Estimation incrémentale pour la détection non supervisée d'anomalies multivariées en imagerie médicale

Geoffroy OUDOUMANESSAH[1,2,3] Carole LARTIZIEN[2] Michel DOJAT[3] Florence FORBES[1]

[1]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France

[2]Univ. Lyon, CNRS, Inserm, INSA Lyon, UCBL, CREATIS, UMR5220, U1294, F-69621, Villeurbanne, France

[3]Univ. Grenoble Alpes, Inserm U1216, Grenoble Institut des Neurosciences, 38000 Grenoble, France

Résumé – La détection non supervisée d'anomalies en imagerie médicale est une tâche complexe. Les modèles de l'état de l'art basés sur des réseaux de neurones artificiels obtiennent des performances raisonnables mais au prix d'une consommation élevée de ressources computationnelles. Nous montrons que des modèles statistiques de mélange de distribution de probabilité, en particulier des mélanges gaussiens multivariés ou de lois de Student généralisées sont des alternatives aussi performantes, frugales et sans paramétrisation excessive. Nous proposons une approche incrémentale originale qui estime séquentiellement les hyperparamètres et permet ainsi de traiter de gros volumes de données. Nous illustrons l'efficacité de cette approche avec un problème de détection d'anomalies subtiles dans des IRM cérébrales de patients Parkinsoniens récemment diagnostiqués.

Abstract – Unsupervised anomaly detection in medical imaging is a complex task. State of the art models based on artificial neural networks obtain reasonable performances but at the cost of a high consumption of computational resources. We show that statistical mixture models of probability distributions, in particular multivariate Gaussian mixtures or generalized Student's distributions, are equally efficient, frugal alternatives without excessive parameterization. We propose an original incremental approach that sequentially estimates the hyperparameters and thus allows to handle large volumes of data. We illustrate the efficiency of this approach with a problem of detecting subtle anomalies in brain MRIs of newly diagnosed Parkinson's patients.

1 Introduction

Dans ce travail, nous nous intéressons à la détection d'anomalies subtiles dans les images médicales, dans un contexte non supervisé généralement appelé *détection d'anomalies non supervisée* (UAD en anglais). Ce formalisme nécessite la construction d'un modèle de *normalité* (ou *référence*). Les *anomalies* sont ensuite détectées comme des valeurs aberrantes, c'est-à-dire comme des échantillons s'écartant de ce modèle de référence. Les réseaux de neurones artificiels (ANN) ont été largement utilisés en UAD [8]. Qu'ils soient basés sur des architectures standards d'autoencodeurs ou sur des architectures plus avancées, les ANN ne parviennent généralement pas à un compromis optimal entre précision et coût computationnel. En guise d'alternative, nous montrons que des approches plus frugales peuvent être utilisées avec des modèles statistiques traditionnels. Les solutions frugales se réfèrent généralement à des stratégies qui peuvent fonctionner avec des ressources limitées, telles que celles d'un seul ordinateur portable. L'apprentissage frugal a été étudié sous plusieurs angles, sous la forme de contraintes sur les données acquises, sur l'algorithme implémenté et sur la nature de la solution proposée [5]. L'angle que nous adoptons est celui de l'apprentissage en ligne ou incrémental, qui se réfère aux approches qui traitent les données de manière séquentielle. Cela permet d'obtenir des solutions plus efficaces en termes d'utilisation de la mémoire vive et de consommation d'énergie. Pour l'UAD, nous proposons d'étudier des mélanges de distributions de probabilité dont l'interprétabilité et la polyvalence ont été largement reconnues pour un certain nombre de tâches, tout en ne nécessitant pas d'effort de conception ou de paramétrisation excessive. En particulier, l'utilisation de mélanges gaussiens multivariés ou de mélanges de lois de Student généralisées a déjà été proposée dans de nombreuses tâches de détection d'anomalies, voir [11]. Toutefois, dans le cadre standard, les mélanges sont difficiles à utiliser avec de grands ensembles de données en raison de l'augmentation considérable du temps et de la mémoire nécessaires à leur estimation, traditionnellement effectuée à l'aide d'un algorithme de maximisation de l'espérance (EM) [10]. Des versions en ligne de l'algorithme EM ont été proposées et étudiées théoriquement dans la littérature, *e.g.* [3], mais avec certaines restrictions sur la classe de mélanges qui peut être traitée de cette façon. Une première approche naturelle consiste à considérer les mélanges gaussiens qui appartiennent à cette classe. Nous présentons donc des résultats obtenus avec des mélanges gaussiens et un EM en ligne dont l'implémentation est améliorée. Nous considérons ensuite des mélanges plus généraux basés sur des lois de Student à échelles multiples plus adaptées à la détection de valeurs aberrantes [6]. Nous montrons que ces mélanges peuvent aussi être intégrés dans le cadre EM en ligne et décrivons les principales étapes de l'algorithme qui en résulte.

