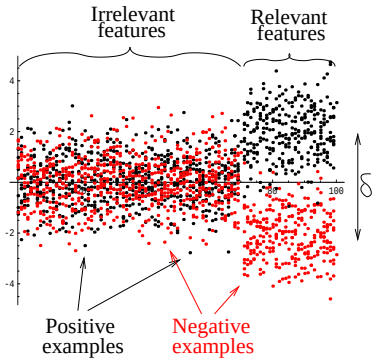

Fig. 4. The values of artificial genes. For each relevant genes, 10 values are drawn from distribution 1, and 10 from distribution 2, while the 20 values for the irrelevant genes are drawn from a unique distribution.

difference $\delta \in [0.1, 5]$ and $\sigma \in [1, 5]$. The smaller δ and the larger σ, the more difficult it is to distinguish the relevant features from the irrelevant ones.

Figure 5 shows the results for $p \in \{50, 200, 400\}$, $\delta = 1$ and $\sigma = 1$ (left), and for $\delta = 2$ (right). Predictively, the curves are much more peaked in the case of a larger signal/noise ratio. Indeed, on the right curves, the value of p can be guessed directly. However, even in the less favorable case (Figure 1, left), the value of p can be retrieved quite accurately using equation 2 and the maximum likelihood principle. After computation, the values $p = 50, 200, 400$ emerge rightly as the more likely ones.

These results obtained in this controlled experiments and others not reported here show the value of considering pairs of experts in this strongly unsupervised context. But is it possible to use more than one pair of experts?

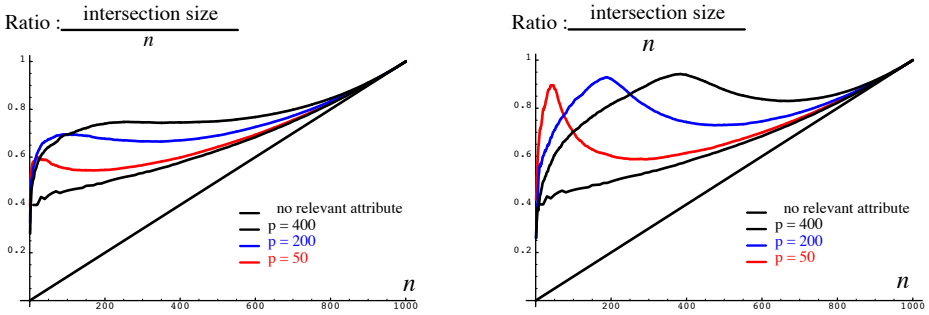

Fig. 5. The correlation curves obtained for various values of p, in a difficult setting (left), and a milder one (right).

4 Towards higher order combinations of experts and rankings

A combination of experts must be weighted according to the value of each expert. But how to proceed when no information is *a priori* available on the value of the various experts for the task at hand?

Measuring the difference in the correlation of pairs of experts on average and on the sample $\mathcal{S}$ as described in Section 3 allows one to make such a distinction. On one hand, experts that are blind to the target regularities won't show any over or under correlation with any other expert on the sample $\mathcal{S}$ since they do not distinguish in any way the positive instances. On the other hand, experts that are not blind and tend either to promote positive instances towards the top (resp. the bottom) of the ranking will show over-correlation on $\mathcal{S}$ with experts that do the same, and under-correlation with experts that do the opposite. Two classes of experts will thus appear, the ones positively correlated with the target ranking and the ones negatively correlated. If, in addition, we assume that the positive instances are a minority in $\mathcal{S}$, it is easy to discriminate the "positive" class form the "negative" one. Because we measure correlation by comparing the top n ranked elements of the experts, the correlation curve rises much more sharply for the positive experts, that put the positive instances towards the top, than for the negative ones that put the rest of the instances towards the top.

Knowing the sign of the weight of each expert is however too crude an information in order to obtain a good combined ranking. We are currently working on a scheme to compute the value of the weight that each expert should be given in order to reflect its alignment with the unknown target evaluation function.

5 Lessons and perspectives

This paper has presented a new totally unsupervised approach for ranking data from experts diversely correlated with the target regularities. One key idea is to

measure and exploit the possible difference between a priori correlation existing in pairs of experts and their correlation on the actual data to be ranked. We have shown how to measure these correlations and how to estimate from them relevant parameters through a maximum a posteriori principle.

The use of pairs of experts in order to overcome the lack of supervised information is, to our knowledge, new. The experimental results obtained so far confirm the practical and theoretical interest of the method. We have also suggested ways to use multiple pairs of experts in a boosting like process. Future work will be devoted to the precise design of such an algorithm and to extensive experimentations. They will include comparisons with other co-learning methods specially designed for unsupervised learning (see for instance [11]).

Acknowledgments. We thank Gilles Bisson for many enlightening discussions, and Romaric Gaudel for helping in finding relevant references to related works and for his participation in the early stages of this study. This work has been supported in part by the ANR-Content Holyrisk project.

References

1. Tusher, V.G., Tibshirani, R. and Chu, G.: Signficance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci.* USA, 98, 5116-5121 (2001)
2. Sahai, H.: *The Analysis of Variance*. Birkhauser, Boston, MA (2000)
3. Pavlidis, P. and Noble, W.: Analysis of strain and regional variation in gene expresion in mouse brain. *Genome Biol.*, 2, 1-14 (2001)
4. Kira, K. and Rendell, L.: A practical approach to feature selection. In Sleeman,D. and Edwards,P. (eds), *Proceedings of the International Conference on Machine Learning*. Morgan Kaufmann, Aberdeen, UK, pp. 249-256 (1992)
5. Kononenko, I.: Estimating attributes: analysis and extensions of RELIEF. *Proc. of the European Conf. on Machine Learning*, ECML-94. pp. 171-182 (1994)
6. Freund, Y. and Iyer, R. and Schapire, R. and Singer, Y.: An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4, 933-969 (2003)
7. Jong K., and Mary J., Cornuéjols A. and Marchiori E. and Sebag M.: Ensemble feature ranking. *Proc. of Principles of Knowledge Discovery in Databases* (PKDD-2004, Pisa, Italy, Sept. 20-24, 2004), Springer-Verlag, LNAI-3202, 267-278 (2004)
8. Agarwal, S. and Cortes, C. and Herbrich, R. (Eds): *Proceedings of the NIPS 2005 Workshop on "Learning to rank"* (2005)
9. Herbrich, R. and Graepel, T. and Obermayer, K.: Large margin rank boundaries for ordinal regression. In Smola, A. and Bartlett, P. and Schökopf, B. and Schuurmans, D. (Eds) *Advances in Large Margin Classifiers*, 115-132. MIT Press (2000)
10. Mercier, G. and Berthault, N. and Mary, J. and Peyre, J. and Antoniadis, A. and Comet, J-P. and Cornuéjols, A. and Froidevaux, Ch. and Dutreix, M.: Biological detection of low radiation by combining results of two analysis methods. *Nucleic Acids Research (NAR)*, Vol.32, No.1, e12 (8 pages), (2004)
11. Grozavu N. and Ghassany M. and Bennani Y.: Learning Confidence Exchange in Collaborative Clustering, in Proc. *IJCNN, IEEE International Joint Conference on Neural Network*, San Jose, California-July 31 - August 5, (2011).