Détection non supervisée d'une sous-population par méthode d'ensemble et changement de représentation itératif

Christine Martin, Antoine Cornuéjols

AgroParisTech, département MMIP et INRA UMR-518
16, rue Claude Bernard
F-75231 Paris Cedex 5 (France)
christine.martin , antoine.cornuejols@agroparistech.fr,
http://www.agroparistech.fr/mia/equipes:membres:page:christine

Résumé. L'apprentissage non supervisé a classiquement pour objectif la détection de sous-populations homogènes (classes) considérées de manière équivalente sans information *a priori* sur celles-ci. Le problème étudié dans cet article est quelque peu distinct. On se focalise ici uniquement sur une sous-population d'intérêt que l'on cherche à identifier avec un rappel et une précision optimales.

Nous proposons, pour cela, une méthode s'appuyant sur les principes suivants : (1) travailler dans l'espace de représentation fourni par des experts faibles pour cette tâche, (2) confronter ces experts pour détecter des seuils de sélection plus pertinents, et (3) les combiner itérativement afin de converger vers l'expert idéal. Cette méthode est éprouvée et comparée sur des données synthétiques.

1 Introduction

De nombreuses tâches de fouille de données ou d'apprentissage impliquent la détection d'un sous-ensemble inconnu d'éléments dans une collection d'éléments non étiquetés.

Classiquement des méthodes de regroupement ou *clustering* sont utilisées dans l'espoir de faire apparaître la classe d'intérêt. Cependant les résultats sont généralement très sensibles à la technique utilisée et aux paramètres choisis. Une autre approche est d'utiliser une *méthode de filtre* (Sahai (2000) ; Kira et Rendell (1992)) afin d'évaluer la qualité des objets en fonction d'un critère supposé pertinent et d'ordonner ainsi les objets en qualité décroissante puis de déterminer un seuil dans ce classement permettant de distinguer les deux classes. Malheureusement, ces méthodes donnent des résultats très variables en fonction, d'une part, de leur adéquation avec les régularités cibles inconnues, et, d'autre part, des seuils de sélection choisis.

Nous proposons de circonvenir ces difficultés en utilisant une méthode d'ensemble basée sur des méthodes de filtres nommées ici *« experts faibles »*. A l'image des méthodes d'ensemble proposées en apprentissage supervisé, telles que le boosting (Freund (1995)), nous combinons des méthodes de filtres, pour converger vers la méthode idéale, c'est-à-dire permettant une détection aisée des objets considérés.

Le problème étudié ici s'apparente donc à celui du tri (*ranking* en anglais) d'un ensemble d'objets. Pour cela, des approches par apprentissage d'une fonction de score évaluant chaque

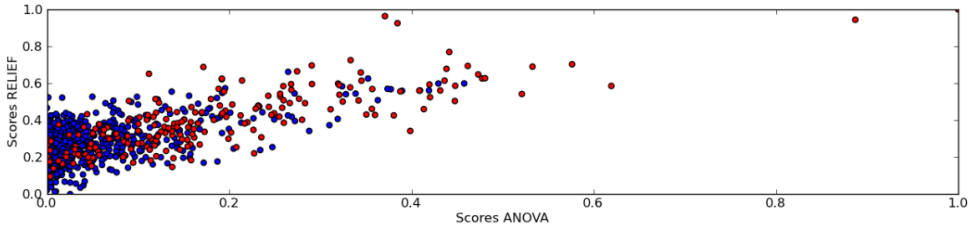

FIG. 1 – *Représentation des objets de $\mathcal{S}$ dans le plan des valeurs des experts ANOVA et RELIEF. La pente du nuage de points reflète la corrélation entre les deux experts.*

objet (voir Herbrich et al. (2000)), d'une part, ou d'une fonction de préférence définie sur les paires d'objets, d'autre part, ont été proposées. Dans le cas du problème du tri bipartie, il y a seulement deux classes d'objets (étiquetés '+' ou '−'), et le but est d'apprendre une fonction de score rangeant les objets positifs avant les objets négatifs. Des méthodes d'ensemble développées pour la classification supervisée ont été adaptées pour traiter ce problème en utilisant un échantillon d'apprentissage de paires ordonnées d'exemples (Freund et al. (2004)). Un autre point de vue sur ce type de problèmes consiste à supposer que des tris sont disponibles, éventuellement imparfaits, partiels ou bruités par rapport au tri idéal (Jong et al. (2004)). Le but est alors de compléter, fusionner ou réconcilier ces tris dans l'espoir de parvenir à un meilleur tri. Cependant, les hypothèses nécessaires à de telles approches sont fortes et discutables.

