Extraction d'information via une méthode d'interprétation de scores

Vincent Lemaire *, Raphaël Féraud *

*France Telecom R&D - 2 avenue Pierre Marzin 22300 Lannion
vincent.lemaire@orange-ft.com

Résumé. Cet article présente une méthode permettant d'interpréter la sortie d'un modèle de classification ou de régression. L'interprétation de la sortie du modèle se base sur deux grandeurs : l'importance de la variable et l'importance de la valeur de la variable. Contrairement à la plupart des méthodes d'interprétation de l'état de l'art, notre approche permet d'interpréter la sortie du modèle pour chaque instance. La compréhension du score délivré par le modèle permet ainsi une aide à la décision immédiate, rapide, efficace et contextuelle. Un autre point fort de cette méthode d'interprétation est qu'elle ne dépend pas du modèle. La méthode proposée est intégrable dans n'importe quel processus ou logiciel de modélisation.

1 Introduction

Les outils de production de scores permettent de projeter sur une population déterminée une information quantifiable. Le score est l'évaluation pour chacune des instances d'une variable à expliquer ou variable cible. Le score est calculé par l'application d'un modèle sur les variables explicatives décrivant chacune des instances. Ces scores sont réinjectés dans le système d'information pour par exemple personnaliser la relation clients. Néanmoins la connaissance extraite sur un phénomène par le score n'est pas toujours exploitable directement. Par exemple, si un modèle identifie un client comme susceptible de passer à la concurrence, il ne dit rien sur l'action à entreprendre pour éviter son départ. Pour réagir, il est nécessaire d'identifier sa fragilité et les causes de cette fragilité. Nous proposons de traiter ce second point en interprétant la classification donnée par le modèle pour chacune des instances. Afin de rendre possible une industrialisation de cette connaissance, nous proposons une méthode entièrement automatique. La méthode proposée est indépendante du modèle utilisé pour construire le score. L'analyste peut utiliser le modèle le plus performant pour le phénomène étudié sans se soucier de la difficulté de son interprétation. Cette méthode d'interprétation pourrait ainsi lever un des freins principaux à l'utilisation dans les services marketing de modèles comme les Support Vector Machines, les Random Forest ou les réseaux de neurones.

2 Positionnement

Notations employées par la suite - Soit V_j : la variable explicative j, X : un vecteur de dimension J, K : le nombre d'instances, X_n : le vecteur représentant l'instance n, X_{nj} : la

composante j du vecteur n, F : le modèle, p : la sortie p du modèle, $F^p(X)$: la valeur de la sortie p du modèle pour le vecteur X et enfin $F_j^p(X_n; X_k)$ désigne la sortie p du modèle étant donné le remplacement de la composante j de l'instance X_n par celle de l'instance X_k.

2.1 Importance d'une variable

Le domaine de l'apprentissage automatique regorge aujourd'hui de techniques capables de résoudre efficacement des problèmes de régression et/ou de classification. Ces techniques construisent un modèle à partir d'une base de données d'apprentissage constituée d'un nombre fini d'instances (des couples de vecteurs d'entrée, sortie). Le modèle construit est utilisé pour associer à une instance, un vecteur, positionnée en entrée une réponse de sortie.

La grande variété des modèles (régression linéaire, réseaux de neurones, naïf bayes, random forest, ...) existant dans la littérature impose souvent d'avoir une méthode d'interprétation propre à chaque modèle. L'interprétation du modèle (si elle existe) est alors souvent basée sur les paramètres et la structure du modèle. On peut citer les interprétations des coefficients des paramètres du modèle (Arcadius et al., 2005), les interprétations géométriques (Iemma et Palm, 2003), les interprétations à base de règles (Thrun, 1995), ... Les interprétations résultantes sont souvent complexes, en moyenne, pour un modèle donnée, pour une tâche donnée (régression ou classification).

