Unsupervised one class identification by selecting and combining ranking functions

Antoine Cornuéjols[1] et Christine Martin[1]

[1]AgroParisTech, département MMIP et INRA UMR-518, 16, rue Claude Bernard , F-75231
Paris Cedex 5 (France)

January 5, 2015

Abstract

We study the problem of identifying a class of interest in an unsupervised data set. Assuming that a set $\mathcal{F}$ of score functions is available, of unknown performance for the task at hand, we propose a method in order to *select useful functions* from the set. Each of these functions induces a ranking over the data set.

We then show how to *combine the base rankings* thus obtained. Experimental results demonstrate that the combined performance is almost as good, or better, than the performance of the best, but unknown, score function in $\mathcal{F}$. In addition, we show, under some simplifying assumptions, how a proper combination of the base rankings allows one to end up with DNF formulas involving the selected score functions that converge to optimal precision and recall with respect to the target concept, if the capacity of $\mathcal{F}$ permits it. Such formulas, easily interpretable, are very desirable in the exploratory context of data mining.

Mots-clef: Unsupervised learning, Ensemble methods.

1 Introduction

Data exploration aimed at discovering interesting classes of patterns is an essential part of scientific discovery or, more mundanely, of data mining. For instance, in bioinformatics, many research works look for the identification of genes that respond to some conditions in the environment, or for finding proteins that could potentially interact with some given target drugs. In a different context, the IRS (Internal Revenue Service) would like to identify the most likely tax evaders. More generally, fraud detection is a growing application area. In each case, there is one class of interest that gathers objects the expert is looking for against the other data points.

In this exploratory setting, it is difficult to come up with informative functions good at distinguishing between the interesting data points versus the non interesting ones. While it might be easy to get candidate evaluation functions from experts or from libraries of functions commonly used in statistics or in Machine Learning, or even to generate such functions automatically, it is difficult in an unsupervised context to assess their merit. Therefore one is left guessing which one(s) of these functions to rely on. Additionally, for many application domains, and especially those where data is described by a large number of features, it is highly desirable that the class of interest be described in an interpretable way. This means that the class of interest should be expressed as much as possible using understandable features. For most experts, understanding and the capacity for reasoning imply descriptions that use combinations of predicates like disjunctive normal forms (DNF). This allows him/her to gain insight in what makes the class of interest apart and how this can be related to the current domain theory, possibly stimulating some revision of the theory.

In this work, we study the following problem. We suppose that there exists a set $\mathcal{S}$ of m data points from the input space $\mathcal{X}$ with no labels: $\mathcal{S} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ that has been generated by an unknown mixture of distributions of which some components, belonging to $\mathsf{P}^+_{\mathcal{X}}$, correspond to the class of interest that we call $\mathcal{S}^+$, and the other components, $\mathsf{P}^-_{\mathcal{X}}$, correspond to the set of the remaining data points $\mathcal{S}^-$. The sets $\mathcal{S}^+$ and $\mathcal{S}^-$, such that $\mathcal{S}^+ \cup \mathcal{S}^- = \mathcal{S}$, are unknown and must be identified as well as possible.

In addition, we suppose that a set $\mathcal{F}$ of evaluation functions (or score functions) is available, each function associating a score to a data point: $f_i : \mathcal{X} \rightarrow \mathbb{R}$. Nothing is assumed *a priori* about the usefulness of each function $f_i \in \mathcal{F}$, and in particular, one does not know if any given function is "aligned" with the target concept, that is if it tends to put the data points of the class of interest toward the top of the induced ranking over the data set $\mathcal{S}$.

We propose a method for identifying useful score functions in $\mathcal{F}$, if some exist, in this completely unsupervised setting. The basic idea is to look at the correlation between the rankings induced by the score functions over $\mathcal{S}$ and to select functions with a particular property. We explain how one can use the base rankings in order to get a combined ranking of the data points in $\mathcal{S}$ with good performances.

We end up by demonstrating, under some simplifying assumptions, how a proper combination of the base rankings allows one to end up with DNF formulas that converges to optimal precision and recall with respect to the target concept, if the capacity of $\mathcal{F}$ permits it.

2 The selection of useful evaluation functions

2.1 Principle of the method

In supervised learning ...

2.2 Correlation measures

A measure of correlation between rankings ...

In the proposed method, ...

Figure **??** depicts a typical difference. Here the evaluation functions are ANOVA and RELIEF [Kon94] and the data corresponds to 6,400 genes. The task was to find out if some genes were sensitive to low radioactivity levels. The upper curve $|\cap_n^{i,j}|$ shows the correlation over the data, while the lower curve with confidence intervals is obtained by computing the intersections $|\cap_n^{i,j}|$ over random samples $\mathcal{S}_0$ (here 100).