Il faut retenir que toutes ces méthodes reposent sur la connaissance d'un échantillon d'apprentissage. Dans ce papier, on suppose qu'il n'en existe pas et que la valeur *a priori* des différents experts disponibles est inconnue.

Le principe de l'approche ainsi que l'algorithme proposés sont décrits en section 2. Ce papier est centré sur une étude expérimentale des propriétés de la méthode servant de preuve de principe du concept. Le protocole expérimental suivi est décrit en section 3. Les résultats expérimentaux obtenus font quant à eux l'objet de la section 4. Enfin, les variations possibles autour de l'approche générique et les directions de recherche à envisager forment la section 5.

2 Le principe de l'approche

Un expert est dit *« faible »* si sa fonction d'évaluation distingue, en probabilité, les objets de la classe '+' et ceux de la classe '−'. Plus formellement, un expert est *« faible »* si le coefficient de Spearman de son tri avec le tri idéal est > 0. Ce n'est pas vérifiable a priori et est supposé vrai par défaut. Deux experts faibles présentent un certain degré de corrélation entre eux qui peut être visuellement apprécié en représentant chaque objet dans le plan de leurs évaluations (voir figure 1). Deux experts totalement corrélés dans leurs évaluations donneront un nuage de points aligné selon une droite, tandis qu'un nuage de points répartis uniformément dans un rectangle trahira deux experts totalement décorrélés.

Nous avons vu que l'utilisation d'un expert seul, se heurte à deux difficultés : adéquation de l'expert aux données, et détermination d'un seuil de décision pertinent c'est-à-dire permettant de décider avec assurance que les objets mieux évalués que ce seuil sont les objets recherchés.

En pratique, l'ensemble des valeurs retournées par l'expert sur la collection d'objets ne permet généralement pas de distinguer les classes (voir par exemple la figure 1 obtenue avec ANOVA sur des données synthétiques décrites plus loin). Si l'utilisation d'un expert seul est problématique, nous montrons dans cet article qu'en les combinant il est possible d'améliorer leurs performances.

Ainsi, s'il est difficile de trouver un critère de décision pour distinguer les objets '+' des objets '$-$' à partir des évaluations d'un seul expert, cela devient plus aisé dans le plan des évaluations de deux experts faibles, présentant de fait un certain degré de corrélation. Les nuages de points, l'un associé aux points '$-$', l'autre aux points '+', ont tendance à se différencier davantage. Cela est généralisable à l'usage de N experts. L'étiquette des objets étant inconnue, il est possible dans cet espace d'avoir recours à une méthode de catégorisation ou clustering pour découvrir ces deux nuages de points.

On pourrait alors s'arrêter là et utiliser la frontière de décision ainsi obtenue dans le plan des évaluations pour étiqueter les objets de $\mathcal{S}$. Nous confrontons d'ailleurs nos résultats à ceux de cette approche dans la section 4. Néanmoins, il est possible d'utiliser cette approche menée par paires dans un processus combinant plusieurs experts de manière itérative convergeant vers un expert cible et modifiant l'effet des experts sur les évaluations en fonction de leur comportement en discrimination des nuages de points. Nous proposons de réaliser ce biais en appliquant une transformation non linéaire aux experts afin d'amplifier la différence entre les évaluations des points détectés comme '+' et ceux détectés comme '$-$', et en conservant approximativement les évaluations des points incertains. On obtient alors un expert combiné dont la fonction d'évaluation tend à concentrer les évaluations sur les deux valeurs extrêmes possibles, minimale et maximale (voir figure 3). Chaque étape combine alors l'expert combiné courant avec un nouvel expert faible jusqu'à ce qu'il n'y ait plus d'expert disponible. L'objectif est de tendre vers l'expert idéal qui évalue les objets '+' à 1 et les autres à 0.