Une autre approche consiste à analyser le modèle telle une boite noire en faisant varier les entrées. Dans les simulations "What if ?", la structure et les paramètres du modèle n'ont de l'importance que dans la mesure où ils permettent de calculer avec une certaine précision les variables à expliquer en fonction des variables explicatives. Cette indépendance permet de proposer des méthodes d'interprétation valables quelque soit le modèle. La mesure d'importance est alors basée sur une analyse de la sensibilité des sorties du modèle. On observe la différence entre la sortie du modèle pour l'instance n et la sortie du modèle en faisant varier la variable V_j, d'une valeur h, pour cette instance n . Lorsque h tend vers zéro, cette mesure correspond à la dérivée partielle du modèle par rapport à la variable V_j. Dans ce cas, la mesure est locale. Elle peut conduire à une mesure d'importance erronée car la dérivée partielle en un point peut être nulle (voir l'exemple illustratif dans (Féraud et Clérot, 2002)) alors que la variable V_j est importante. Lorsque h est plus grand cette mesure peut être trompeuse si F n'est pas monotone. Le problème est le même quand les ces mesures sont moyennées sur tout l'ensemble de données (voir (Réfénes et al., 1994; Moody, 1994) pour les dérivées partielles et (Baxt et White, 1995) pour les différences).

(Féraud et Clérot, 2002) ont proposé une méthode permettant de mesurer l'importance d'une variable se basant sur l'intégrale des variations des sorties du modèle (la mesure n'est donc plus locale mais globale). La mesure ainsi obtenue est bien adaptée aux fonctions non monotones et permet de résoudre les inconvénients (supposition d'une fonction F monotone, utilisation d'extremums non robustes) de l' intéressante méthode d'interprétation proposée dans (Främling, 1996) . Le principal défaut de la méthode de Féraud et al. est qu'elle ne tient pas compte de la distribution des instances pour définir l'intervalle d'intégration. Leo Breiman (Breiman, 2001) corrige cet aspect en montrant comment la prise en compte de la distribution des données peut se révéler intéressante. Les aspects positifs de (Féraud et Clérot, 2002) et (Breiman, 2001) ont été associés dans une autre méthode de mesure d'importance des variables dans (Lemaire et Clérot, 2004). Cette méthode se base sur l'intégrale des variations des sorties du modèle étant donné la distribution empirique des données. Elle a été positionnée et

testée avec succès pour des problèmes de classification dans (Lemaire et Clérot, 2004) et de régression dans (Lemaire et Féraud, 2006).

Dans la suite de cet article, nous allons détailler comment cette méthode de mesure d'importance peut être utilisé *instance par instance*. La méthode d'interprétation détaillée par la suite est donc naturellement liée à l'utilisation de cette méthode itérative de sélection de variables. Le processus d'interprétation est réalisé après une phase de sélection de variables qui permet de supprimer les variables fortement corrélées pour ne garder que les variables pertinentes. Cette phase préliminaire de sélection permet de ne pas "diluer" l'importance, et donc l'interprétation, d'une variable.

2.2 Interprétation

La mesure de l'importance d'une variable permet de sélectionner un sous-ensemble de variables pertinentes pour un problème donné. Cette sélection de variables permet d'augmenter la robustesse des modèles et de faciliter l'interprétation d'un modèle. Cependant, la notion d'importance d'une variable pour une instance n'est pas suffisante pour interpréter la classification. Pour la compléter, il est nécessaire d'analyser l'importance de la valeur de cette variable sur la sortie du modèle. Qu'indique la valeur de la variable V_j pour une instance X_n ? Est-il possible de faire changer de classe l'instance X_n en modifiant la valeur de V_j ? Voilà autant de questions que l'on propose de résoudre à l'aide d'une mesure de l'importance de la valeur d'une variable pour une instance donnée. Dans tout ce qui suit nous appellerons l'importance de la valeur d'une variable pour une instance donnée son influence.

