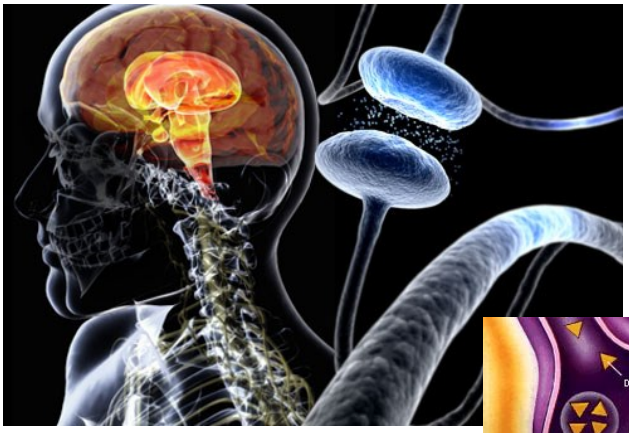
Le génie pour l'industrie

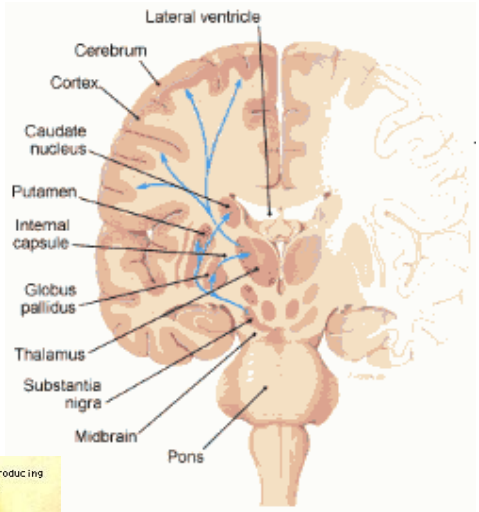
Diagnostic précoce de la maladie de Parkinson

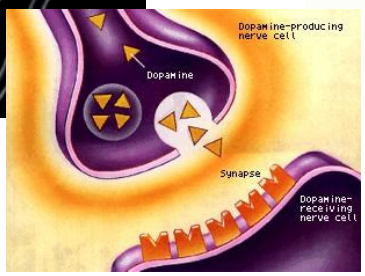
Par classification automatique et modèle de prédiction

Loïc TETREL

01/12/2014







La détection précoce et précise de la maladie de Parkinson est importante pour la prise en charge rapide des patients, ainsi que pour diminuer les coûts liés aux traitements médicaux. Les nouvelles techniques d'imagerie (SPECT) avec 123I-Ioflupane (DaTSCAN) ont montrées leurs performances pour détecter des stages précoces de la maladie. En utilisant les caractéristiques SBR (obtenues avec la base PPMI), je propose divers techniques de classification automatique du patient : SVM gaussien, Régression logistique et des classifieurs paramétriques et non paramétriques basés sur la règle de Bayès. Le classifieur SVM gaussien offre un taux d'erreur de 6%, le MLR et les classifieurs bayésien ont un taux d'erreur au-dessus de 10%. Mais ces derniers ont l'avantage de proposer une probabilité de la maladie pour chaque patient. Bien que l'étude supporte que ces techniques ont le potentiel d'aider les cliniciens dans la détection de la maladie, une application multi-classifieurs pourrait être proposée sous condition d'indépendance des classifieurs.

Table des matières

Introduction

La maladie de Parkinson est une maladie neurodégénérative, caractérisée par la perte des cellules produisant de la dopamine [1]. La dopamine est un produit chimique qui fait voyager les signaux entre les neurones du cerveau. Actuellement, il n'y a pas de tests définitifs qui permettent le diagnostic de la maladie. La maladie évolue à un rythme dépendant de la personne. Et c'est en analysant la progression de ces bio-marqueurs que l'on peut tirer des hypothèses sur l'existence de la maladie sur un patient ou non.

La détection précoce est cruciale car elle permet une prise en charge du patient dès l'apparition de la maladie. Cela permet d'éviter des examens médicaux inutiles ainsi que leurs coûts associés. Et éviter au patient des effets secondaires liées à la prise de grandes quantités de médicament, pour sa propre sécurité. L'utilisation d'outils d'apprentissage supervisée permet ainsi d'assister le clinicien dans le diagnostic précoce de la maladie.

Ce papier est découpé en plusieurs parties, la première explique le contexte global du sujet et ce qui m'a permis d'arriver à bout du projet. Une petite section sur l'Etat de l'Art permettra d'appuyer les points forts de ma méthode. Découlera ensuite plusieurs sections sur les analyses des données et des différents classifieurs. Une dernière partie sur la comparaison des différents classifieurs pour finir avec une conclusion avec entre autre ma contribution scientifique en comparant aux résultats de l'article (qui m'a servi de fil conducteur).

Matériel et Méthodes

Pour permettre de diagnostiquer la maladie, il est nécessaire d'obtenir des caractéristiques, selon qu'un patient soit malade ou non. Pour ce projet, 4 caractéristiques « Striatal Binding Ratio » (SBR) ont été utilisées. Ces SBR sont en fait des ratio sur les bio-marqueurs liés à la maladie de parkinson pour chacune des 4 régions striatal du cerveau :

- SBR du Caudate droit
- SBR du Caudate gauche
- SBR du Putamen droit
- SBR du Putamen gauche

Ces caractéristiques ont été calculées à partir d'images médicales « 123-I Ioflupane SPECT Imaging Processing » qui est une nouvelle technique d'extraction.

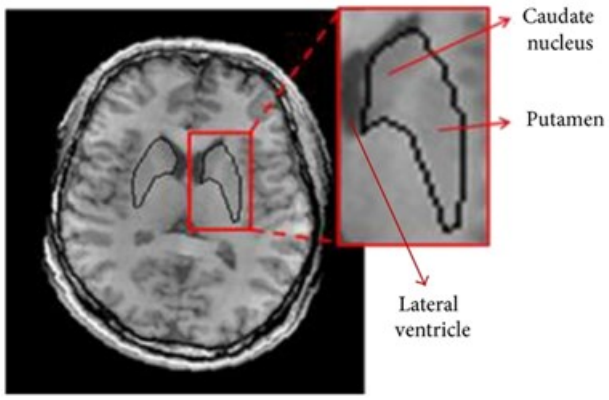


Figure 1 - Striatum du basal ganglia

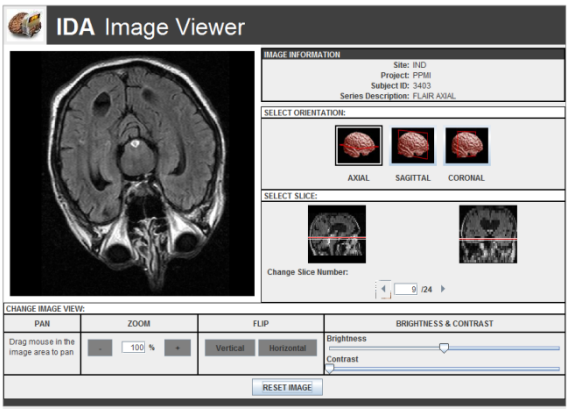


Figure 2 - Image médicale type

La base de donnée « Parkinson's progression markers initiative » (PPMI) propose des données de patients atteints ou non de la maladie de Parkinson. C'est la première étude globale, à très grande échelle et internationale à identifier la progression de la maladie de Parkinson [2]. Cette étude permet d'améliorer la compréhension de la maladie ainsi que l'efficacité des traitements sur le patient.
Elle met en avant si le patient est atteint de la maladie, avec bien évidemment les caractéristiques associées. Elle propose aussi des images pour chacun de ces patients (voir

Figure 2).

J'ai téléchargé la base de données en Octobre 2014 en m'inscrivant sur le site http://www.ppmi-info.org/data. Elle est composée au total de 736 patients (dont 71.33% de Parkinson), on se retrouve donc dans un cas de classification 2D (Parkinson ou non).

Mes méthodes sont basées sur un article existant [3] auquel j'apporte ma contribution. Ainsi, après avoir analysé la disparité entre les 2 classes pour chacune des caractéristiques, j'ai implémenté 4 classifieurs au total :

- Un linéaire et non linéaire SVM
- Un par modèle de régression multinomiale logistique
- Une approche bayésienne non-paramétrique
- Une approche bayésienne paramétrique

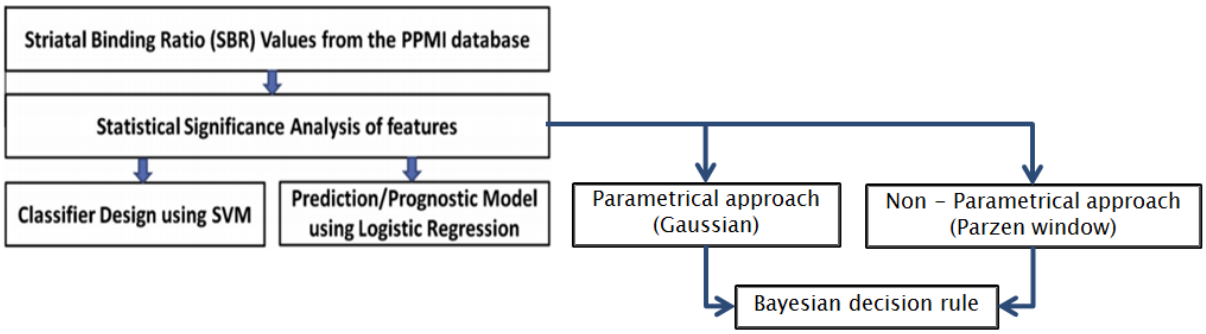


Figure 3 - Etapes du projet

L'utilisation de ces différents classifieurs sont toutes basées sur des méthodes différentes, ce qui permet de bien comparer les résultats de l'un avec les autres. De plus, ces autres approches apportent des informations complémentaires sur l'état du patient.

