Approche bayésienne variationnelle pour l'agrégation de modèles en classification.

Stevenn Volant, Marie-Laure Martin-Magniette, Stéphane Robin

Approche bayésienne variationnelle pour l'agrégation de modèles en classification.

Stevenn Volant[1] , Marie-Laure Martin-Magniette [1,2] et Stéphane Robin[1]

[1] *UMR AgroParisTech/INRA MIA 518, 16 rue Claude Bernard, PARIS Cedex 05.*
[2] *UMR INRA 1165 - UEVE ERL CNRS 8196 URGV, 2 rue Gaston Crémieux, EVRY.*

Résumé Nous nous intéressons au cas d'un mélange entre deux populations dont l'une est connue et facilement identifiable. Plusieurs modèles ont été développés pour modéliser la distribution inconnue. Nous proposons une alternative qui consiste à prendre un mélange de plusieurs distributions gaussiennes de moyennes et variances inconnues. Chaque modèle apporte une information plus ou moins pertinente sur l'estimation des paramètres. Nous suggérons alors d'utiliser une approche BMA pour prendre en compte l'incertitude relative à chacun des modèles ainsi que de s'affranchir du choix du nombre de composants. En moyennant sur un ensemble de modèles, le BMA permet de calculer un estimateur agrégé à partir de l'information apportée par la collection de modèles, pondérée par le poids du modèle concerné. Dans la pratique, ce poids est estimé à partir du BIC mais la qualité de l'approximation pour obtenir ce critère est discutable. Ainsi, nous nous intéressons au cadre bayésien variationnel qui permet de définir naturellement une distribution a posteriori des paramètres et d'obtenir un poids pour chacun des modèles. Nous proposons dans ce travail la définition de poids de chaque modèle à partir de la minimisation de la divergence de Kullback-Leibler entre la distribution estimée des poids et la vraie. Une étude de simulation permet d'évaluer le comportement de notre estimateur agrégé.

Abstract We consider a mixture between two populations, one is known and easily identifiable and the other is unknown. Several models have been developed to fit this unknown distribution. We propose an alternative that involves taking a mixture of Gaussian distributions with unknown parameters. Each model provides more or less relevant information to parameter estimation. We suggest to use BMA aproach which takes model uncertainty into account and eliminates the choice of the number of components. Based on a weighted average over a set of models, BMA calculates an aggregated estimator from information provided by each model. In practice, the weight is estimated using the BIC criterion, but the quality of this approximation is questionable. Thus, we focus on the variational bayesian framework which allows to define a posterior distribution over the parameters and a natural weight for each model. In this work, we propose a definition of model weight based on the minimization of the Kullback-Leibler divergence between the estimated distribution and the true one. A simulation study allows to evaluate the aggregate estimator behaviors.

Mots clés BMA, bayésien variationnel, modèle de mélange.

Nous nous intéressons à un mélange entre deux populations dont l'une est connue et facilement identifiable et l'autre totalement inconnue. La densité de probabilité au point x est alors définie par :

$$g(x) = af(x) + (1-a)\Phi(x), \tag{1}$$

où la proportion a et la fonction de densité f sont inconnues et Φ est une fonction de densité connue.

Ce mélange peut s'interpreter comme la loi marginale de X obtenue à partir du couple (X, Z) où Z désigne le groupe d'appartenance non observé. Dans un objectif de classification, on s'intéresse à τ_0 la probabilité d'appartenir au groupe caractérisé par la fonction de densité connue Φ :

$$\tau_0(x) = P(Z = 0|X = x) = \frac{af(x)}{g(x)}. \tag{2}$$

Le modèle donné en Equation (1) avec les variables latentes Z indépendantes et distribuées selon une loi multinomiale a été proposé par Efron (2001) pour calculer un False Discovery Rate local (FDR local) : les observations sont la transformation probit des probabilités critiques brutes et Φ est la densité d'une gaussienne centrée réduite. La qualité de l'estimation de τ_0 dépend fortement de la modélisation de la fonction f et les méthodes existantes se sont alors concentrées sur l'estimation de cette fonction. Efron (2001) a mis en place une méthode bayésienne empirique fondée sur le calcul de scores. McLachlan et al. (2006) ont proposé de prendre une distribution gaussienne de paramètres inconnus pour modéliser f. L'estimation de la variable τ_0 est alors obtenue par un simple mélange de deux distributions gaussiennes dont l'une est connue. Par ailleurs, Robin et al. (2007) se sont placés dans un cadre non-paramétrique avec un estimateur à noyau.

Dans un cadre plus général, les variables latentes ne sont pas forcément indépendantes et elles sont alors distribuées suivant une chaîne de Markov. Cette modélisation appelée HMM(Hidden Markov Model) permet de prendre en compte une dépendance spatiale car le groupe d'une observation dépendra du groupe de l'observation précédente. Dans ce contexte, la probabilité τ_0 ne peut pas être obtenue directement. Un algorithme avant-arrière (Forward-Backward en anglais) nous donne la possibilité d'obtenir cette probabilité de manière itérative. Ce type de modèle peut notamment être utilisé dans le cadre des données transcriptomes de type tiling array, pour lesquelles la dépendance entre les sondes est avérée. La généralisation aux données multidimensionnelles est directement obtenue via les modèles HMRF(Hidden Markov Random Fields) qui peuvent être utilisés pour traiter des données de type NGS (New Generation Sequencing).