Notre objectif est de proposer une méthode qui produit automatiquement (sans aide humaine, sans analyste) une interprétation du score pour chaque instance. C'est pourquoi nous proposons de construire une mesure d'importance pour chaque instance. Parmi les méthodes de l'état de l'art les plus proches de la méthode proposée dans cet article on notera la méthode proposée Kary Framling (Främling, 1996) qui fournit une mesure de l'influence de la variable V_j sur la sortie du modèle. Les inconvénients de cette mesure sont les mêmes que précédemment. La mesure est basée sur les extremums de la variable, et elle est donc très sensible au bruit et aux 'outliers'. Nous proposons une mesure plus robuste qui se base sur la distribution des données.

3 Description de la méthode

On définit ici les notions "d'importance (I) d'une variable pour une instance" et "d'influence (I_v) d'une variable pour une instance" pour l'une des variables V_j en entrée du modèle sur l'une des variables de sortie p du modèle. Ces définitions sont rigoureusement les mêmes pour toutes les variables en entrée et en sortie du modèle.

3.1 Importance à l'instance d'une variable d'entrée

Etant donné le modèle F, l'instance considérée X_n , la variable explicative V_j du modèle et la variable à expliquer p du modèle, on définit la sensibilité du modèle $S(V_j/F, X_n, p)$ par : la moyenne des variations mesurées en sortie du modèle lorsqu'on perturbe l'instance considérée X_n en fonction de la distribution de probabilité de la variable V_j.

La sortie perturbée du modèle F, pour une instance X_n est la sortie obtenue en échangeant la valeur de la variable V_j pour l'instance n par sa valeur pour l'instance k. La variation mesurée, pour l'instance X_n est donc la différence entre la "vraie sortie" du modèle $F_j(X_n)$ pour l'instance n et la "sortie perturbée" du modèle $F_j(X_n, X_k)$. La sensibilité du modèle pour l'exemple X_n à la variable V_j est alors la moyenne des $||F_j(X_n) - F_j(X_n, X_k)||^2$ sur la distribution de probabilité de la variable V_j. En approximant la distribution de probabilité des données à l'aide de la distribution empirique de K exemples on a :

$$S(V_j|F, X_n, p) = \frac{1}{K}\sum_{k=1}^{K} ||F_j\left(X_n\right) - F_j\left(X_n; X_k\right)||^2 \quad (1)$$

En réalisant cette mesure de sensibilité sur la sortie p et quelque soit la variable d'entrée[1] j on possède une distribution des sensibilités. On définit alors l'importance de la variable V_j à l'instance X_n, $I(V_j|F, X_n, p)$, comme étant le rang, o, de la sensibilité du modèle $S(V_j|F, X_n, p)$ parmi l'ensemble des sensibilités $S(V_j|F, X_i, p)\ \forall i, j$. tel que :

$$I(V_j|F, X_n, p) = P[(S(V_j|F, X_i, p)\forall i, \forall j) \leq (S(V_j|F, X_n, p)] \geq o. \quad (2)$$

Cette mesure[2] fournit l'importance d'une variable d'entrée pour une instance relativement à toutes les autres instances et toutes les autres variables. Cette mesure relative permet de se concentrer sur les seules informations pertinentes pour chaque instance.

3.2 Influence à l'instance d'une variable d'entrée

Une variable peut "tirer vers le haut" (valeur forte) ou "tirer vers le bas" (valeur faible) la sortie du modèle. Pour l'exemple X_n la valeur "naturelle" de la sortie p ' du modèle est par définition $F(X_n)$ (que l'on peut encore notée $F_j(X_n, X_n)$). La valeur "perturbée" de la sortie du modèle pour l'exemple et en perturbant la variable d'entrée V_j est $F_j(X_n, X_k)$. C'est simplement la valeur de la sortie du modèle p étant donné l'exemple X_n mais pour lequel on a remplacé la valeur de sa j^{ieme} composante par la valeur d'un autre exemple k.