Pour ce projet j'ai utilisé le logiciel *MATLAB* qui propose de nombreux outils mathématiques faciles d'accès, avec une grande communauté sur internet.

Etat de l'Art

Beaucoup d'auteurs ont travaillés sur la problématique de classification précoce de la maladie de Parkinson. On retrouve par exemple Segovia et al. [4], Towey et al. [5] ou encore Rojas et al. [6].

Segovia et al. extrait les voxels correspondant au striatum et a décomposé les données en utilisant une classification par moindre carré et un SVM. Rojas et al. ont travaillés sur l'influence de la sélection des voxels. Ils ont utilisés un SVM (avec toutes les caractéristiques) et un autre SVM avec une sélection par analyse des composantes principales (PCA) au préalable. La méthode avec PCA donne les meilleurs résultats. Enfin Towey et al. utilisent les informations de tous les voxels et extrait les caractéristiques par décomposition en valeur singulière suivi par une classification SVM.

Ainsi, contrairement aux références mentionnées ci-dessus, ce papier propose différents avantages :

- Pas de perte de l'information car je ne fais pas de sélection de caractéristique
- Ma base de données est beaucoup plus grande
- Rajout de l'information sur la probabilité qu'un patient soit affecté par parkinson

Analyse statistique des caractéristiques

La première étape de tout projet de reconnaissance de forme est l'analyse des caractéristiques. Dans le cadre de mon projet, je n'ai pas réduit la dimensionnalité de ma base de données (dû au trop faible nombre de caractéristiques). Parmi les méthodes de sélection on retrouve parmi les plus citées dans la littérature PCA, ou encore Fisher.

J'ai commencé par calculer le taux de recouvrement inter-classes pour chacune de mes caractéristiques. Pour calculer ce taux, plusieurs méthodes existent, j'ai décidé d'appliquer la méthode du 1-NN, c'est-à-dire un K-nn à un seul voisin. Cette méthode a pour avantage de ne pas nécessité d'hypothèses à priori sur la modélisation de notre classe. Ensuite une fois le recouvrement calculée, j'ai utilisé le critère de Fisher qui donne une information sur la divergence entre les classes cette-fois–ci. Il est défini comme suit :

$$F_{ij} = \frac{|\mu_i - \mu_j|^2}{v_i^2 + v_j^2}$$

i et j correspondent respectivement aux 2 classes choisies. La moyenne μ a été calculée en utilisant la formule habituelle pour la moyenne d'une classe. La variance v^2 est calculée à partir de la formule d'un écart sans biais, $\bar{x}$ étant l'estimation de la moyenne :

$$v^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Le résultat nous donne un scalaire, plus celui-ci est élevé meilleure est la divergence.

Par soucis de lisibilité, j'ai aussi tracé la fonction de densité des classes, pour cela j'ai utilisé les « Parzen Windows ». Ces fenêtres glissantes vont permettre de modéliser une classe sans en connaître la loi qui régit la classe. J'ai utilisé un noyau gaussien, avec comme volume de boule :

$$V_{nGaus} = 0.05$$

Ainsi qu'un noyau carré :

$$V_{nBox} = 1/\sqrt{n}$$

Couplé à des « boites à moustache » qui vont permettre de rajouter un œil critique aux résultats de façon assez claire. Vous retrouverez pour chacune des caractéristiques en haut mes résultats et en bas ceux de l'article, par soucis de comparaison.