Nous illustrons notre approche avec des données issues d'IRM cérébrale de patients atteints de la maladie de Parkinson (MP) à des stades précoces (*de novo*), où les anomalies cérébrales sont subtiles et difficilement visibles dans des IRM standard pondérées en T1 ou de diffusion. L'évolution de la pathologie est caractérisée par un score sur l'échelle de Hoehn et Yahr (HY), qui décrit l'évolution des symptômes. Notre approche non supervisée permet une détection des modifications structurales fines des structures sous-corticales les plus touchées aux stades 1 et 2 de l'échelle HY. Nos résultats

sont cohérents avec les connaissances physio-pathologiques de l'évolution de la MP. En l'absence de vérité terrain au niveau du voxel, l'utilisation de l'échelle HY est une validation indirecte originale et pertinente. La consommation en énergie et en mémoire est également rapportée, pour l'EM standard et en ligne afin de confirmer le compromis performance/coût obtenu.

2 UAD avec modèles de mélanges

Dans ce travail, nous montrons comment les modèles de mélanges peuvent être utilisés dans le contexte de l'UAD en construisant un modèle de référence et une règle de décision.

Apprentissage d'un modèle de référence. Nous considérons un ensemble Y_H de caractéristiques au niveau voxel pour un certain nombre de sujets contrôles (*e.g.* sains, définissant la normalité), $\mathsf{Y}_H = \{\mathbf{y}_v; v \in \mathsf{V}_H\}$ où V_H représente les voxels de tous les sujets contrôles et $\mathbf{y}_v \in \mathsf{R}^M$ est extrait de chaque voxel v des images de différentes modalités ou de caractéristiques de représentation fournies par un ANN effectuant une tâche prétexte. Entrainer un modèle de mélange avec des données contrôles Y_H, revient à construire un modèle de référence de densité f_H qui dépend de paramètres notés $_H = \{ _k; _k; k = 1 : K_H\}$. Nous considérons deux types de mélanges, à savoir les mélanges gaussiens et les mélanges de lois de Student multi-échelles (MST) qui sont plus appropriés lorsque les données présentent des sous-groupes allongés et fortement non elliptiques [6]. :

$$f_H(\mathbf{y}; \boldsymbol{\Theta}_H) = \sum_{k=1}^{K_H} \pi_k f(\mathbf{y}; \boldsymbol{\theta}_k), \quad (1)$$

avec $_k \in [0; 1]$, $\sum_{k=1:K_H} _k = 1$ et K_H le nombre de composantes, chacune caractérisée par une distribution $f(; _k)$. L'algorithme EM est généralement utilisé pour estimer la valeur de $_H$ qui correspond le mieux à Y_H tandis que K_H peut être estimé en utilisant *l'heuristique de pente* [1].

Mesure de proximité. Étant donné un modèle de référence (1), il faut choisir une mesure de proximité $r(\mathbf{y}_v; _H)$ du voxel v (de valeur $\mathbf{y}_v$) à f_H. Pour utiliser la structure du mélange, nous proposons de prendre en compte les distances aux composantes respectives du mélange par le biais de certains poids agissant comme des distances de Mahalanobis inverses. Les distributions MST sont des généralisations des lois de Student multivariées où la variable d'échelle (poids) univariée de la distribution standard est remplacée par une variable multivariée $\mathbf{W} = (W_m)_{m=1:M} \in \mathsf{R}^M$,

$$f_{MST}(\mathbf{y};) = \int_{[0;1]^M} N_M(\mathbf{y}; ; D \ _{\mathbf{w}} A D^T) \prod_{m=1}^{M} G\left(w_m; \tfrac{m}{2}\right) dw_{1:M} \quad (2)$$