La distribution des $F_j(X_n, X_k)$ représente ce qu'aurait pu être la valeur de la sortie du modèle pour l'instance X_n si sa variable V_j avait été différente. La position de sa sortie "naturelle" au sein de cette distribution renseigne sur la nature de la valeur de sa variable V_j. On définit alors l'influence de la variable V_j à l'instance X_n, $I_v(V_j|F, X_n, p)$, comme étant le rang, r, de la sortie "naturelle" parmi l'ensemble de ses sorties potentielles tel que :

$$I_v(V_j|F, X_n, p) = P[(F_j(X_n, X_k)\forall k) \leq F(X_n)] \geq r. \quad (3)$$

Cette mesure fournit l'influence d'une variable d'entrée pour une instance relativement à toutes les autres valeurs "potentielles" de la variable. Par exemple, pour un problème de classification à deux classes (-1 ;$+1$), un rang important de I_v dénote d'une influence positive par rapport à la classe $+1$ et négative par rapport à la classe -1. Réciproquement un très faible rang de I_v dénote d'une influence positive par rapport à la classe -1 et négative par rapport à la classe $+1$.

[1] L'importance n'est pas intrinsèque à une variable mais relativement à l'ensemble des variables. La distribution est donc établie quelle que soit la variable d'entrée et sur l'ensemble des instances que l'on possède.

[2] dans l'équation 1 une autre distance aurait pu être choisie. On utilise ici la distance L2 (1) du fait que le calcul utilisant cette distance permet d'obtenir explicitement les coefficients d'une régression linéaire (dans le cas d'un modèle linéaire gaussien, voir Diagne et Lemaire (2006))

4 Illustration

Cette méthode[3] n'ayant pas encore été testée sur une base de données dont les résultats sont publiables on illustre ici son application sur un exemple jouet ; on peut néanmoins trouver son évaluation théorique sur des modèles linéaires gaussiens dans Diagne et Lemaire (2006). Cet exemple de jouet de test, présenté figure 1 (classe '-1' est en noir, classe '+1' et en gris), a été construit dans le but de mesurer et d'illustrer pour chaque instance l'importance et l'influence de chacune des variables prises indépendamment.

4.1 L'exemple jouet

La figure 2 illustre les zones d'influence 'a priori' des deux dimensions : en haut à gauche et à droite les points où les deux dimensions V_1 et V_2 influencent la classe, en haut au centre les points où seul V_1 influence la classe, en bas à gauche et à droite les points où seul V_2 influence la classe, en bas au centre les points où aucune dimension n'influence la classe ('-1' quoi qu'il advienne pour V_1 et V_2).

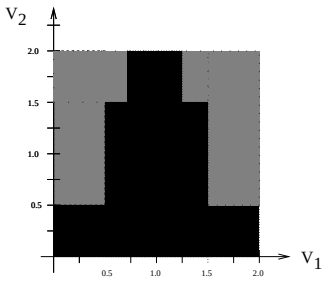


FIG. 1 – *Les deux classes*

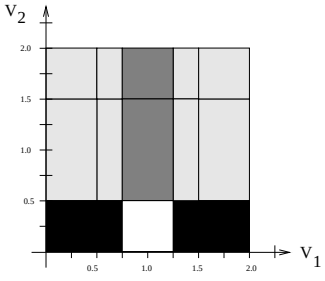


FIG. 2 – *Les influences*

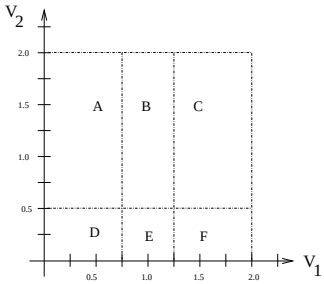


FIG. 3 – *6 des points de test*

1000 exemples ont été tirés aléatoirement pour réaliser un ensemble d'apprentissage et 1000 autre exemples pour réaliser un ensemble de test. Deux types de modèles ont été testé sur cet exemple jouet : 1) un réseau de neurones (Dreyfus, 2002) à une couche cachée utilisant des fonctions d'activation sigmoïdales (MLP). Le nombre de neurones Q de la couche cachée à été fixé en utilisation une validation croisée (Q=4). L'algorithme d'apprentissage est la rétro-propagation de l'erreur quadratique en version stochastique. 2) une fenêtre de Parzen (Parzen, 1962) utilisant un noyau gaussien de norme L2 dont le paramètre σ à été fixé en utilisation une validation croisée (σ=0.1). Dans les deux cas les données ont été centrées et réduites avant "l'apprentissage" (en utilisant les statistiques de l'ensemble d'apprentissage).