The difference in ...

2.3 A theoretical analysis

In this section, we develop a simple model in order to allow us (in Section 5) to devise a strategy for discovering interpretable expressions of the hidden regularities in the data.

We start by assuming that the evaluation functions are characterized by a positive (or negative) propensity

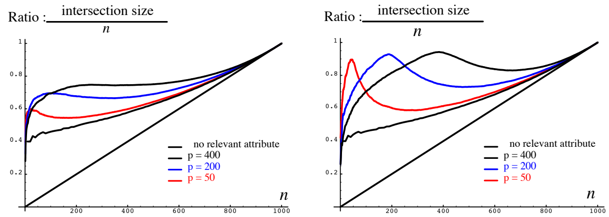

Figure 2: Correlation curves between rankings of an artificial data set of 1,000 elements for various numbers of elements of class '+', here 50, 200 and 400. The peaks are accentuated on the right graph which corresponds to an easier problem.

to put the elements of class '+' at the top of their ranking. This propensity can be modeled by a ROC curve [Fla12].

(...)

In the simple analysis reported here, we suppose that we consider two evaluation functions f_i and f_j of the same strength (defined by ε_x and ε_y), that is they share a common ROC curve. The theoretical study with functions exhibiting different ROC curves does not change qualitatively the results.

Let us compute the size of the intersection of the top$_n$ elements: $|\cap_n^{i,j}|$. Let x be the number of false positive elements. Therefore, x varies on the FP axis. Let m^+ be the number of positive elements in $\mathcal{S}$ and m^- be the number of negative elements. Then, we have two phases to consider.

1. 1st phase: $x \leq \varepsilon_x$. One finds:

$$\left\{ \begin{array}{lcl} n & = & x\, m^- + \frac{1-\varepsilon_y}{\varepsilon_x}\, x\, m^+ \\ & & \\ |\cap_n^{i,j}| & = & x^2\, m^- \ + \left(\frac{1-\varepsilon_y}{\varepsilon_x}\right)^2 x^2\, m^+ \end{array} \right. \tag{1}$$

giving, for the first part of the curve, the equation:

$$\begin{aligned} \frac{|\cap_n^{i,j}|}{n} &= \frac{x^2\, m^- \ + \left(\frac{1-\varepsilon_y}{\varepsilon_x}\right)^2 x^2\, m^+}{x\, m^- + \frac{1-\varepsilon_y}{\varepsilon_x}\, x\, m^+} \\ &= x\, \frac{m^- \ + \left(\frac{1-\varepsilon_y}{\varepsilon_x}\right)^2 m^+}{m^- + \frac{1-\varepsilon_y}{\varepsilon_x}\, m^+} \end{aligned} \tag{2}$$

For the special value $x = \varepsilon_x$ (point P), we get:

$$\left\{ \begin{array}{lcl} n & = & \varepsilon_x\, m^- + (1-\varepsilon_y)\, m^+ \\ & & \\ |\cap_n^{i,j}| & = & \varepsilon_x^2\, m^- \ + \ (1-\varepsilon_y)^2\, m^+ \end{array} \right. \tag{3}$$

corresponding to the value on the y-axis:

$$\frac{|\cap_n^{i,j}|}{n} = \frac{\varepsilon_x^2\, m^- \ + \ (1-\varepsilon_y)^2\, m^+}{\varepsilon_x\, m^- + (1-\varepsilon_y)\, m^+} \tag{4}$$

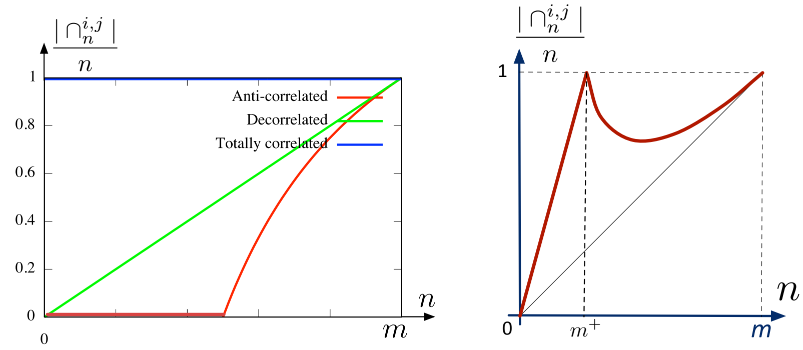

Figure 1: The curve $|\cap_n^{i,j}|/n$ function of n. Two independent draws should approximately result in the diagonal law. (Left) Two maximally correlated draws give $|\cap_n^{i,j}|/n = 1 \ (\forall n)$. Two draws maximally inversely correlated give the red curve at the bottom. All possible behaviors fall between these two extreme curves. (Right) The characteristic curve for two rankings from uncorrelated but perfectly informed evaluation functions.