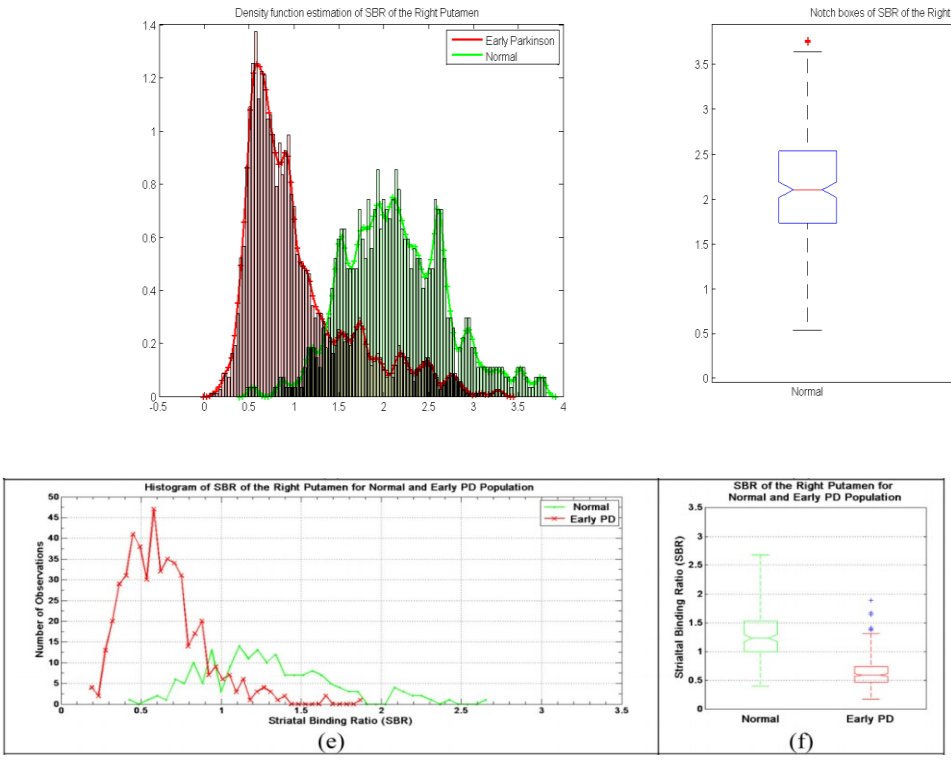


Figure 4 - SBR du Putamen droit

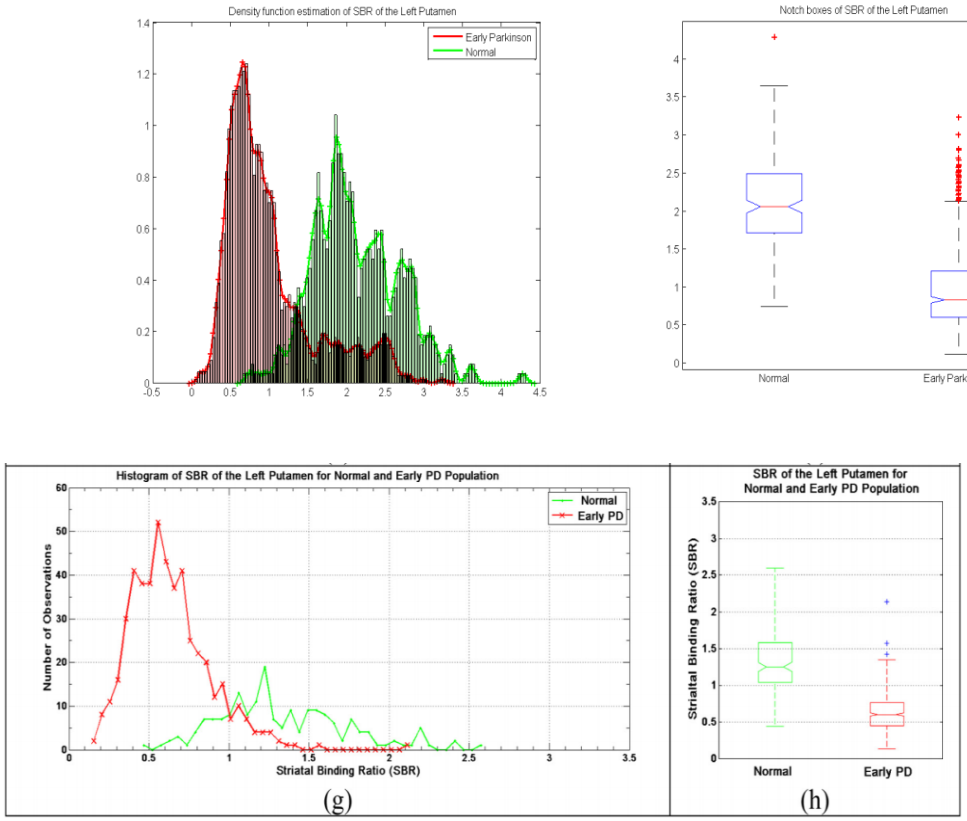


Figure 5 - SBR du Putamen gauche

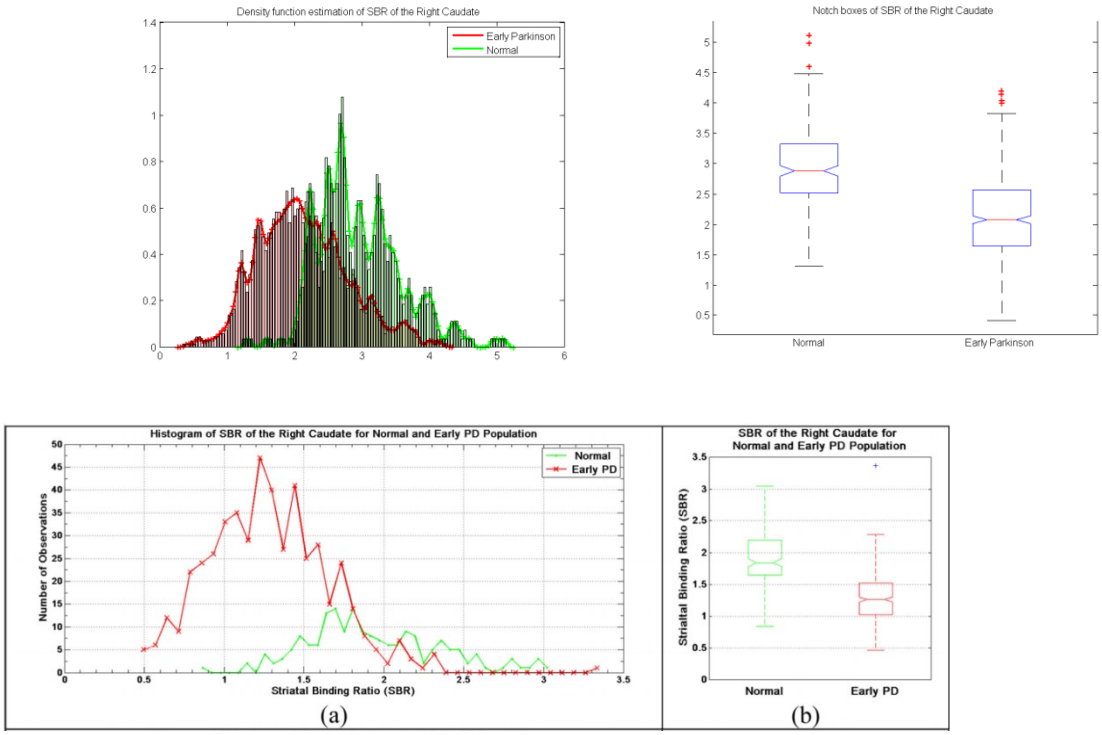


Figure 6 - SBR du Caudate droit

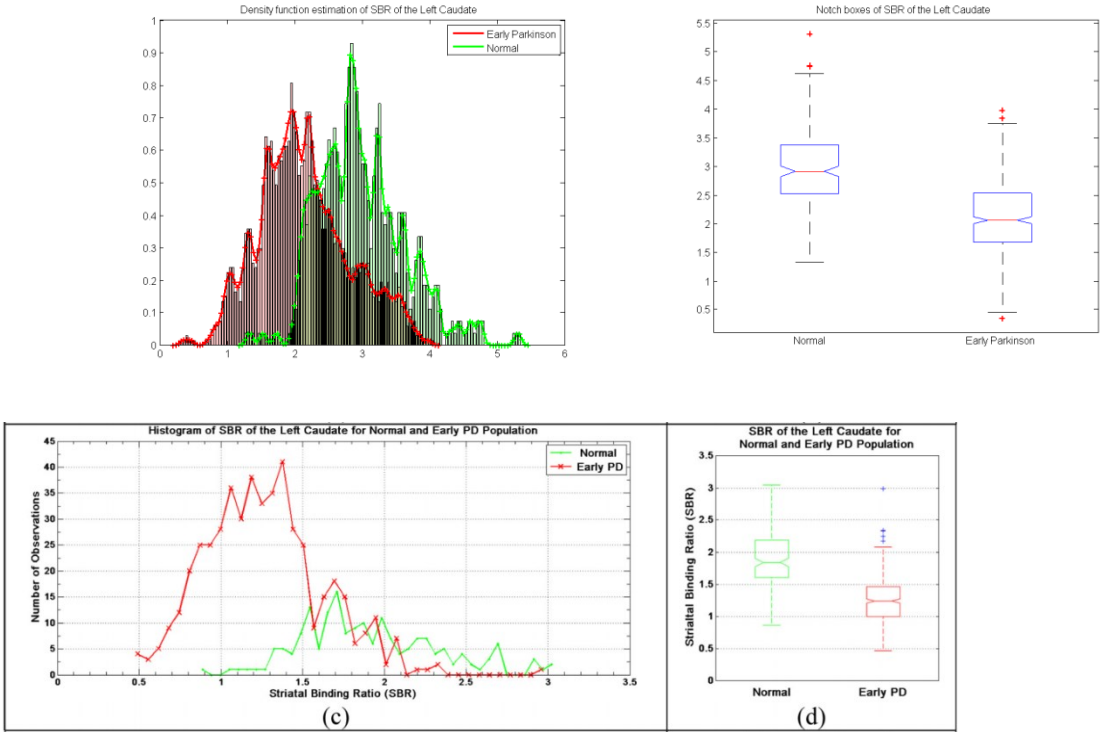


Figure 7 - SBR du Caudate gauche

	Caudate droit	**Caudate gauche**	**Putamen droit**	**Putamen gauche**
Recouvrement	27.5815%	28.6685%	20.7880%	21.4674%
Divergence	0.8157	0.8447	1.7557	1.8807

On remarque globalement que les caractéristiques sur le Putamen sont meilleures que celles sur le Caudate. La divergence est assez critique pour le Caudate ce qui peut se voir sur la modélisation des 2 classes. Globalement, par rapport à l'article, on remarque que les boîtes à moustache ont tendance à être plus rapprochées pour mes données.
Néanmoins les résultats restent satisfaisants et permettent de confirmer la divergence des caractéristiques.

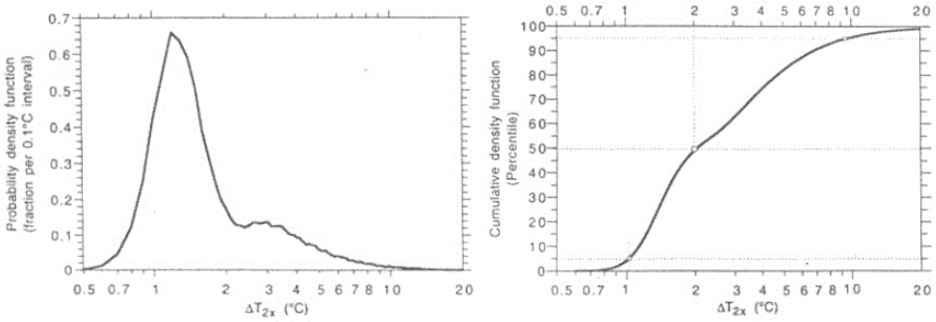
Pour la suite de ce papier, je détaillerai chaque classifieur que j'ai implémenté. En commençant à chaque fois par une explication mathématique de l'ensemble. Je passerai ensuite au design du classifieur et la validation.

Régression logistique multinomiale

La régression logistique est un modèle de régression binomiale (réponse 1 ou 0). On veut prédire la probabilité d'un évènement binaire (parkinson ou non) en connaissant les variables aléatoire (caractéristiques SBR du patient). Cette méthode est très répandue en médecine pour trouver des facteurs qui caractérisent un groupe de malade par exemple [7].