où $G(; \frac{m}{2})$ désigne la densité gamma avec le paramètre $(\frac{m}{2}; \frac{m}{2}) \in \mathsf{R}^2$ et N_M la distribution normale multivariée de moyenne $\in \mathsf{R}^M$ et de covariance $D \ _{\mathbf{w}} A D^T$ montrant la pondération par les W_m via une matrice diagonale $_w = diag(w_1^{-1}; \ldots; w_M^{-1})$, avec $D \in O(M) \quad \mathsf{R}^{M \times M}$ orthogonale et $A = diag(A_1; \ldots; A_M)$ diagonale. L'ensemble complet de paramètres est $= \{ ; A; D; (_m)_{m=1:M}\}$. La variable d'échelle W_m pour la dimension m peut être interprétée comme tenant compte du poids de cette dimension et peut être utilisée pour dériver une mesure de proximité. Après avoir ajusté un mélange (1) avec des composantes MST à Y_H, nous définissons $r(\mathbf{y}_v; _H) = \max_{m=1:M} w_m^{\mathbf{y}_v}$, avec $w_m^{\mathbf{y}} = \mathsf{E}[W_m | \mathbf{y}; _H]$. La proximité r est généralement plus grande lorsqu'au moins une dimension de $\mathbf{y}_v$ est bien expliquée par le modèle. Une mesure de proximité similaire peut également être dérivée pour les mélanges gaussiens.

Règle de décision. Un seuil sur les scores de proximité peut être calculé de manière empirique en décidant d'un taux de faux positifs (FPR) acceptable ; est la valeur telle que $P(r(\mathsf{Y}; _H) <) = $, lorsque Y suit la distribution de référence f_H. Tous les voxels v dont la proximité $r(\mathbf{y}_v; _H)$ est inférieure à sont alors étiquetés comme anormaux. En pratique, la distribution de $r(\mathsf{Y}; _H)$ n'est pas connue mais il est facile de la simuler pour estimer .

Malheureusement, pour de grands volumes de données, l'estimation de f_H avec EM est impossible. Une première solution est d'augmenter la puissance de calcul. Dans la section suivante nous présentons une solution moins energivore utilisant une version en ligne de l'algorithme EM.

3 Estimation en ligne d'un mélange

L'estimation en ligne fait référence à des procédures capables de traiter des données acquises séquentiellement. Nous considérons l'EM en ligne de [3] qui appartient à la famille des algorithmes d'approximation stochastique [2]. Cet algorithme a été bien étudié et étendu théoriquement. Cependant, il est conçu uniquement pour des distributions qui admettent une vraisemblance complète appartenant à la famille exponentielle.

EM en ligne pour les mélanges MST. Dans un premier temps il faut montrer que (2) admet une forme exponentielle et expliciter la statistique exhaustive et le parametre naturel. Le cas du mélange se déduit alors du cas à une composante, *c.f.* [12]. La vraisemblance complète pour une distribution (2) s'écrit (avec $\mathbf{x} = (\mathbf{y}; \mathbf{w})$),

$$f_c(\mathbf{x};) = N_M(\mathbf{y}; ; D \ _{\mathbf{w}} A D^T) \prod_{m=1}^{M} G\left(w_m; \tfrac{m}{2}\right) :$$

Elle admet la forme exponentielle suivante :

$$f_c(\mathbf{x};) = h(\mathbf{x}) \exp\left([s(\mathbf{x})]^{>} (; D; A;) \quad (; D; A;)\right)$$

avec $s(\mathbf{x})$ vecteur obtenu par concatenation des sous-vecteurs :

$$s(\mathbf{x}) = \left[w_1 \mathbf{y}; w_1 vec(\mathbf{y}\mathbf{y}^{>}); w_1; \log w_1; \ldots; w_M \mathbf{y}; w_M vec(\mathbf{y}\mathbf{y}^{>}); w_M; \log w_M\right]$$

$$(; D; A;) = [_1; \ldots; _M]^T, \text{ avec}$$

$$_m = \left[\frac{d_m d_m^T}{A_m}; \ \frac{vec(d_m d_m^T)}{2A_m}; \ \frac{vec(d_m d_m^T)^T vec(\ ^T)}{2A_m} \quad \frac{m}{2}; \frac{1 + m}{2}\right]$$

$$(; D; A;) = \sum_{m=1}^{M} \left(\frac{\log A_m}{2} + \log (\tfrac{m}{2}) \quad \tfrac{m}{2} \log(\tfrac{m}{2})\right)$$