Dans ce travail, pour la modélisation de f, nous proposons un compromis entre les approches de Robin et al. (2007) et McLachlan et al. (2006) en considérant une collection de mélange de gaussiennes dont le nombre de composants varie. La variable d'intérêt étant τ_0, il n'est pas judicieux de se focaliser sur l'estimation de ce nombre de composants comme le proposent les méthodes de sélection de modèles. Nous proposons une alternative qui consiste à agréger l'information des modèles de la collection afin d'obtenir un estimateur $\tilde{\tau}_0$ de variance plus faible. Cette méthode d'agrégation appelée le BMA (Bayesian Model Averaging) a été proposée par Raftery et al. (1999) et permet de prendre en compte de l'incertitude relative à chaque modèle de la collection. Des résultats théoriques sur le gain apporté par le BMA ont été démontrés dans Madigan et Raftery (1994) ; Madigan et al. (1995) ; Raftery et al. (1997).

Dans notre cas, la collection définie par l'ensemble des mélanges de 2 à K composants est notée $\mathcal{M} = \{M_1, \ldots, M_{K-1}\}$ et $\hat{\tau}_0^{(k)}$ l'estimateur de τ_0 dans le modèle M_k. Sachant que chaque modèle possède un degré d'incertitude, l'estimateur agrégé $\tilde{\tau}_0$ est défini par :

$$\tilde{\tau}_0 = \sum_{k=1}^{K-1} \hat{\tau}_0^{(k)} P(M_k|X), \tag{3}$$

où X caractérise les données observées. $P(M_k|X)$ représente la probabilité a posteriori du modèle M_k, i.e. le poids du modèle M_k par rapport aux autres modèles de la collection. Pour obtenir cet estimateur agrégé il faut donc estimer τ_0 dans chaque modèle de la collection et définir le poids du modèle. Concernant le poids, dans la pratique ce calcul est réalisé à partir d'une approximation BIC qui approche $\log P(M_k|X)$ (Raftery et al. (1997)). Bien que cette approximation permet de bien choisir un nombre de composants elle ne permet pas de juger du poids relatif d'un modèle. Nous avons donc proposé de nouveaux poids quand l'objectif est l'agrégation de modèles.

Pour l'estimation de τ_0 dans un modèle donné, nous avons opté pour une approche bayésienne variationnelle où les paramètres du modèle sont considérés comme des variables aléatoires, distribués selon une distribution de probabilité, appelée distribution a priori. Cette dernière permet de prendre en compte une information connue sur les paramètres du modèle. Le but du bayésien est de réviser cette connaissance a priori à l'aide de l'information apportée par les données disponibles, on parle alors de distribution a posteriori. L'hypothèse du bayésien variationnel consiste alors à supposer que, conditionnellement aux données, la variable latente Z et des paramètres sont indépendants. Cette approximation variationnelle peut être vue comme une alternative aux algorithmes de type MCMC dont l'utilisation pour des échantillons de grande taille n'est pas recommandée (Wang et Titterington 2004)). Pour l'estimation des paramètres, on se place dans le cadre des modèles exponentiels conjugués (Beal et Ghahramani (2003)) et on utilise un algorithme

itératif type EM. La convergence ces estimateurs a été démontrée par Wang et Titterington (2004) et Wang et Titterington (2002).

L'approche bayésienne variationnelle permet de construire de manière naturelle le poids du modèle M_k à partir de la minimisation de la divergence de Kullback-Leibler entre une distribution approchée et la vraie $P(M|X)$. On peut ainsi réaliser une agrégation de modèle sans passer par un critère BIC dont l'approximation de $\log P(M_k|X)$ n'est pas satisfaisante. Nous proposons trois types de poids et nous avons étudié dans une étude de simulation les trois estimateurs dérivés en simulant plusieurs jeux de données où les variables latentes sont soit indépendantes soit distribuées selon une chaîne de Markov.

Bibliographie

[1] M.J. Beal et Z. Ghahramani (2003) , The Variational Bayesian EM Algorithm for Incomplete Data : with Application to Scoring Graphical Model Structures, Gatsby Computational Neuroscience Unit.

[2] B. Efron, R. Tibshirani, J.D. Storey et V. Tusher (2001) *Empirical Bayes analysis of a microarray experiment*, Journal of the American Statistical Association.

[3] D. Madigan et A.E. Raftery (1994) *Model Selection et Accouting for Model Uncertainty in Graphical Models Using Occam's Window* , Journal of the American Statistical Association.

[4] D. Madigan et J. Gravin et A.E. Raftery (1995) *Elicting Prior Information to Enhance the Predictive Performance of Bayesian Graphical Models* , Theory et Methods.

[5] G. McLachlan et R.W. Bean et L. Ben-Tovim Jones (2006) *A simple implementation of a normal mixture approach to differentiate gene expression in multiclass microarrays*, Bioinformatics.

[6] JA.E. Raftery et D. Madigan et J.A. Hoeting (1997) Selection et Accouting for Model Uncertainty in Linear Regression Models , Journal of the American Statistical Association.

[7] A.E. Raftery et D. Madigan et C.T. Volinsky (1997) *Bayesian Model Averaging in Proportional Hazard Models : Assessing the Risk of a Stroke*, Applied Statistics.

[8] A.E. Raftery, J.A. Hoeting, D. Madigan et C.T. Volinsky (1999) *Bayesian Model Selection : A Tutorial* , Statistical Science.

[9] S. Robin et A. Bar-Hen et J.J. Daudin et L. Pierre (2007) *A Semi-parametric Approach for Mixture Models : Application to Local False Discovery Rate Estimation*, Computational Statistics & Data Analysis.

[10] B. Wang et D.M. Titterington (2002) *Variational Bayes Estimation of Mixing Coefficient*, Lecture Notes in Computer Science, University of Glasgow.

[11] B. Wang et D.M. Titterington (2004) *Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model*, Proc. 20th Conf. Uncertainty in Artificial Intell., University of Glasgow.