Le but est d'étudier non plus la fonction de densité mais son intégrale, c'est-à-dire la fonction de répartition (la probabilité) :

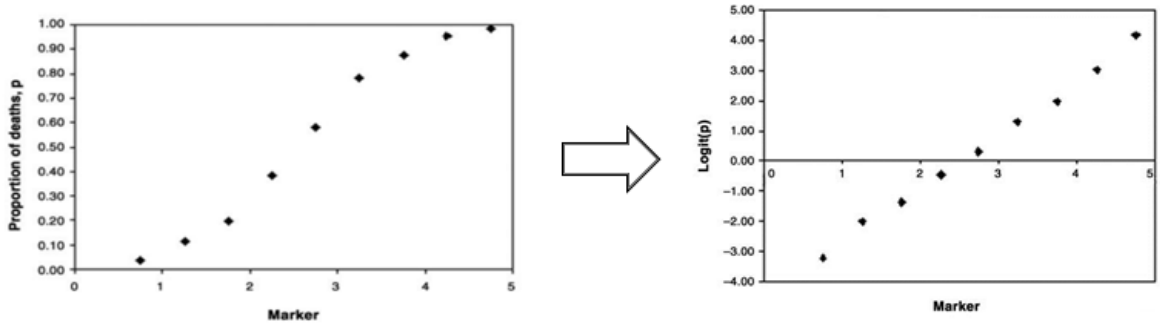
$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx$$



On retrouve en abscisse le bio-marqueur concerné qui sera dans mon cas la caractéristique SBR. La probabilité que le patient définit par les caractéristiques X soit atteint de la maladie de parkinson est appelé le Likelihood:

$$\pi = P(y = 1|X)$$

L'idée est de linéariser la fonction de répartition pour récupérer facilement les probabilités, pour cela on passe par le modèle dit « logit ».



$$\text{Logit}(\pi) = \ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta x$$

Comme on est dans un cas multivarié, l'équation devient pour un patient i :

$$\text{Logit}(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_1 \times SBR_{RCi} + \beta_2 \times SBR_{LCi} + \beta_3 \times SBR_{RPi} + \beta_4 \times SBR_{LPi}$$

α et β_k sont calculés en utilisant la méthode du maximum de vraisemblance de Likelihood. On fait varier les paramètres α et β_k et on cherche la probabilité maximale par méthode itérative de recherche de maximum. Une fois les paramètres trouvés, la probabilité qu'un patient i soit atteint de Parkinson est de :

$$\pi_i = \frac{1}{1 + e^{-logit(\pi_i)}}$$

La première étape est d'analyser statistiquement la qualité de nos prédicteurs. Pour cela j'ai travaillé sur toute ma base de données. L'analyse statistique permet de conclure sur la qualité de nos prédicteur, au début je me suis concentré sur un seul paramètre qui est la p-valeur :

Prédicteur	$\boldsymbol{\beta}$	$\boldsymbol{p}$**-valeur**
Constante	3.8529	$2.6157e^{-14}$
Right Caudate	0.2561	0.4989
Left Caudate	0.6056	0.1173
Right Putamen	-1.3250	$3.4438e^{-4}$
Left Putamen	-1.9993	$2.0568e^{-7}$

[1]

On voit que pour les SBR du Caudate, on a des p-valeurs extrêmement élevés (on veut habituellement $p < 0.05$). L'idée est alors de multiplier ces valeurs entre elles et espérer obtenir une meilleure p-valeur ($0.4989 \times 0.1173 = 0.0585$ ce qui est déjà plus acceptable).

Finalement, la base de données finale ne contient plus que 3 caractéristiques :

- $SBR_{RC} \times SBR_{LC}$
- SBR_{RP}
- SBR_{LP}

J'ai donc analysé mes nouveaux prédicteurs avec cette nouvelle base de donnée :

Prédicteur	$\boldsymbol{\beta}$	$\boldsymbol{p}$**-valeur**	**Erreur standard**	$\boldsymbol{t}$**-Test de Student**
Constante	3.8529	$2.6157e^{-50}$	0.3313	14,8753
Right Caudate × Left Caudate	0.2561	$8.2207e^{-4}$	0.0559	3,3453
Right Putamen	0.6056	$2.7789e^{-5}$	0.3435	4,1909
Left Putamen	-1.3250	$1.9637e^{-8}$	0.3528	5,6152

[1] Les résultats correspondent à la probabilité $P(y = 0|X)$ et donc la probabilité que le patient soit normal

On remarque que cette fois-ci, les p-valeurs sont heureusement plus faibles que précédemment, on reste en dessous du seuil fatidique de 5%. Les erreurs ne sont pas si élevées, les marges de valeur sur les prédicteurs ne sont ainsi pas beaucoup impactées.
On peut aussi calculer le degré de liberté, qui nous donne une information sur le nombre d'équations en plus que l'on a. Il est de 732 ce qui est assez élevé mais logique au vu de notre base de donnée (736 patients) pour 4 prédicteurs. Le fait d'avoir remplacé 2 caractéristiques par 1 seule augmente le degré de liberté ce qui peut s'avérer néfaste au niveau algorithmique.

Le t-Test ou la statistique de $Wald$ donne une information sur l'importance du prédicteur, plus cette valeur est élevée, plus le prédicteur est important pour la prédiction finale. En utilisant la table de la loi de $Student$ avec 732 degrés de liberté, on remarque que les prédicteurs ont tendance à ne pas être si performant que ça.

La matrice de corrélation permet de nous dire si nos prédicteurs sont bien indépendants entre eux :

Correlation Matrix	β_{cte}	$\beta_{RC\times LC}$	β_{RP}	β_{LP}
β_{cte}	1	$0{,}0776$	$-0{,}2569$	$-0{,}2945$
$\beta_{RC\times LC}$	$0{,}0776$	1	$-0{,}3581$	$-0{,}3773$
β_{RP}	$-0{,}2569$	$-0{,}3581$	1	$-0{,}5902$
β_{LP}	$-0{,}2945$	$-0{,}3773$	$-0{,}5902$	1

Elle donne des résultats satisfaisant (surtout pour la constante) mais on remarque que le prédicteur du Left Putamen à tendance à dépendre au Right Putamen (ce qui est physiquement logique si les marqueurs sont situés dans la zone entre les 2).

Enfin, le coefficient de détermination $R^2 = 0.8756$ (idéalement 1) indique que nos données ont bien tendance à être incluses par la régression.

Cette analyse démontre la qualité globale des prédicteurs, pour étudier la qualité du classifieur, j'ai choisi d'utiliser une validation croisée. En fait on découpe notre base de donnée en k boites, dans chaque boite le dernier élément sert de test. Les premiers eux permettent d'entrainer l'algorithme (récupérer β_k et α). J'ai choisi d'utiliser 50 boites ce qui permet d'avoir 50 tests. [2]

[2] La taille des boîtes influe sur le résultat et montre l'impact du « training » sur le classifieur. Ainsi dans le cas du Logit, on remarque que moins il y a de boîtes (donc chaque boite possède plus de points) meilleurs sont les résultats.

Pour analyser les résultats, une méthode courante est l'utilisation de la matrice de confusion. Cette matrice est très intéressante car elle permet de voir concrètement les erreurs types de notre classifieur. Il est essentiel pour l'étape de comparaison de classifieurs.

Label/ prediction	Normal	Parkinson
Normal	10	3
Parkinson	2	35

Figure 8 - Résultat du MLR

30 % d'erreurs sur les normaux

5.71% d'erreurs sur parkinson

Total de 11.11 % d'erreur

Le fait de multiplier ensemble les 2 SBR du Caudate améliore les performances de quelques pourcents globalement. J'ai utilisé un « cuttoff » de 0.5, c'est-à-dire que lorsque ma probabilité retourné est au-dessus de ce seuil, alors je classe le patient en tant que Parkinson.

Il peut aussi être intéressant d'analyser les différentes fonctions de répartition, ci-dessous un exemple pour le SBR du Putamen droit :

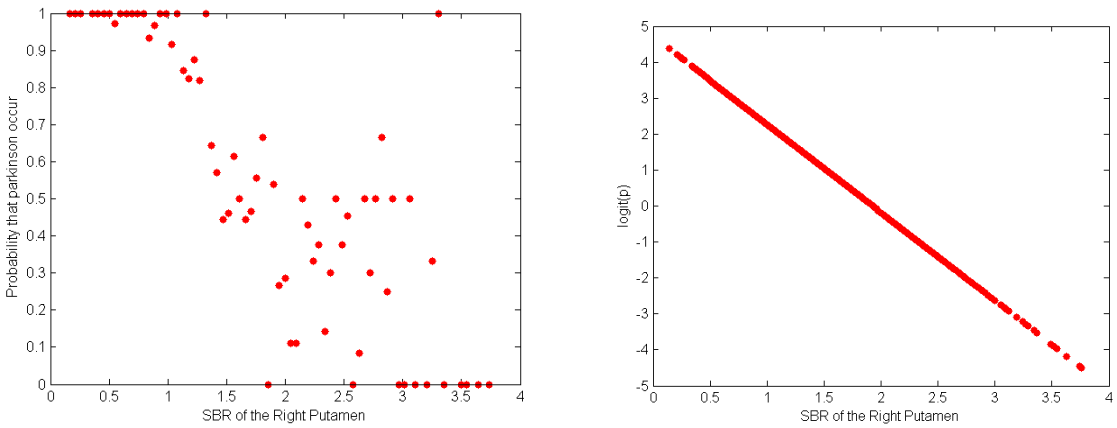


Figure 9 - Fonctions de répartition

Ainsi on remarque que dû au recouvrement entre les classes, la fonction de répartition à tendance à avoir un comportement assez chaotique. Ce qui explique un taux d'erreur au-dessus de 10%.