où d_m désigne la $m^{\text{ième}}$ colonne de D et $vec( )$ l'opérateur de vectorisation, qui convertit une matrice en un vecteur colonne. La forme exacte de h n'est pas importante pour l'algorithme. Pour estimer, à partir d'observations $(\mathbf{y}_i)_{i=1:N}$, le paramètre d'une loi MST, l'EM en ligne consiste à calculer à chaque itération (i) une estimation courante $^{(i)} = (s^{(i)})$ où $s^{(i)}$ est une statistique mise à jour sequentiellement par l'addition d'un nouveau terme s pondéré par un taux $_i$ (voir [3]) : $s^{(i)} = (1 \quad _i) s^{(i \ 1)} + _i s(\mathbf{y}_i; ^{(i \ 1)})$. La première quantité importante pour mettre en oeuvre l'EM en ligne est donc (s) qui est défini comme l'unique maximiseur de la

fonction $Q(s;\ ) = s^T\ (\ )\quad (\ )$ où s est un vecteur de la dimension de $(\ )$. L'annulation des gradients de Q conduit à $(s) = (\ (s); A(s); D(s);\ (s))$ dont l'expression, non détaillée ici, permet la mise à jour des paramètres.

Une seconde quantité importante est $s(y;\ )$ qui permet la mise à jour de la statistique $s^{(i)}$. Cette quantité est définit par $s(y;\ ) = E[s(X)|Y = y;\ ]$. Elle nécessite le calcul des espérances suivantes, $E[W_m|Y = y;\ ]$ et $E[\log W_m|Y = y;\ ]$, pour tout m. Ces espérances doivent être calculées pour $y = y_i$ et $^{(i\ 1)}$. Pour une MST on a :

$$E[W_m|y_i;\ ^{(i\ 1)}] = \ ^{(i\ 1)}_m = \ ^{(i\ 1)}_m$$

$$E[\log W_m|y_i;\ ^{(i\ 1)}] = \ ^{(0)}(\ ^{(i\ 1)}_m)\quad \log\ ^{(i\ 1)}_m$$

où $^{(i\ 1)}_m = \frac{^{(i-1)}_m+1}{2}$ et $^{(i\ 1)}_m = \frac{^{(i-1)}_m}{2} + \frac{\left(d_m^{(i-1)T}(\mathbf{y}_i\quad ^{(i-1)})\right)^2}{2A_m^{(i-1)}}$.

EM en ligne pour les mélanges gaussiens. Ce cas est présent dans des travaux antérieurs *e.g.* [13] mais à notre connaissance, l'optimisations de son implémentation n'est jamais vraiment abordée. Or, le cas multivarié implique de nombreuses inversions et calculs de déterminants de grandes matrices. Le coût de ces calculs peut être diminué en appliquant la formule de Sherman–Morrison et le lemme du déterminant matriciel.

4 Application à la détection d'anomalies cérébrales en IRM 3D

L'objectif est d'évaluer l'apport des modèles de mélanges de distributions gaussiennes et MST à la détection d'anomalies en imagerie par résonnance magnétique (IRM) 3D de patients atteints de formes précoces de la maladie de Parkinson.

Données. La base de données Parkinson's Progression Markers Initiative (PPMI) [9] contient des examens IRM 3D de patients atteints de la maladie de Parkinson (PP), ainsi que de sujets sains (SS). Pour notre illustration, nous utilisons 108 sujets sains et 260 patients. Pour chaque sujet, nous disposons d'une image 3D pondérée en T1 (T1w), de volumes d'anisotropie fractionnelle (FA) et de diffusivité moyenne (MD). Les groupes SS et PP sont appariés en âge (âge médian : 64 ans) avec un ratio homme/femme égal à $6:4$. Nous nous concentrons sur les structures sous-corticales, principalement touchées au stade précoce de la maladie : Globus Pallidus externe et interne (GPe et GPi), Nucleus Accumbens (NAC), Substantia Nigra reticulata (SNr), Putamen (Pu), Caudate (Ca) et Extended Amygdala (EXA) [4]. Leur position est déterminée à l'aide de l'atlas CIT168 [14].