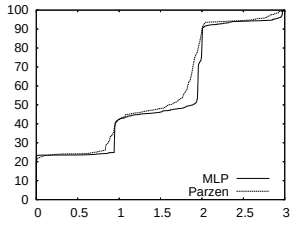
4.2 Construction des éléments de l'interprétation

Parmi les 1000 exemples de test 6 exemples représentatifs des zones d'influence des variables V_1 et V_2 (zones de la figure 2) ont été choisis pour illustrer la méthode. Leur emplacement est indiqué dans la figure 3 et ils sont nommés de A à F. L'interprétation de ces 6 instances nécessite de calculer ($n \in \{A, B, C, D, E, F\}$) : (A) pour l'établissement de $I(V_j/F, X_n, p)$: (A.1) calcul des $S(V_j/F, X_i, p)$ $\forall j, \forall i$ (A.2) établissement de la statistique d'ordre des $S(.)$

[3]Cette méthode a fait l'objet d'un dépôt de brevet numéro 06 53609 en date du 7 septembre 2006.

calculées en A.1 et détermination du rang de $S(V_j/F, X_n, p)$; (B) pour l'établissement de $I_v(V_j/F, X_n, p)$: (B.1) calcul des $F(X_n, X_k)\ \forall k$; (B.2) établissement de la statistique d'ordre $F(X_n, X_k)$ calculées en B.1 et détermination du rang de $F(X_n)$;

La figure 4 présente les distributions des S (équation 1) obtenue pour le MLP et la fenêtre de Parzen sur l'ensemble des exemples ayant servis à l'élaboration du modèle. La valeur en abscisse représente une valeur de sensibilité et sa correspondance en ordonnée délivre le rang de cette valeur. On note sur cette figure que l'ordre de grandeur des rangs des sensibilités est le même pour le MLP et la fenêtre de Parzen. Les rangs des sensibilités, aussi bien sur V_1 que V_2, progressent par palier. Les distributions des sensibilités sont naturellement et logiquement constituées de quelques modalités importantes étant le problème de classification étudié et les modèles employés. Ces distributions sont la concaténation des sensibilités individuelles, la concaténation des zones d'influences : des zones où la variable n'est d'aucun intérêt, des zones où elle revêt le plus grand intérêt et des zones transitoires.



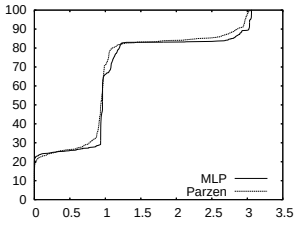


FIG. 4 – *Distribution ordonnée des sensibilités, à gauche pour V_1 et à droite pour V_2.*

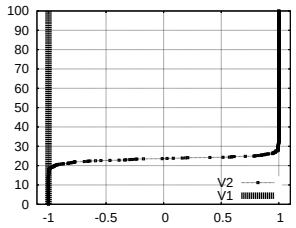


FIG. 5 – *Sorties potentielles de l'instance 'F'*

La figure 5 exhibe les distributions des sorties potentielles pour l'instance F et la variable V_1 ou V_2 pour la fenêtre de Parzen. Sur cette figure on observe que la distribution des sorties potentielles de l'instance 'F' est unimodale lorsqu'on utilise pour le perturber la distribution de probabilité empirique de la variable V_1. On a $F(X_n, X_k) = -1.0\ \forall k$. Ce résultat est cohérent avec le fait que cette variable n'est d'aucun intérêt pour cette instance. Lorsqu'on utilise pour perturber F la distribution de probabilité empirique de la variable V_2 on observe une distribution des $F(X_n, X_k)$ à trois modes : $F(X_n, X_k) = -1$, $-1 \leq F(X_n, X_k) \leq +1$, $F(X_n, X_k) = +1$.