2. 2nd phase: $\varepsilon_x < x$.

$$\begin{cases} n &= x\,m^- + \left[(1-\varepsilon_y) + \frac{\varepsilon_y}{1-\varepsilon_x}(x-\varepsilon_x)\right] m^+ \\ |\cap_n^{i,j}| &= x^2\,m^- \ + \ \left[(1-\varepsilon_y) + \frac{\varepsilon_y}{1-\varepsilon_x}(x-\varepsilon_x)\right]^2 m^+ \end{cases} \tag{5}$$

giving, for the second part of the curve, the equation:

$$\frac{|\cap_n^{i,j}|}{n} = \frac{x^2\,m^- \ + \ \left[(1-\varepsilon_y) + \frac{\varepsilon_y}{1-\varepsilon_x}(x-\varepsilon_x)\right]^2 m^+}{x\,m^- + \left[(1-\varepsilon_y) + \frac{\varepsilon_y}{1-\varepsilon_x}(x-\varepsilon_x)\right] m^+} \tag{6}$$

These equations give the most probable value for $\frac{|\cap_n^{i,j}|}{n}$, as shown on the right hand side of Figure **??**. While computed from an idealized model, this curve is in good accordance with empirical observations.

2.4 The algorithm

The selection of the useful base scoring functions is done according to algorithm1. ...

Algorithm 1: Selection of "good enough" base scoring functions

```
Input: The data set $\mathcal{S}$
       The set $\mathcal{F}$ of the base scoring functions
Output: A subset $\mathcal{F}'' \in \mathcal{F}$ of base functions

Generation of $N$ random samples $\mathcal{S}_0$;

forall the pairs of scoring functions $(f_i, f_j)_{(i \neq j)}$ $\in \mathcal{F}$ do
    compute the over-correlation of $(f_i, f_j)$ on $\mathcal{S}$ compared to the mean correlation on the samples $\mathcal{S}_0$
end forall

Select the scoring functions $f_i \in \mathcal{F}$ with over-correlation $\geq \tau_{\text{min\_overcor}}$: producing $\mathcal{F}'$

Initialization : $\mathcal{F}'' = \emptyset$

forall the $f_i \in \mathcal{F}'$ do
    if $\sum_{j \neq i} overcorr(f_i, f_j) \geq \tau_{sum\_overcorr}$ then
        Put $f_i$ in $\mathcal{F}''$
end forall
```

3 Experimental studies

These experiments address the question as to which extent the proposed method is able to select relevant evaluation functions in $\mathcal{F}$, ...

In order to test for this, we have realized experiments with artificial data. The data where generated using two probability distributions over the input space $\mathbb{R}^d$ (here $d = 20$): distribution $\mathbf{P}^+_{\mathcal{X}}$ for the '+' instances and distribution $\mathbf{P}^-_{\mathcal{X}}$ for the '+' instances. In the experiments reported here we have used two Gaussian distributions with means separated by a euclidian distance of 3. The difficulty of the task was controlled by adding noise of varying standard deviation σ to the data points ($\sigma = 1.5$, 2.5, 3.5 and 4.5).

(...)

Table 1 reports the minimal AUC (auc_m) and the maximum AUC (auc^M) for the functions in $\mathcal{F}$. Likewise, it reports the minimal AUC (auc_m), the maximal AUC (auc^M) and the mean AUC ($\overline{auc}$) for the functions selected by the method: in $\mathcal{F}''$. Finally, the last column gives the AUC obtained by combining the results of the evaluation functions selected in $\mathcal{F}''$ (see Section 4 for an explanation).

(...)

4 A method for combining results

(...)

5 Towards interpretable combinations of selected features

Assuming that there exists a class of m^+ objects of interest from a distribution $\mathbf{P}^+_{\mathcal{X}}$ and a class of m^- other objects in the data set $\mathcal{S}$ from a distribution $\mathbf{P}^-_{\mathcal{X}}$, is there any hope of identifying the objects of the class '+'? It all depends on the number and properties of the evaluation functions contained in $\mathcal{F}$.