Support Vector Machines

SVM est une technique d'apprentissage supervisé qui offre une méthode pour définir l'hyper plan de décision, ceci sans supposition de distribution à priori de nos classes. L'idée est de trouver l'hyperplan qui maximise la marge, c'est-à-dire l'hyperplan qui « fait s'éloigner » les points les plus critiques (appelés vecteurs de support). En effet, ce sont ces vecteurs de support qui pose le plus de problème. [8]

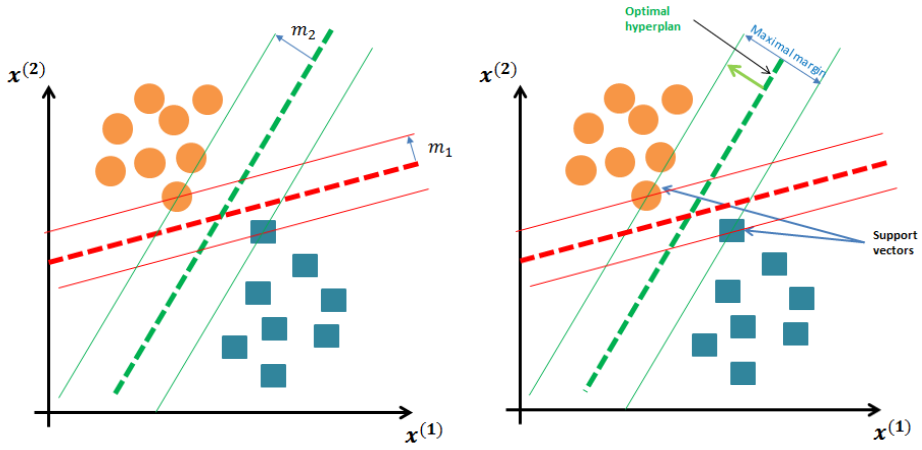


Figure 10 - Hyperplans de décision

Ainsi comme le montrent les figures ci-dessus, il est préférable de choisir la frontière verte comme hyper-plan de décision plutôt que la frontière rouge. Les avantages du SVM comparé à d'autres techniques est qu'il est flexible (avec l'utilisation de noyaux non-linéaires), on n'a pas besoin de connaitre la loi qui régit les densités et le SVM retourne une solution unique (le meilleur hyper-plan). [9]

L'équation de l'hyper plan est défini telle que :

$$w.x_i + w_0$$

SVM propose ainsi de résoudre le problème de minimisation suivant, avec C qui définit la pénalité de l'erreur :

$$\min \frac{1}{2} w^t w + C \sum_{y_i=0}^{n} \xi_i$$

Avec la contrainte (dû au recouvrement) :

$$y_i(w^t x_i + b) \geq 1 - \xi_i \,, \quad \xi_i > 0 \;\; \forall i$$

La décision y_i ∈{-1,1} sur la classe du point x est donnée par la fonction :

$$Classe(y) = sign(wx + w_0) = sign\left(\sum_{i=1}^{l} \alpha_i y_i \, x_i . x + w_0\right)$$

Si on veut projeter x_i dans un espace de caractéristique plus élevé, on peut utiliser la fonction noyau $\phi(x_i)$, la décision devient alors :

$$sign\left(\sum_{i=1}^{l} \alpha_i y_i \, K(x_i . x) + w_0\right), \quad K(x_i . x) = \phi(x)^t \times \phi(x_i)$$

$\phi(x_i)$ permet ainsi d'obtenir des frontières de décision non linéaire, comme par exemple avec un noyau gaussien.

Pour la suite, j'ai décidé d'utiliser plusieurs sortes de noyaux. Pour déterminer le meilleur noyau, il est tout d'abord nécessaire pour chacun de trouver les meilleurs paramètres. J'ai décidé d'effectuer ce choix des paramètres sur tout ma base de donnée, ainsi je prends $2/3$ de cette base qui servira de training pour le SVM et $1/3$ servira pour la validation.

Je ressors une erreur pour chaque paramètre, et c'est la plus petite erreur qui va me permettre de déterminer le meilleur paramètre. On retrouve en général 2 paramètres :

- Paramètre C
- Paramètres liés au noyau

C apparaît lorsque l'on a des problèmes de recouvrement. C'est lui qui va définir le compromis entre la marge maximale et les erreurs de classification appelées la pénalité de l'erreur. J'ai choisi de faire varier C de $1e^{-5}$ à $1e^{5}$ ce qui permet d'avoir une idée assez globale de la meilleure valeur. [10]

Pour définir les meilleurs paramètres pour chaque noyau, je fais à chaque fois varier tous les paramètres un à un. Je stocke les erreurs de classification ressorties par le SVM dans une hyper-matrice d'erreur de dimension $[\, n_1 \times n_2 \times \ldots \times n_i \,]$, avec n le nombre de variations sur les paramètres et i le nombre de paramètres (Figure 11).

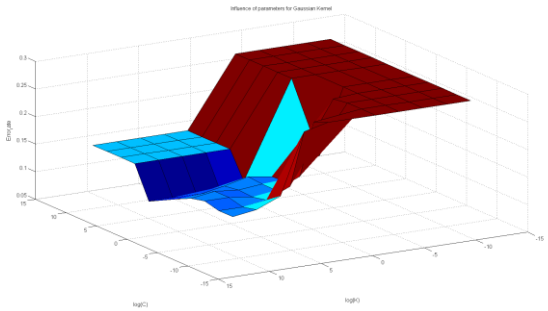


Figure 11 – Hyper-matrice pour le SVM gaussien

J'ai opté pour 4 noyaux différents pour la suite.

Noyau linéaire :

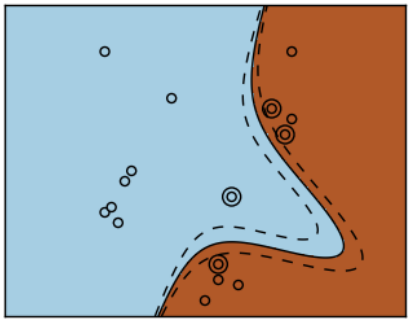
$$K(x_i, x) = x^t . x_i$$

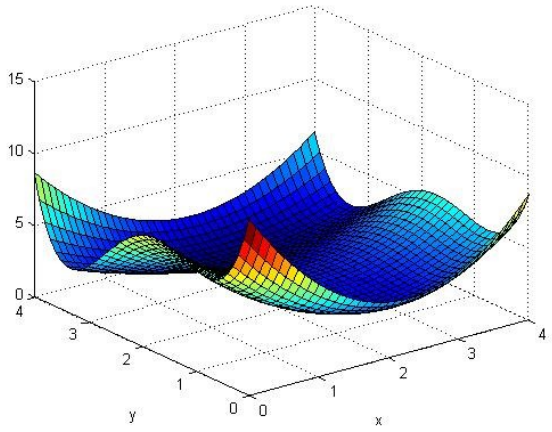
C'est le SVM de base où aucune transformation sur notre espace de caractéristique n'est effectuée. Il n'y a pas de paramètre spécifique lié au noyau, le seul paramètre est C.

Noyau polynomial :

$$K(x_i, x) = (\alpha \times x^t . x_i + c\ )^d$$

C'est un noyau non-stationnaire qui est plutôt utilisé dans des problématiques de régression où les données sont normalisées.





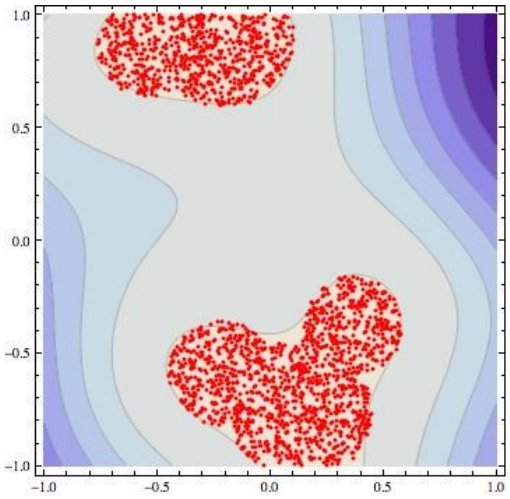
On retrouve 3 paramètres ajustables :

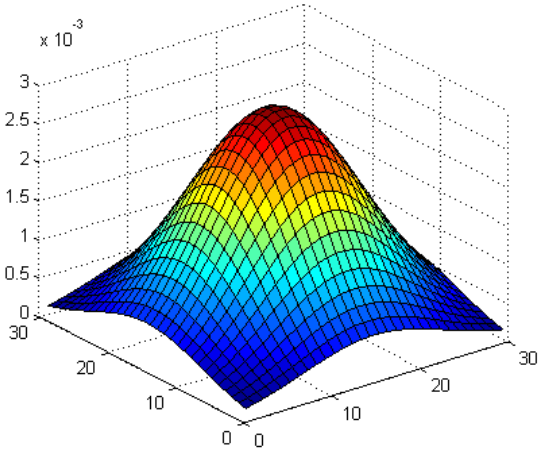
- Le degré du polynôme d de 1 à 10
- L'offset c de $1e^{-5}$ à $1e^{5}$
- Le facteur α de $1e^{-5}$ à $1e^{5}$

Noyau gaussien :

$$K(x_i, x) = \exp(-\gamma \|x - x_i\|^2)$$

C'est un des noyaux les plus populaires car il permet de classifier sous des formes d'ellipses.