L'analyse des figures 4 et 5 montre qu'il est utile d'attribuer des notions plus robustes non pas à un rang dans la distribution étudiée mais à une plage de rang. On peut, par exemple, attribuer à chaque rang un quintile (Q) dans l'ordre croissant Q_1, Q_2, Q_3, Q_4 et Q_5 et respectivement les notions "Très faible", "Faible", "Moyenne", "Forte" et "Très Forte". Chaque rang appartient à l'un des quintiles (les intervalles correspondant à Q_1 et Q_5 sont respectivement ouvert à gauche et à droite). Ils se voient de par cette appartenance attribués l'une des 5 notions énoncées ci-dessus (valeurs de Q dans le tableau 1).

Si on observe les valeurs obtenues dans le tableau 1 pour chacun des 6 points caractéristiques à la lumière des figures 4 et 5, on observe alors une totale cohérence dans les résultats obtenus : les faibles sensibilités ont des petits rangs, les très petites sensibilités qualifient les variables de très peu importantes (Q_1), etc. Les importances des valeurs des variables doivent être aussi lues à la lumières des importances (si $I = 0$ il est inutile de lire l'I_v correspondant). Les variables 'très peu' importantes pour une instance ne devraient pas être utilisées dans l'in-

terprétation. Les valeurs de I et I_v sont alors inutilisées (absence de la ligne correspondante dans le tableau 1).

	F=MLP				F=Parzen			
V_j, X_n	S	I	$F(X_n)$	I_v	S	I	$F(X_n)$	I_v
V_1, X_A	1.24	Q_4 (o=63)	+1.00	Q_5 (r=99)	1.16	Q_4 (o=63)	+0.99	Q_4 (r=74)
V_2, X_A	0.96	Q_3 (o=49)	+1.00	Q_5 (r=99)	0.97	Q_3 (o=53)	+0.99	Q_4 (r=74)
V_1, X_B	2.70	Q_5 (o=89)	-1.00	Q_1 (r=14)	2.28	Q_5 (o=89)	-0.99	Q_2 (r=25)
V_1, X_C	1.24	Q_4 (o=63)	+1.00	Q_5 (r=99)	1.16	Q_4 (o=63)	+0.99	Q_4 (r=75)
V_2, X_C	0.93	Q_2 (o=31)	+1.00	Q_5 (r=99)	0.90	Q_2 (o=35)	+0.99	Q_4 (r=67)
V_2, X_D	3.03	Q_5 (o=95)	-1.00	Q_2 (r=22)	2.96	Q_5 (o=96)	-0.99	Q_1 (r=12)
V_2, X_F	3.05	Q_5 (o=98)	-1.00	Q_2 (r=21)	3.02	Q_5 (o=90)	-0.99	Q_1 (r=12)

TAB. 1 – *Interprétation des points caractéristiques pour le MLP et la fenêtre de Parzen*

4.3 Interprétation obtenue

On présence deux exemples de l'interprétation obtenue à l'aide du tableau 1. Pour l'instance A et en utilisant la fenêtre de Parzen on obtient que le point A est de la classe '+1' avec une probabilité de 0.99 (la valeur de $F(X_A)$) car : (1) V_1 qui est fortement importante indique qu'il est fortement de la classe '+1' et (2) V_2 qui est moyennement importante indique qu'il est fortement de la classe '+1'. Pour l'instance D[4] et en utilisant le MLP on obtient que le point D est de la classe '-1' avec une probabilité de 1.00 (la valeur de $F(X_D)$) car V_2 qui est très fortement importante indique qu'il est fortement de la classe '-1'.

L'examen des interprétations obtenues dans le tableau 1 sur l'ensemble des points de la figure 3 montre que l'on obtient une interprétation cohérente, voir identique, quel que soit le modèle utilisé. Cet aspect, désiré, de la méthode est obtenu sur les deux modèles testés. L'interprétation produit aussi une autre connaissance : l'importance et l'influence des variables étant connues il est possible d'essayer de faire changer de classe une instance, un individu. Par exemple pour l'instance A, il suffit de regarder la distribution de ses sorties potentielles et de remarquer que si on lui fait adopter une valeur de V_1 différente il sera possible de le faire changer de classe.