(...)

One can compute the ROC curve obtained when considering the intersection $\frac{|\cap_n^{i,j}|}{n}$ of the top$_n$ of each function.

Using the equations of Section 2.3, one obtains:

For $x \leq \varepsilon_x$:

$$|\cap_n^{i,j}| \;=\; \underbrace{x^2 m^-}_{FP} \;+\; \underbrace{\left[\frac{1-\varepsilon_y}{\varepsilon_x}\right]^2 x^2 m^+}_{TP} \tag{7}$$

and for $x > \varepsilon_x$:

$$|\cap_n^{i,j}| \;=\; \underbrace{x^2 m^-}_{FP} \;+\; \underbrace{\left[(1-\varepsilon_y) + \frac{\varepsilon_y}{1-\varepsilon_x}(x-\varepsilon_x)\right]^2 m^+}_{TP} \tag{8}$$

(...)

What is interesting is that while the AUC of the function $\frac{|\cap_n^{i,j}|}{n}$ is not much larger than the AUC of each base function, its slope in the left part of the curve is much steeper. That means that the precision in this part seems very much improved. Does the theoretical analysis confirms this? Let us see how the precision and recall evolve when one goes from a random selection of objects in $\mathcal{S}$ (stage 0), to using the base score function (stage 1), up to using the function $|\cap_n^{i,j}|$ (stage 2).

1. *Stage 0.* We suppose that a fraction η of the m objects are randomly selected in $\mathcal{S}$ and are assigned to the class '+'. We let: $m^- = \alpha\, m^+$, with $\alpha \geq 0$ and $\varepsilon_x = \beta\,(1-\varepsilon_y)$ with $0 \leq \beta < 1$ (note that $0 \leq \beta < 1$ entails an AUC > 0.5 while $\beta > 1$ entails an AUC < 0.5). Then, we get the precision (*prec.*) and recall:

$$\begin{aligned}
\text{prec.} &= \frac{\text{TP}}{\text{TP}+\text{FP}} = \frac{\eta\, m^+}{\eta(m^+ + m^-)} = \frac{1}{1+\alpha} \\
\text{recall} &= \frac{\text{TP}}{\text{TP}+\text{FN}} = \frac{\eta\, m^+}{m^+} = \eta
\end{aligned}$$

2. *Stage 1.* We look at the point on the ROC curve that maximizes precision and recall: $x = \varepsilon_x$ on Figure **??**.

$$\begin{aligned}
\text{prec.} &= \frac{(1-\varepsilon_y)\, m^+}{(1-\varepsilon_y)\, m^+ + \varepsilon_x\, \alpha\, m^+} \\
&= \frac{1-\varepsilon_y}{1-\varepsilon_y+\alpha\,\beta\,(1-\varepsilon_y)} = \frac{1}{1+\alpha\,\beta} \\
\text{recall} &= \frac{(1-\varepsilon_y)\, m^+}{m^+} = 1-\varepsilon_y
\end{aligned}$$

3. *Stage 2.* We now use the function $|\cap_n^{i,j}|$, again at the point with best precision and recall.

$$\begin{aligned}
\text{prec.} &= \frac{(1-\varepsilon_y)^2\, m^+}{(1-\varepsilon_y)^2\, m^+ + {\varepsilon_x}^2\, \alpha\, m^+} \\
&= \frac{(1-\varepsilon_y)^2}{(1-\varepsilon_y)^2+\alpha\,\beta^2\,(1-\varepsilon_y)^2} = \frac{1}{1+\alpha\,\beta^2} \\
\text{recall} &= \frac{(1-\varepsilon_y)^2\, m^+}{m^+} = (1-\varepsilon_y)^2
\end{aligned}$$

It is apparent that at each stage one looses on the recall, meaning that a smaller part of the class '+' gets recognized. At the same time, ...

$$\begin{aligned}
\text{prec.} &= \frac{(1-\varepsilon_y)^k\, m^+}{(1-\varepsilon_y)^k\, m^+ + {\varepsilon_x}^k\, \alpha\, m^+} &&= \frac{1}{1+\alpha\,\beta^k} \\
\text{recall} &= \frac{(1-\varepsilon_y)^k\, m^+}{m^+} &&= (1-\varepsilon_y)^k
\end{aligned}$$

(...)

The method is therefore in principle able to "invent" new predicates and to produces expressions, DNF, that are conducive to easier interpretation.

However, ...