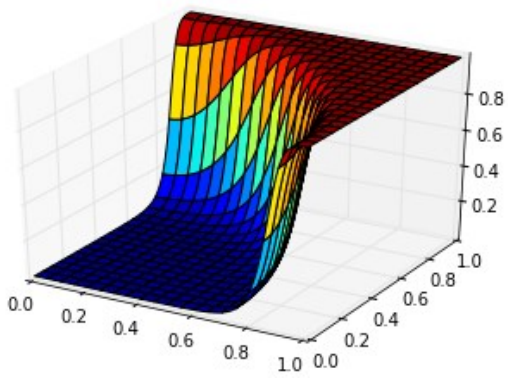


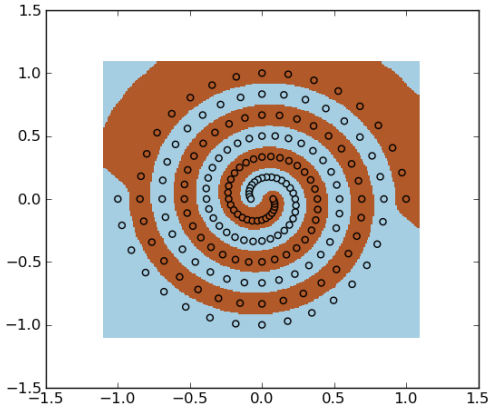
On ne retrouve qu'un seul paramètre ajustable ce qui en fait un noyau relativement simple à analyser. Il s'agit de la constante $\gamma > 0$ que j'ai choisi de faire varier de $1e^{-5}$ à $1e^{5}$.

Noyau sigmoid :

$$K(x_i, x) = \tanh(\alpha \times x^t . x_i + c)$$

Il vient des réseaux de neuronnes pour permettre leur activation et l'utilisation de ce noyau revient au perceptron d'un réseau de neuronne à 2 couches. Il est indiqué qu'il offre de bons résultats en pratique. [8]



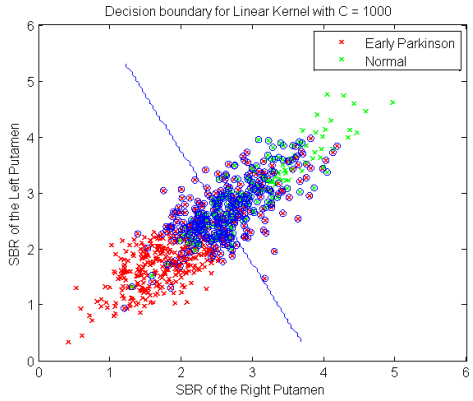


On retrouve 2 paramètres ajustables :

- L'offset c de $1e^{-5}$ à $1e^{5}$
- Le facteur α de $1e^{-5}$ à $1e^{5}$

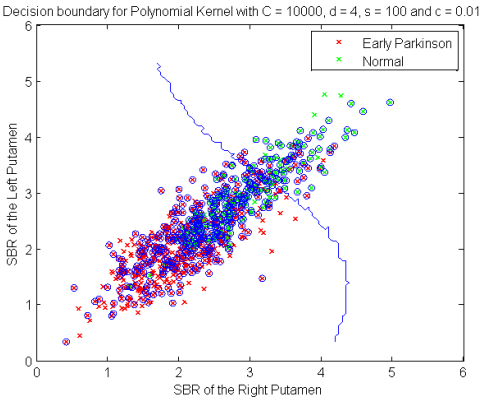
Pour valider les différents noyaux, j'utilise encore une fois une validation croisée avec 50 boîtes. Voici les différents résultats pour chacun des noyaux :

Linéaire :



Label \ prediction	Normal	Parkinson
Normal	10	2
Parkinson	4	33

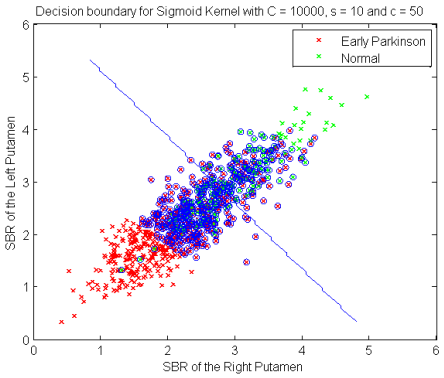
Polynomial :



Label \ prediction	Normal	Parkinson
Normal	10	3
Parkinson	1	36

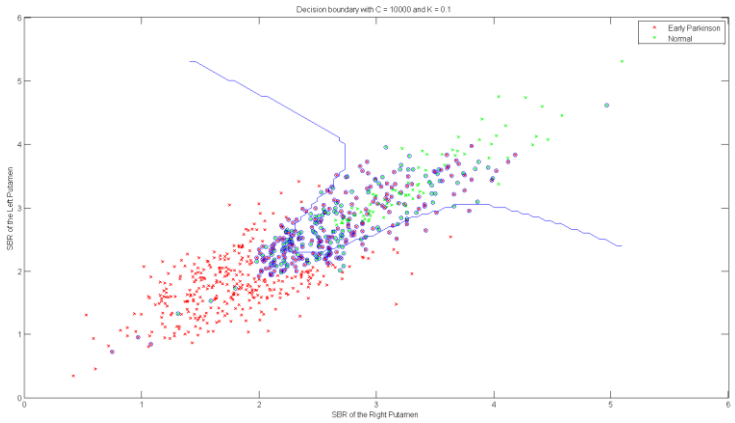
Figure 12 - Résultats du SVM linéaire et polynomial

Sigmoid :



Label \ prediction	Normal	Parkinson
Normal	10	3
Parkinson	2	35

Gaussien :



Label \ prediction	Normal	Parkinson
Normal	11	2
Parkinson	1	36

Figure 13 - Résultats du SVM Sigmoid et gaussien

Au final, c'est le noyau gaussien qui offre les meilleurs résultats, avec un taux d'erreur des plus faibles (6%). On remarque de façon générale que le SVM à tendance à prendre beaucoup de vecteurs de support (entourées en bleu), ce qui explique les très long temps de calcul pour le training. Pour ce qui est du noyau gaussien, le fait d'augmenter les paramètres conduit à surentraîner le classifieur SVM, comme le montre la Figure 14.

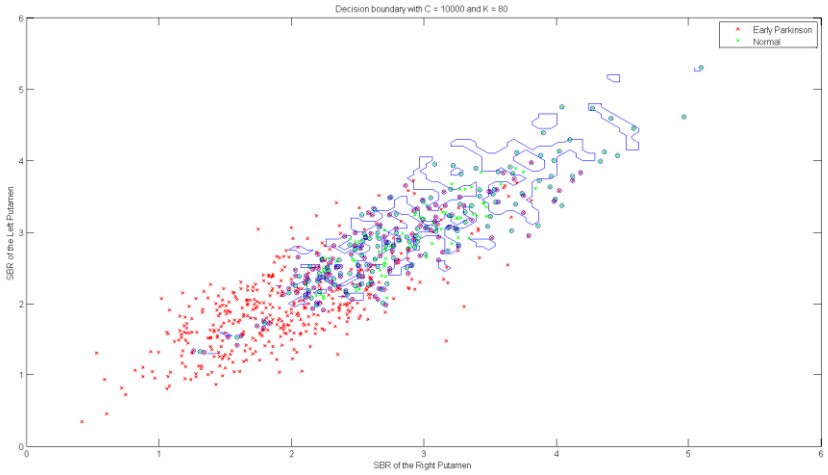


Figure 14 - Surentraînement du SVM

On peut aussi remarquer l'influence de la base de données sur les frontières de décision. Comme notre base contient plus de Parkinson que de patients normaux, la frontière a tendance à être du côté des Parkinsons pour « équilibrer » les zones. Cette frontière peut être modifié en jouant sur les paramètres, pour classer plus de normaux par exemple.

Classifieurs Bayésien

Pour le classifieur bayésien, il y a plusieurs il y a plusieurs suppositions à effectuer :

- On connait l'information à priori $P(\omega_j)$ qui va nous aider à la décision
- Les densités conditionnelles des classes sont connues (dans le cas paramétrique)
- On suppose aussi que l'on arrive à mesure notre vecteur de caractéristique X

La règle de Bayès nous dit que :

$$P(\omega_j|X) = \frac{p(X|\omega_j) \times p(\omega_j)}{p(X)}$$

$$p(X) = \sum_{i=1}^{n} P(X|\omega_i) \times P(\omega_i)$$

Il y a plusieurs approches pour modéliser notre densité de probabilité $P(X|\omega_j)$, l'approche paramétrique ou non-paramétrique.