Méthode et résultats. Nous suivons les sections 2 et 3 en utilisant les volumes T1w, FA et MD comme caractéristiques ($M = 3$) et un FPR $= 0;02$. L'experience est répétée 10 fois pour une validation croisée. Chaque fold est composé de 64 images SS pour l'entraînement (70M voxels), 44 SS et la totalité des 260 PP pour le test, les images saines étant à chaque fois sélectionnées au hasard dans la base de données SS. Pour le modèle de référence, nous testons les mélanges gaussiens et MST, avec respectivement $K_H = 14$ et $K_H = 8$, estimés avec l'heuristique de pente. Les voxels anormaux sont ensuite détectés pour tous les sujets testés, sur la base de leur proximité avec le modèle de référence appris, comme détaillé dans la section 2.

PPMI ne fournit pas d'informations sur les anomalies structurelles au niveau du voxel (vérité terrain). Il s'agit d'un problème récurrent en UAD, qui limite les validations à des évaluations qualitatives. Pour une évaluation plus quantitative, nous proposons de recourir à une tâche auxiliaire dont le succès est susceptible d'être corrélé à une bonne détection des anomalies. Nous considérons la classification des sujets testés en sujets sains et sujets parkinsoniens sur la base de leurs pourcentages globaux (sur l'ensemble du cerveau) de voxels anormaux. Nous exploitons la disponibilité des valeurs HY pour diviser les patients en deux groupes HY=1 et HY=2, représentant les deux premiers stades de la progression de la maladie. Les résultats de la classification donnent une moyenne *gmean* [7], pour le stade 1 par rapport au stade 2, respectivement de 0,59 contre 0,63 pour le modèle de mélange gaussien et de 0,63 contre 0,65 pour le mélange MST. La capacité des deux mélanges à mieux différencier les patients de stade 2 des patients de stade 1 de l'échelle HY est cohérente avec la progression de la maladie. Il convient de noter que les différences structurelles entre ces deux stades de la MP restent subtiles et difficiles à détecter, ce qui démontre l'efficacité des modèles.

Le modèle de mélange MST semble mieux identifier les patients atteints de la MP au stade 2 sur la base de leurs voxels anormaux. Pour mieux l'appréhender, nous présentons, dans la Figure 1, les pourcentages d'anomalies détectées dans chaque structure sous-corticale, pour les groupes SS et PP de stade 1 et de stade 2. Pour chaque structure et pour les deux modèles de mélange, le nombre d'anomalies augmente du groupe témoin au groupe PP de stade 1 et au groupe PP de stade 2. Comme attendu, le mélange MST montre une meilleure capacité à détecter les anomalies avec des différences significatives entre les groupes SS et PP, tandis que pour le modèle gaussien, les pourcentages ne s'écartent pas beaucoup de ceux du groupe témoin. Dans l'ensemble, conformément aux connaissances courantes en physio-pathologie, les résultats suggèrent que toutes les structures sous-corticales sélectionnées sont de bons marqueurs potentiels de la progression de la maladie à ces stades précoces, GPe, GPi, EXA et SNr étant les plus impactées.

En ce qui concerne le coût computationel de notre approche, la consommation d'énergie est mesurée à l'aide de la bibliothèque PowerAPI. A des fins de comparaison, les données PPMI sont réduites à un sous-ensemble plus petit de 300 000 voxels pour permettre l'exécution de l'EM standard sur une station de travail standard (Intel i7-4790 @ 3,60 Ghz). Les deux modèles montrent une réduction significative de la consommation de l'unité centrale, les versions en ligne consomment moins d'énergie que leurs homologues en EM standard, de 20 kJ à 7 kJ pour les mélanges gaussiens, et de 40 kJ à 15 kJ pour les mélanges MST. Les algorithmes en ligne, qui traitent les données de manière séquentielle, affichent également une consommation d'énergie DRAM inférieure, de 2 kJ à 1 kJ pour les mélanges gaussiens, et de 5 kJ à 1,8 kJ pour les mélanges MST. À titre de comparaison, nous implémentons un autoencodeur convolutionnel à 8 couches basé sur des patchs, dont la consommation de DRAM est similaire à celle des algorithmes en ligne, mais dont la consommation sur CPU est supérieure de 30 kJ. Il est aussi important de noter que la comparaison n'est pas tout à fait juste car l'autoencodeur nécessite beaucoup plus de données pour converger. En ce qui concerne le coût mémoire, les résultats des pics de DRAM, mesurés par