5 Conclusion et discussion

On a présenté dans cet article une méthode permettant d'interpréter les résultats délivrés par un modèle prédictif de classification ou de régression sur deux types de modèle un réseau de neurones artificiel et une fenêtre de Parzen. La méthode a été posée et illustrée sur un exemple jouet se prêtant bien à son illustration pas à pas. Pour la partie interprétation nous montrerons dans de futurs travaux qu'elle peut être appliquée à d'autres modèles dont le réseau bayésien naïf et ce sur des bases de données benchmark libres d'accès. Nous espérons aussi montrer que la notion de rang de l'importance à l'exemple peut être utilisée dans un processus d'apprentissage actif et/ou dans un nouveau concept 'de fragilité à l'exemple' d'un modèle. L'interprétation automatique proposée dans cet article, intelligible par le plus grand nombre, peut être industrialisée et fournie avec le score afin d'aider à la personnalisation des services.

[4]Pour l'instance D de classe '-1' et réciproquement au point 'A' de classe '+1' c'est un très faible rang de I_v qui dénote d'une influence positive par rapport à la classe -1 et négative par rapport à la classe $+1$. Voir section 3.2

Références

Arcadius, Y., J. Akossou, et R. Palm (2005). Conséquences de la sélection de variables sur l'interprétation des résultats en régression linéaire multiple. In *Biotechnol. Agron. Soc. Environ.*, Volume 9, pp. 11–18.

Baxt, W. G. et H. White (1995). Bootstrapping confidence intervals for clinical inputs variable effects in a network trained to identify the presence of acute myocardial infraction. *Neural Computation 7*, 624–638.

Breiman, L. (2001). Random forest. *Machine Learning 45*(1), 5–32.

Diagne, G. et V. Lemaire (2006). Sélection de variables et méthode d'interprétation des résultats obtenus par un modèle "boite noire". Master's thesis, Université de Versailles Saint-Quentin en Yvelines.

Dreyfus, G. (2002). *Réseaux de neurones - Méthodologies et Applications*. Eyrolles.

Främling, K. (1996). *Modélisation et apprentissage des préférences par réseaux de neurones pour l'aide à la décision multicritère*. Ph. D. thesis, INSA de Lyon.

Féraud, R. et F. Clérot (2002). A methodology to explain neural network classification. *Neural Networks 15*(2), 237–246.

Iemma, A. F. et R. Palm (2003). *Notes de statistique : Interprétation géométrique de la regression*. Université des Sciences agronomiques agronomiques. `www.fsagx.ac.be/si/`.

Lemaire, V. et F. Clérot (2004). An input variable importance definition based on empirical data probability and its use in variable selection. In *International Joint Conference on Neural Networks IJCNN*, Volume 2, pp. 1375–1380.

Lemaire, V. et R. Féraud (2006). Driven forward features selection : a comparative study on neural networks. In *International Conference on Neural Information Processing, ICONIP (2)*, Hong-Kong, pp. 693–702.

Moody, J. (1994). *Prediction Risk and Architecture Selection for Neural Networks*. From Statistics to Neural Networks-Theory and Pattern Recognition. Springer-Verlag.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics 33*, 1065–1076.

Réfénes, A. N., A. Zapranis, et J. Utans (1994). Stock performance using neural networks : A comparative study with regression models. *Neural Network 7*, 375–388.

Thrun, S. (1995). Extracting rules from artifcial neural networks with distributed representations. In M. Press (Ed.), *Advances in Neural Information Processing Systems*, Volume 7, Cambridge, MA. G. Tesauro, D. Touretzky, T. Leen.

Summary

This paper presents a method allowing to interpret the output of a predictive model of classification or regression. The intepretation is based on two measures : the importance of the value of the variable and the importance of the variable. Contrary to most of state of the art methods of interpretation, our approach allows to intrepret the output model for every instances. The understanding of the score delivered by the model, for a instance, allows a help to a immediate, fast, effective and contextual decision.