Dans l'approche paramétrique, on connaît la loi qui régit nos classes, on supposera ainsi une loi normale d'après la forme des densités (confirmé par un test t de Student). En pratique (par soucis de complexité algorithmique) on préfèrera calculer la fonction discriminante associée à chaque classe. Ainsi, dans l'hypothèse à priori cas où toutes les matrice de covariance forme une hyper-sphère $\Sigma_i = \sigma^2 I$:

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i) - \frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_i| + \log P(\omega_i)$$

Dans le cadre du projet, le résultat fonctionne comme une probabilité. On suppose que le point associé au patient (les caractéristiques SBR) qui donne le $g_i(x)$ le plus élevé correspond à la classe i associée.

Les matrices de covariance Σ_i et les moyennes μ_i sont calculées à partir des estimateurs les plus performants. Ces estimateurs sont donnés par la méthode du maximum de vraisemblance et on retrouve pour une loi normale $\mathcal{N}(\theta_1, \theta_2)$:

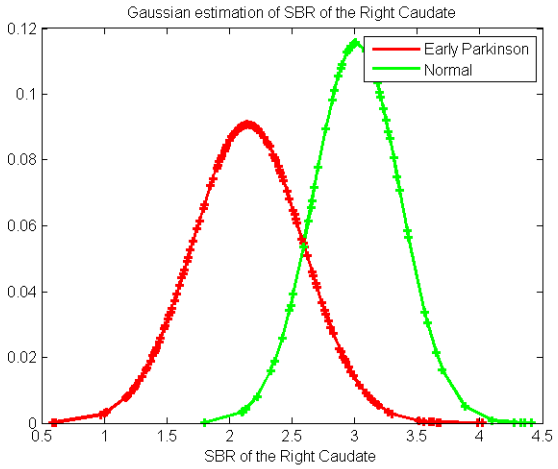
$$\widehat{\theta_1} = \frac{1}{n}\sum_{i=0}^{n} X_i \;\; ; \quad \widehat{\theta_2} = \frac{1}{n-1}\sum_{i=0}^{n}(X_i - \bar{X})^2$$

Dans l'approche non-paramétrique, il est nécessaire de définir notre loi. On suppose alors que cette loi « n'existe pas » mathématiquement parlant, et on va la définir nous-même. Pour estimer la densité de notre loi une méthode existe, les fenêtres de Parzen. Dans la veine d'une convolution en analyse fréquentielle, l'idée est d'utiliser une fenêtre glissante qui va estimer notre loi au fur et à que

cette fenêtre avance. Cette fenêtre, appelé aussi noyau, peut-être de forme différente.
Pour le projet j'ai utilisé un noyau gaussien, avec un noyau de largeur défini par $1/\sqrt{n_{train}}$. Cette largeur est le paramètre optimal, et a été testé sur l'ensemble d'apprentissage défini par $n_{train} = 2n/3$. Le noyau carré a aussi été testé mais retournait de trop faibles résultats.

Une fois la densité calculée, il ne reste plus qu'à calculer la probabilité correspondante à chacune des classes.

Pour l'estimateur Bayésien paramétrique, j'ai utilisé $2/3$ de ma base de donnée en tant qu'apprentissage (pour estimer les paramètres de la gaussienne). Le reste servira de test et retourne les résultats suivants :



label/ prediction	Normal	Parkinson
Normal	132	9
Parkinson	55	295

Figure 15 - Résultats du Bayes paramétrique

Par soucis de lisibilité, je n'ai tracé que le SBR du Caudate droit (qui possède le recouvrement le plus élevé), les densités correspondent donc bien évidemment à des gaussiennes :

$\bar{X}$	Right Caudate	Left Caudate	Right Putamen	Left Putamen
Normal	2.911	2.919	2.095	2.096
Parkinson	2.125	2.123	1.099	1.029

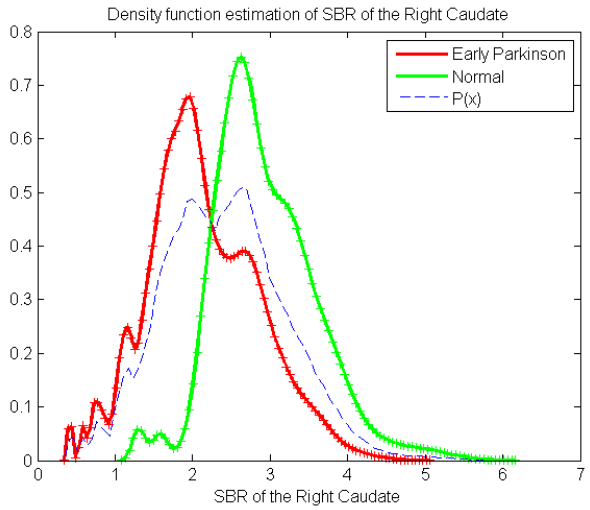
Au niveau des matrices de covariance :

Cov Normal	Right Caudate	Left Caudate	Right Putamen	Left Putamen
Right Caudate	0.330	0.312	0.246	0.219
Left Caudate	0.312	0.367	0.268	0.229
Right Putamen	0.246	0.268	0.322	0.259
Left Putamen	0.219	0.229	0.259	0.276

Cov Parkinson	Right Caudate	Left Caudate	Right Putamen	Left Putamen
Right Caudate	0.473	0.389	0.353	0.272
Left Caudate	0.386	0.467	0.314	0.328
Right Putamen	0.353	0.314	0.406	0.319
Left Putamen	0.272	0.328	0.319	0.365

On remarque à partir des matrices que les différentes caractéristiques ont bien tendance à être indépendantes entre elles. Il y a une certaine redondance entre les variances du même type (Caudate ou Putamen). De plus les variances entre Putamen/Caudate par exemple sont heuresement plus faibles qu'entre Caudate/Caudate. Ceci peut se comprendre physiquement, les 2 zones n'étant pas les même, les caractéristiques ont donc tendance à être plus indépendantes au niveau de la variation des bio-marqueurs.
On retrouve un taux d'erreur à 13.03% ce qui est relativement assez faible au vu de la simplicité de la technique.

Pour l'estimateur Bayésien non-paramétrique, j'ai utilisé une fois de plus $2/3$ de ma base comme entraînement (pour estimer mes 2 lois). Le reste de ma base retourne les résultats suivants :



label/ prediction	Normal	Parkinson
Normal	105	25
Parkinson	31	330

Figure 16 - Résultat du Bayes non-paramétrique

Le tracé de la probabilité $P(X)$ correspond à la probabilité qu'un patient appartienne à une des deux classes. Elle est intéressante pour l'analyse car elle permet de voir comment se répartissent les patients au niveau de leurs critères SBR. Il permet aussi de confirmer le choix de 2 classes.
Le taux d'erreur ressorti par ce classifieur est de 11.41% .

Comparaison des classifieurs

Après avoir implémenté tous ces classifieurs, il est nécessaire de les comparer pour en déduire le plus performant. Pour cette partie j'ai décidé de comparer les classifieurs selon plusieurs points à priori indépendants :

- Bien évidemment le taux d'erreur de l'algorithme
- Le temps de calcul pour tester un point rentrant
- La complexité algorithmique, ainsi que la difficulté de compréhension et de mise en place
- La consommation en mémoire

Tous les résultats peuvent être récapitulés dans le tableau ci-dessous :

	SVM	MLR	Bayes param	Bayes non param
Error rate	6.0% (++)	11.11% (+)	13.03% (−)	11.41% (+)
Time	(−−)	(=)	(++)	(+)
Complexity	(−−)	(−)	(++)	(+)
Memory consumption	23 MB (in MATLAB) (=)	31 MB (−)	9 MB (++)	50 MB (−−)

Figure 17 - Comparaison des différents classifieurs

Si l'on décide que le seul et unique critère qui compte est le taux d'erreur, alors c'est vraisemblablement le SVM qui remporte la partie. Avec 6% de taux d'erreur, le SVM gaussien est plus performant de moitié comparé aux autres classifieurs.
Néanmoins, il est nécessaire de prendre en compte que c'est le classifieur le plus complexe et le plus long à traiter. C'est dans ces domaines que la Bayès paramétrique peut être intéressant, en effet mis à part le taux d'erreur, il reste le meilleur classifieur sur tous les autres critères. Il pourrait ainsi être intéressant de l'appliquer en amont pour une supposition rapide sur l'état du patient.
Les 2 autres classifieurs sont assez modestes dans l'ensemble et ont l'avantage tous les deux d'offrir la probabilité d'apparition de la maladie au clinicien. Soulignons tout de même la grande consommation en mémoire du Bayes paramétrique, due au transport de toutes les estimations de densité.