Notre approche est donc fondée sur trois idées principales : utiliser la représentation des objets dans l'espace des évaluations retournées par des paires d'experts ; déterminer, dans ce cadre, une transformation non linéaire visant une meilleure séparation des classes pour chaque expert considéré ; approcher l'expert idéal par combinaison itérative des experts.

3 Protocole expérimental

Description des données synthétiques et des paramètres de l'étude Afin d'évaluer la méthode présentée, nous avons utilisé des données synthétiques dont nous contrôlons les propriétés ce qui est nécessaire en apprentissage non supervisé. L'échantillon étudié contient m objets (m = 1000 dans nos expérimentations). Ces objets sont caractérisés par d mesures (d= 40 ici). Le nombre d'objets positifs p est minoritaire dans les m objets (p=200 dans nos expériences).

Par ailleurs, pour les objets '+', $d/2$ mesures sont distribuées selon une loi normale de moyenne μ_0 et d'écart-type σ, soit $\mathcal{N}(\mu_0, \sigma)$, tandis que les $d/2$ autres mesures sont distribuées selon une loi normale $\mathcal{N}(\mu_1, \sigma)$. Les objets '$-$' ont, quant à eux, leurs d mesures distribuées selon la loi normale $\mathcal{N}(\mu_1, \sigma)$.

Dans nos expériences $\mu_0 = 2$, $\mu_1 = 0$ et σ prend sa valeur dans $\{0, 0.5, 1.0, 1.5, \ldots, 4.5\}$ ce qui permet de contrôler la distinction entre les objets positifs (dont certains descripteurs présentent une distribution particulière) et les objets négatifs. Plus σ est grand, plus la distinction est difficile.

L'ensemble des experts faibles Idéalement, il faudrait disposer d'une base d'experts faibles à la fois diversifiée et peu redondante. Dans la pratique, nous avons, pour cette première étude, utilisé les experts décrits ci-dessous.

- L'analyse de la variance (ANOVA) (Sahai (2000)) permet de mesurer la corrélation entre une variable continue à expliquer et une ou plusieurs variables explicatives catégorielle.
- RELIEF (Kira et Rendell (1992)), semblable à ANOVA, est une méthode non paramétrique, pouvant être préférable à ANOVA lorsque les groupes ne suivent pas des distributions normales.
- L'expert faible nommé SEUILS_OPTI a été développé pour cette étude. Il utilise l'information sur les expériences '+' et attribue à chaque exemple la valeur minimale observée pour celui-ci sur les attributs correspondants.

Comparaison des experts. Dans notre cadre, un expert est meilleur qu'un autre s'il permet de mieux distinguer les objets '+' des '$-$'. Idéalement, les scores qu'il attribue aux objets devraient permettre à une méthode non supervisée de déterminer sans erreurs les deux classes. Ce principe est à la base de notre mesure de performance. La méthode k-moyenne a été utilisée en fixant le nombre de classes à 2 et en utilisant une initialisation basée sur une classification hiérarchique ascendante afin de maximiser les performances du clustering.

La classe des données synthétiques étant connue, nous pouvons évaluer le nombre d'individus bien classés par l'algorithme des k-moyennes. Les nombres de vrais positifs (notés TP), de vrais négatifs (notés TN), de faux positifs (noté FP) et de faux négatifs (noté FN) sont ainsi calculés. Afin d'obtenir une mesure unique qui agrège le rappel ($TP/(TP+FN)$) et la précision ($TP/(TP+FP)$), nous avons comparé les experts en fonction de la F-mesure obtenue ($2 \cdot rappel \cdot precision/(rappel + precision)$).

Les experts cités précédemment et leurs combinaisons sont comparés par cette méthode. Une dernière comparaison avec la méthode de recherche directe des objets '+' et des objets '$-$' dans l'espace des experts disponibles permet de mesurer si les transformations non linéaires et le processus de combinaison itératif permettent de gagner en performance par rapport à une méthode ne reposant que sur le changement de représentation.