σ	$\frac{m^+}{m}$	Before selection		After selection			
		auc_m	auc^M	auc_m	auc^M	$\overline{auc}$	AUC comb
1.5	$\frac{40}{320}$	0 ± 0	1 ± 0	$\mathbf{0.92} \pm 0.03$	1 ± 0	0.98 ± 0.01	$\mathbf{1} \pm 0$
	$\frac{80}{320}$	0 ± 0	1 ± 0	$\mathbf{0.87} \pm 0.06$	1 ± 0	0.97 ± 0.01	$\mathbf{1} \pm 0$
	$\frac{120}{320}$	0 ± 0	1 ± 0	$\mathbf{0.84} \pm 0.07$	1 ± 0	0.95 ± 0.01	$\mathbf{1} \pm 0$
2.5	$\frac{40}{320}$	0.02 ± 0.01	0.98 ± 0.01	$\mathbf{0.94} \pm 0.03$	0.98 ± 0.00	0.96 ± 0.02	$\mathbf{0.98} \pm 0.01$
	$\frac{80}{320}$	0.03 ± 0.01	0.98 ± 0.01	$\mathbf{0.85} \pm 0.05$	0.98 ± 0.01	0.91 ± 0.02	$\mathbf{0.97} \pm 0.01$
	$\frac{120}{320}$	0.03 ± 0.01	0.98 ± 0.01	$\mathbf{0.76} \pm 0.03$	0.98 ± 0.01	0.88 ± 0.02	$\mathbf{0.97} \pm 0.01$
	$\frac{160}{320}$	0.03 ± 0.01	0.98 ± 0.01	$\mathbf{0.73} \pm 0.04$	0.97 ± 0.01	0.85 ± 0.02	$\mathbf{0.95} \pm 0.01$
3.5	$\frac{40}{320}$	0.09 ± 0.02	0.91 ± 0.02	$\mathbf{0.75} \pm 0.06$	0.90 ± 0.03	0.83 ± 0.01	$\mathbf{0.90} \pm 0.03$
	$\frac{80}{320}$	0.09 ± 0.02	0.92 ± 0.02	$\mathbf{0.65} \pm 0.05$	0.92 ± 0.02	0.79 ± 0.02	$\mathbf{0.90} \pm 0.02$
	$\frac{120}{320}$	0.09 ± 0.02	0.91 ± 0.01	$\mathbf{0.64} \pm 0.04$	0.91 ± 0.01	0.77 ± 0.02	$\mathbf{0.89} \pm 0.02$
	$\frac{160}{320}$	0.10 ± 0.01	0.91 ± 0.02	$\mathbf{0.63} \pm 0.03$	0.91 ± 0.02	0.76 ± 0.02	$\mathbf{0.88} \pm 0.02$
4.5	$\frac{40}{320}$	0.13 ± 0.02	0.86 ± 0.02	$\mathbf{0.67} \pm 0.03$	0.86 ± 0.02	0.76 ± 0.02	$\mathbf{0.86} \pm 0.02$
	$\frac{80}{320}$	0.15 ± 0.02	0.85 ± 0.02	$\mathbf{0.65} \pm 0.03$	0.84 ± 0.03	0.75 ± 0.02	$\mathbf{0.84} \pm 0.03$
	$\frac{120}{320}$	0.15 ± 0.02	0.84 ± 0.02	$\mathbf{0.62} \pm 0.06$	0.84 ± 0.02	0.73 ± 0.03	$\mathbf{0.84} \pm 0.02$
	$\frac{160}{320}$	0.15 ± 0.01	0.85 ± 0.01	$\mathbf{0.61} \pm 0.03$	0.85 ± 0.01	0.72 ± 0.02	$\mathbf{0.83} \pm 0.03$

Table 1: Experimental results in function of the noise parameter σ and the proportion of the class '+'.

6 Related works

Ensemble methods have first been studied in the context of supervised learning (see [SF12, Zho12] for comprehensive studies). It is indeed . . .

7 Conclusion and future works

In this paper, we addressed . . .

Acknowledgments. Part of this work has been supported by the French ANR project "Coclico" (2013-2016).

References

[Fla12] Peter Flach. *Machine learning: the art and science of algorithms that make sense of data.* Cambridge University Press, 2012.

[Kon94] Igor Kononenko. Estimating attributes: analysis and extensions of relief. In *Machine Learning: ECML-94*, pages 171–182. Springer, 1994.

[SF12] Robert E Schapire and Yoav Freund. *Boosting: Foundations and Algorithms.* MIT Press, 2012.

[Zho12] Zhi-Hua Zhou. *Ensemble methods: foundations and algorithms.* CRC Press, 2012.