Au niveau des matrices de confusion que vous pouvez retrouver Figure 8, Figure 12, Figure 15 et Figure 16, on remarque que l'erreur commune est de classer des normaux en tant que Parkinson. Ceci est lié au fait qu'il y a plus de Parkinson dans la base de données, le résultat est donc biaisé par la probabilité à priori. C'est tout de même le SVM qui commet le moins d'erreur, là où la Bayes paramétrique classe plus de Parkinson en tant que normal. Le classifieur non-paramétrique classe plus de normaux en tant

que Parkinson et la régression logistique est dans la moyenne. Ce que l'on peut souligner avec le SVM, c'est sa capacité à être amovible de par ses noyaux et ses paramètres. Ainsi, en modifiant les paramètres γ et C du noyau gaussien on peut jouer sur la forme de la frontière :

- Vaux-il mieux classer à tord des malades en tant que normaux ? (la frontière est du côté des Parkinsons : $C = 10\,000,\ \lambda = 0.1$)
- Ou accepter de classer des normaux en tant que malades ? (on décale la frontière dans la zone des normaux : $C = 8500,\ \lambda = 0.7$)

J'ai choisi la deuxième option, en effet classer un malade en tant que normal est beaucoup plus dangereux que l'inverse.

Une dernière chose intéressante que l'on peut souligner, est le fait que comme ces classifieurs sont **à priori** indépendant l'un de l'autre on peut tout à fait tous les prendre en compte. On pourrait ainsi procéder par vote, si 3 des classifieurs donnent le même résultat, alors la probabilité que ce soit effectivement ce résultat est assez élevée. On pourrait imaginer une intelligence artificielle basé sur l'organigramme suivant :

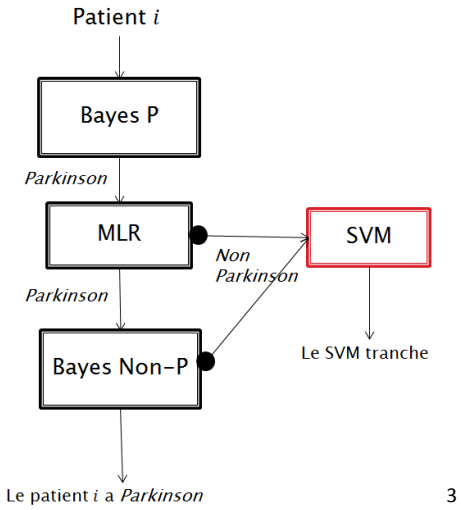


Figure 18 - IA d'apprentissage machine

Avec cet exemple, la probabilité d'erreur lorsqu'un patient est passé par 3 blocs est d'au plus $13.03\% \times 11.41\% \times 11.11\% = 0.1651\ \%$ (tous les blocs noirs). Ce qui permet de diminuer drastiquement le taux d'erreur au détriment du temps de calcul (car on utilise plusieurs classifieurs). Bien évidemment avec l'hypothèse que tous les classifieurs soient complètement indépendants (ce qui nécessite une étude approfondie). Cette méthode permettrait de gommer le désavantage du SVM (temps de calcul) et lui permettre de donner le dernier mot en cas de litige. On pourrait aussi imaginer d'autres blocs en utilisant d'autres types de noyaux pour le SVM par exemple.

[3] Si le premier classifieur bayésien ressort que le patient n'a pas Parkinson, il suffit d'inverser tous les états de l'organigramme.

Conclusion

L'article qui m'a inspiré pour ce projet a été très intéressant sur le pan technique et sur ma culture générale. Il m'a permis de mettre un premier pied dans l'univers du biomédical, où je souhaite m'orienter pour le futur. J'ai aussi pu mettre en application beaucoup de techniques vues en cours, avec des implémentations concrète en *MATLAB*.

Au niveau de l'article, je tiens à souligner leurs différents résultats : Ils ont obtenu un taux d'erreur de 3.86% (j'ai obtenu 6%) pour le SVM gaussien, et un taux d'erreur de 9.2% (j'ai obtenu 11%) pour la régression logistique. Leurs matrices de confusion retourne le même genre d'erreur pour le MLR, mais ils ont tendance à avoir plus de « faux calme » (Parkinson en tant que normal) pour le SVM.

Label \ prediction	Normal	Parkinson
Normal	172	9
Parkinson	17	476

Label \ prediction	Normal	Parkinson
Normal	140	41
Parkinson	21	472

Figure 19 - Matrices de confusion pour SVM et MLR (article)

Concrètement mes résultats sont donc globalement moins bons mais ceci s'explique par plusieurs choses :

- Tout d'abord ma base de données est plus récente (de 2 ans) et je possède moins de Parkinson dans ma base (en gras). Ce qui explique que mes résultats sont moins bons puisqu'il est plus difficile de classer correctement des normaux lorsqu'il y a plus de Parkinson.

  163 ➔ **211 patients normaux**

  482 ➔ **525 patients parkinson**

  77.15% de PD ➔ **71.33% de PD**

- Dans leur base de données, on retrouve certains patients qui ont étés analysés plusieurs fois (à plusieurs étapes de la maladie de Parkinson). Les résultats de ces personnes ne sont donc plus indépendants (car ils dépendent de la même personne), on ne peut donc pas considérer plusieurs points différents. Dans ma base, j'ai trié ces cas et j'ai pris en compte que le stade le plus précoce de la maladie (lorsqu'il est plus difficile de détecter la maladie). C'est pourquoi j'ai plus de recouvrement et donc de moins bons résultats.
- Enfin, ils entraînent leurs SVM et MLR sur toute la base de données ce qui explique de meilleurs résultats que les miens. En effet j'effectue une validation croisée pour mes tests SVM et MLR.

Finalement, j'ai contribué à l'article sur plusieurs points :

- Tout d'abord le fait de tester sur une base plus récente permet de valider (ou non) les méthodes proposées.
- J'ai testé d'autres classifieurs en montrant leurs avantages et désavantages. Qui peuvent tout à fait s'insérer dans la boucle de classification.
- Plus particulièrement pour le SVM, j'ai confirmé le choix des auteurs quant au noyau gaussien ; Je propose aussi une méthode pour gérer ses erreurs, en choisissant d'avoir plus de « faux calme » ou plus de « fausse alarme ».
- Je propose de mixer 4 classifieurs et procéder par vote, en proposant un organigramme adéquat pour ce travail (avec plusieurs hypothèses associées).

Enfin, on pourrait imaginer divers axes de recherche pour compléter on travail notamment :

- Tester les différents algorithmes avec une librairie très fiable. Car j'ai en majeure partie utilisé *MATLAB* pour récupérer mes résultats, ce qui n'est pas forcément le plus rigoureux.
- Etudier l'impact de la taille de la base de données sur les résultats. En jouant sur la taille des boîtes des validations croisées par exemple.
- Analyser statistiquement l'indépendance des classifieurs pour confirmer la méthode proposée.
- Rajouter d'autres classifieurs si l'indépendance est bien confirmée, pour avoir un résultat beaucoup plus fiable.
- Tester une toute nouvelle approche graphique (par graphe ou grammaire) en travaillant directement sur les images des patients.
- Créer une application de niveau commerciale en *C/C++* (avec les bonnes optimisations algorithmiques).

Table des figures

Références

[1] S. P. Canada, «Parkinson,» Décembre 2014. [En ligne]. Available: http://www.parkinson.ca/site/c.jpIMKWOBJoG/b.5187667/k.4DB6/Qu8217estce_la_maladie_de_Parkinson.htm. [Accès le Novembre 2014].

[2] K. e. a. Marek, «Long-term follow-up of patients with scans without evidence of dopaminergic deficit (SWEDD) in the ELLDOPA study,» *Neurology,* vol. A274, n° %164, 2005.

[3] R. e. a. Prashanth, «Automatic classification and prediction models for early Parkinson's,» *Expert Systems with Applications,* vol. 41, p. 3333_342, 2013.

[4] F. e. a. Segivia, «Improved parkinsonism diagnosis using a partial least squares based,» *Medical Physics,* n° %139, p. 4395–4403, 2012.

[5] D. e. a. Towey, «Automatic classification of 123I-FP-CIT,» *Nuclear Medicine Communications,* n° %132,, p. 699–707, 2011.

[6] A. e. a. Rojas, «Application of empirical mode decomposition (EMD) on DaTSCAN,» *Expert Systems with Applications,* n° %140, p. 2756–2766, 2013.

[7] V. e. a. Bewick, «Statistics review 14 : Logistic regression,» *Critical Care,* vol. 9, pp. 112 - 115, 2005.

[8] scikit-learn developers, «Support Vector Machines,» Septembre 2014. [En ligne]. Available: http://scikit-learn.org/stable/modules/svm.html. [Accès le Novembre 2014].

[9] L. e. a. Auria, «Support Vector Machines (SVM) as a Technique for Solvency Analysis,» German Institute for Economic Research, Berlin, 2008.

[10] K. Chin, «Trade-off between Maximum Margin and Classification Errors,» Septembre 1998. [En ligne]. Available: http://mi.eng.cam.ac.uk/~kkc21/thesis_main/node29.html. [Accès le Novembre 2014].