4 Résultats

L'écart-type σ des distributions des deux conditions expérimentales associées respectivement aux 20 premières et aux 20 dernières colonnes joue le rôle de paramètre de contrôle car il détermine la difficulté à distinguer les objets '+' des '$-$'. Nous avons fait varier cette valeur systématiquement de 0.5 (tâche très simple) à 4.5 (tâche difficile) (*cf.* figure 2).

La série d'expériences réalisées (100 fois pour chaque valeur) montre la supériorité de la méthode proposée sur tous les experts de base (ANOVA, RELIEF et Expert_Seuil) ainsi que sur la méthode opérant par k-moyenne directement dans l'espace des évaluations des experts de base (voir Tableau 1). En particulier, les performances mesurées sur les tâches difficiles chutent nettement moins rapidement que les celles des autres méthodes.

Par ailleurs, ces premières expérimentations permettent de vérifier que les distributions de scores tendent effectivement vers des distributions bipolaires ce qui est souhaité pour approcher l'expert idéal (cf. figures 3).

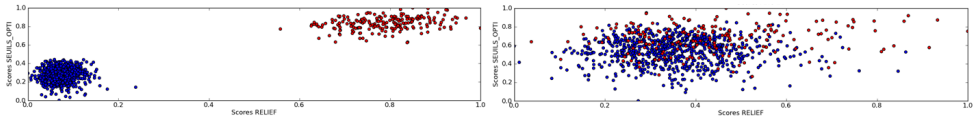

FIG. 2 – *Vue du nuage de point généré dans la représentation des experts utilisés et pour une valeur de σ faible à gauche et une valeur de $sigma$ forte à droite.*

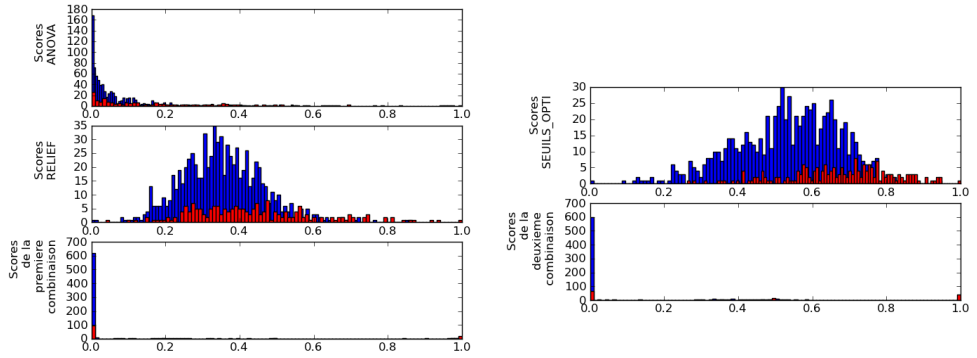

FIG. 3 – *Influence de la méthode sur la distribution des scores.*

5 Conclusion et perspectives

Il existe de nombreuses situations dans lesquelles seule existe une forte présomption de présence d'une sous-population d'intérêt dans une collection d'objets. Il s'agit alors d'un problème d'apprentissage non supervisé d'une nature particulière, peu étudié jusqu'ici. Dans ce papier, nous avons mis en avant les principes d'une approche combinant itérativement des *« experts faibles »* pour s'approcher de l'expert idéal. Sur les jeux de données contrôlés utilisés, les résultats obtenus montrent systématiquement une amélioration des performances par rapport aux *« experts faibles »* et à une méthode de k-moyenne dans l'espace des experts. Néanmoins, si cela fournit une preuve de principe de la méthode, ils ne suffisent pas à en établir les propriétés générales. Et les travaux à venir comprendront notamment une analyse formelle des bases de l'approche, de la forme des transformations non linéaires licites et de la manière de combiner les experts par paires en fonction, entre autres, de leur corrélation (Cornuéjols et Martin, 2011) afin de déterminer des ensembles de définition et des optimums.

