Une nouvelle stratégie d'Apprentissage Bayésienne

Alexis Bondu*, Vincent Lemaire**
Marc Boullé**

*EDF R&D ICAME/SOAD, 1 avenue du Général de Gaulle, 92140 Clamart.
alexis.bondu@edf.fr,
**OrangeLabs, 2 avenue Pierre Marzin, 22300 Lannion.
prenom.nom@orange-ftgroup.com

Résumé. Dans cet article, une nouvelle stratégie d'apprentissage actif est proposée. Cette stratégie est fondée sur une méthode de discrétisation Bayésienne semi-supervisée. Des expériences comparatives sont menées sur des données unidimensionnelles, l'objectif étant d'estimer la position d'un échelon à partir de données bruitées.

1 Notations

Les données D sont composées de deux sous-ensembles T et U qui correspondent respectivement aux données étiquetées et non-étiquetées, avec $D = T \cup U$. L'ensemble T contient des couples (x, y), où $x \in \mathbb{R}$ et $y \in \mathbb{Y}$ est une valeur discrète représentant la classe de l'exemple x. L'ensemble U contient des réels. Les notations suivantes sont adoptées : N, le nombre d'exemples observables ($N = |D|$) ; N^l, le nombre d'exemples étiquetés ($N^l = |T|$) ; J, le nombre de classes observées dans les données ($J = |\mathbb{Y}|$).

2 Discrétisation semi-supervisée Bayésienne

L'approche MODL discrétise les variables explicatives dans le but d'estimer les distributions conditionnelles aux classes. Le problème de la discrétisation d'une variable numérique est transposé en un problème de sélection de modèles. Un modèle de discrétisation $M(I, \{N_i\}, \{N_{ij}\})$ est défini par les paramètres suivants : i) I est le nombre d'intervalles ; ii) $\{N_i\}$ est le nombre d'exemples dans chaque intervalle qui définit les bornes du modèle ; iii) $\{N_{ij}\}$ est le nombre d'exemples de chaque classe dans chaque intervalle, qui définit les distributions conditionnelles localement à chaque intervalle. Une démarche Bayésienne maximisant $P(M|D)$ est appliquée pour sélectionner le meilleur modèle de discrétisation, noté $\mathcal{M}_{map}$ *(Maximum a posteriori)*. Cette démarche revient à maximiser $P(M)P(D|M)$. La distribution a priori des modèles $P(M)$ et la vraisemblance des données $P(D|M)$ sont calculées analytiquement en exploitant le caractère discret de la famille de modèles, et en adoptant des hypothèses faiblement informatives sur les données. Finalement, le $\mathcal{M}_{map}$ minimise l'Equation 1 dont les deux termes correspondent à la distribution a priori des modèles et à la vraisemblance des données.

$$\mathcal{C}(M) = \overbrace{\log(N) + \log C_{N+I-1}^{I-1} \sum_{i=1}^{I} \log C_{N_i+J-1}^{J-1}}^{-logP(M)} + \underbrace{\sum_{i=1}^{I} \log\left(\frac{N_i!}{\sum_{j=1}^{J} N_{ij}!}\right) - \sum_{i=1}^{I} \log\left(\frac{N_i^u!}{\sum_{j=1}^{J} N_{ij}^u!}\right)}_{-logP(D|M)} \quad (1)$$

3 Une nouvelle méthode d'apprentissage actif

Cette section présente une stratégie originale d'apprentissage actif fondée sur la méthode de discrétisation semi-supervisée décrite à la Section 2. La qualité d'un modèle de discrétisation est donnée par la probabilité du modèle connaissant les données. Le critère $\mathcal{C}_{semi\ super}$ est une expression analytique de $P(M|D)$, au sens des hypothèses de modélisation de l'approche MODL (Boullé, 2006). Notre stratégie d'apprentissage actif cherche à étiqueter l'exemple qui maximisera la qualité du futur modèle, sans connaître la classe du nouvel exemple et sans connaître le meilleur modèle à l'itération suivante. Notre démarche prend en compte ces incertitudes en menant un calcul d'espérance sur tous les cas possibles. Un critère dont l'optimisation désigne l'exemple $x_{t+1} \in U$ qui maximise l'espérance de $P(M|D, x_{t+1})$ est établi.

$$\mathcal{C}_{actif}(x_{t+1}) = \sum_{M \in \mathbb{M}} \left[P(M)P(D|M) \times \sum_{y \in \mathbb{Y}} \left[\frac{P(y|M, D, x_{t+1}) \times P(M) \times P(D, x_{t+1}, y|M)}{\sum_{M' \in \mathbb{M}} P(M') \times P(D|M') \times P(y|D, M', x_{t+1})} \right] \right] \quad (2)$$

4 Conclusion

t Notre stratégie active est évaluée dans le cadre de l'estimation d'une fonction échelon à partir de données bruitées (Castro et Nowak, 2008) et est comparée à la dichotomie probabiliste (Horstein, 1963). Contrairement à notre approche, la dichotomie probabiliste doit être renseignée du niveau de bruit présent dans les données. Les deux approches donnent des résultats comparables lorsque le niveau de bruit est connu. Dans le cas général, cette information n'est pas disponible, c'est pourquoi la dichotomie probabiliste est renseignée d'un niveau de bruit erroné lors de son évaluation. Dans ces conditions, notre stratégie est plus performante que la dichotomie probabiliste. Finalement, notre stratégie est plus générique que la dichotomie probabiliste.

Références

Boullé, M. (2006). MODL: A bayes optimal discretization method for continuous attributes. *Machine Learning 65*(1), 131–165.

Castro, R. et R. Nowak (2008). *Foundations and Application of Sensor Management*, Chapter Active Learning and Sampling. Springer-Verlag.

Horstein, M. (1963). Sequential decoding using noiseless feedback. In *IEEE Transmition Information Theory*, Volume 9, pp. 136–143.

Summary

In this article, a new active learning strategy is proposed. Comparative experiments are conducted on unidimensional data, the aim is to estimate the location of a step-function from a noisy sample.