À côté de ces études théoriques, il est nécessaire de poursuivre les expériences avec une base d'experts plus conséquente afin de déterminer s'il existe un nombre optimal d'experts à combiner, un ordre préférenciel et une base optimale d'experts faibles.

Pour terminer, de nombreux choix d'implémentation sont à analyser. Par exemple, la combinaison des experts doit-elle être symétrique, comme dans l'implémentation actuelle, ou favoriser l'un des experts en lui appliquant par exemple une fonction non linéaire plus marquée ?

Les résultats très encourageants obtenus motivent de futurs travaux sur ces perspectives.

	$\sigma = 1.0$	$= 2.0$	$= 2.5$	$= 3.0$	$= 3.5$	$= 4.0$	$= 4.5$
ANOVA	94.9 ± 1.4	73.7 ± 4.7	62.2 ± 4.4	55.2 ± 3.9	49.3 ± 4.2	44.0 ± 3.7	40.4 ± 3.7
RELIEF	97.1 ± 0.8	73.4 ± 2.7	61.6 ± 3.1	51.6 ± 2.8	44.9 ± 2.4	40.3 ± 2.0	37.7 ± 2.2
Comb1	96.9 ± 1.0	74.6 ± 3.2	63.4 ± 3.6	55.2 ± 3.2	48.4 ± 3.5	43.3 ± 3.7	39.8 ± 2.9
Exp_S.	95.3 ± 1.2	53.2 ± 12.3	20.2 ± 20.9	14.7 ± 13.8	11.9 ± 7.8	11.9 ± 2	12.7 ± 2.0
Comb2	**99.4** ± 0.4	**84.5** ± 1.7	**74.0** ± 2.2	**65.2** ± 2.4	**58.0** ± 2.8	**53.0** ± 2.5	**48.6** ± 2.4
Direct	98.8 ± 0.6	84.0 ± 2.5	74.0 ± 3.3	63.3 ± 9.1	52.8 ± 13.9	45.8 ± 14.1	38.6 ± 14.3

TAB. 1 – *Résultats obtenus sur 100 expériences pour chaque valeur*

Références

Cornuéjols, A. et C. Martin (2011). Unsupervised object ranking using not even weak experts. In *International Conference on Neural Information Processing (ICONIP 2011)*, Volume LNCS 7062, vol.1, Shanghai, China, pp. 608–616. Springer-Verlag.

Freund, Y. (1995). Boosting a weak learning algorithm by majority. *Information and Computation 121*, 256–285.

Freund, Y., R. Lyer, R. Schapire, et Y. Singer (2004). An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research 4*(6), 933–969.

Herbrich, R., T. Graepel, et K. Obermayer (2000). Large margin rank boundaries for ordinal regression. In A. Smola, P. Bartlett, B. SchŽlkopf, et D. Schuurmans (Eds.), *Advances in learge margin classifiers*, pp. 115–132. MIT Press.

Jong, K., J. Mary, A. Cornuéjols, E. Marchiori, et M. Sebag (2004). Ensemble feature ranking. In *Principles of Knowledge Discovery in Databases (PKDD-04)*, Volume LNAI-3202, Pisa, Italy, pp. 267–278. Springer-Verlag.

Kira, K. et L. Rendell (1992). A practical approach to feature selection. In *Int. Conf. on Machine Learning (ICML-92)*, pp. 249–256. Morgan Kaufmann.

Sahai, H. (2000). *The Analysis of Variance*. Birkhauser, Boston, MA.

Summary

Unsupervised learning seeks the detection of somewhat homogeneous sub-populations of a given set when no label is available. Each class or group is considered as equivalent. In this paper, we tackle a different, if related, problem. Based solely on the assumption that there exists a sub-population that interests us in the set of instances, we want to identify its member as well as possible.

The method presented here is based on three ideas: (1) using a change of representation whereby the objects are represented in the space of the evaluations of "weak experts", (2) through the study of experts in pairs, to apply a non linear transformation of the evaluations of each expert in order to amplify their tendency to discriminate objects of interest from the other, and (3) to combine iteratively the available "weak experts" to get a better final combined expert. The method has been tested and compared on synthetic data showing improvements in all cases.