Estimation des performances de classifieurs d'ensembles à partir des propriétés des classifieurs individuels

André SMOLARZ[1], Yuan DONG[1], Pierre BEAUSEROY[1]

[1]UMR 6279 "Sciences et Technologies pour la Maîtrise des Risques"
Équipe Modélisation et Sûreté des Systèmes – Université de Technologie de Troyes,
12, Rue Marie Curie, BP 2060, 10010 Troyes Cedex
andre.smolarz@utt.fr, yuan.dong@utt.fr, pierre.beauseroy@utt.fr

Résumé – En considérant le diagnostic d'un système comme un problème de décision à deux classes, « fonctionnement normal » et « fonctionnement anormal », nous proposons d'étudier la performance globale d'une règle de décision en modélisant et en fusionnant les sorties d'un ensemble de classifieurs en fonction des performances individuelles de chacun d'entre eux et de leurs corrélations. Dans ce contexte nous considérons des mesures issues d'un ensemble de capteurs et chaque classifieur est alors défini sur un sous-espace de mesures issues d'un sous ensemble de capteurs choisis au hasard parmi tous les capteurs.

Abstract – In this paper, system diagnosis is considered as a two classes (normal and abnormal) problem. We study a decision rule based on ensemble method. The global decision taken by the rule is obtained by modelizing and combining numbers of individual decisions taken by a set of classifiers. The measurements are obtained from a set of sensors and each classifier is trained using a randomly selected subset of these sensors. This work is a first attempt to study the relation between the global performance of the decision rule and the individual classifiers performances, taking into account the correlation between classifiers.

1 Introduction

On considère un problème à deux classes pour lequel on met en œuvre une règle de décision résultant de la fusion d'un sous-ensemble de classifieurs choisis au hasard [4] au sein d'un ensemble complet de classifieurs. Les données à classer sont représentées dans un espace de représentation par un vecteur d'attributs que nous considèrerons issus de capteurs. Chaque classifieur prend une décision dans un sous-espace issu d'une projection de l'espace d'attributs initial. Comme dans beaucoup de méthodes d'ensemble ([3], [2]), la décision finale est prise par combinaison des décisions des classifieurs et nous avons opté ici pour la règle reposant sur le vote majoritaire. Pour optimiser la règle de décision globale, de nombreux paramètres interviennent : le nombre de classifieurs, les réglages de chaque classifieur en termes de compromis entre les erreurs de première et de seconde espèce, la nature de la combinaison et cette phase de « réglage » est en général assez lourde. De ce fait, si l'espace de représentation initial vient à changer à un moment donné (par exemple, à la suite de la panne d'un ou plusieurs capteurs), il faut à nouveau procéder à une phase de « réglage ». Pour contourner cet aspect, nous proposons d'évaluer au départ la loi conjointe d'un vecteur aléatoire décisionnel dont les composantes sont les sorties de l'ensemble complet de tous les classifieurs considérés. Les sous-espaces de représentation des différents classifieurs n'étant généralement pas disjoints, les composantes du vecteur décisionnel constituent une suite de variables de Bernoulli corrélées. La loi de probabilité conjointe de telles variables est utile dans de nombreuses situations ([6], [5]). Bahadur [1] a proposé un modèle permettant d'obtenir la loi conjointe d'un tel vecteur, mais sa mise en œuvre n'est plus possible lorsque la dimension du vecteur est trop importante, à cause du nombre de corrélations à estimer. Une alternative consiste à évaluer la loi conjointe du vecteur décisionnel au moyen d'une méthode reposant sur l'entropie qui a été proposée par Van der Geest [5].

Nous allons tout d'abord présenter la formalisation du problème en précisant les notations ainsi que les caractéristiques et la définition des règles de décision. Dans la partie suivante, nous présenterons quelques simulations et résultats. Enfin nous conclurons et détaillerons quelques éléments de perspectives.

2 Approche et notations

Si on considère K capteurs C_i recevant une donnée $\mathcal{D}$ et délivrant chacun un attribut représenté par une variable aléatoire X_i, l'espace de représentation initial est alors de dimension K et on va considérer l'ensemble des $N = \mathsf{C}_K^d$ sous-espaces d'attributs de dimension d. Le problème étant à deux classes ω_0 et ω_1 (ou deux hypothèses H_0 et H_1), en définissant un classifieur h_ℓ dans chaque sous-espace, on obtient un vecteur décisionnel constitué de N variables de Bernoulli. Ce principe général est résumé par le schéma du tableau 1.

Si on adopte la règle $Y_\ell = 1$ pour la sortie de chaque classifieur s'il décide en faveur de ω_1 et $Y_\ell = 0$ pour ω_0, on a alors

les probabilités suivantes de fausse alarme et de non détection pour chaque classifieur :

$$\begin{aligned} P[Y_\ell = y_\ell \mid \omega_0] &= \alpha_\ell^{y_\ell}(1-\alpha_\ell)^{1-y_\ell} \quad y_\ell \in \{0,1\} \\ P[Y_\ell = y_\ell \mid \omega_1] &= \beta_\ell^{1-y_\ell}(1-\beta_\ell)^{y_\ell} \quad \ell = 1,\ldots,N \end{aligned} \tag{1}$$

TABLE 1 – Processus décisionnel avec K capteurs et N sous-ensembles de dimension $d = 2$

donnée	sorties capteurs	sous-espaces d'attributs	classifieurs	sorties classifieurs (variables décisionnelles)
	$C_1(\mathcal{D}) = X_1$	$SE_1 = (X_1, X_2)$	h_1	$Y_1 = h_1(X_1, X_2)$
$\mathcal{D}$ (signal, image...)	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$C_K(\mathcal{D}) = X_K$	$SE_N = (X_{K-1}, X_K)$	h_N	$Y_N = h_N(X_{K-1}, X_K)$
	vecteur d'attributs $\boldsymbol{X} = [X_1, \ldots, X_K]^t$			vecteur décisionnel $\boldsymbol{Y} = [Y_1, \ldots, Y_N]^t$

La règle de décision finale du vote majoritaire bâtie sur l'ensemble complet des classifieurs, s'écrit alors

$$D_1 \quad (\text{décision } \omega_1) \quad \text{si} \quad \sum_{\ell=1}^{N} Y_\ell > \frac{N}{2} \tag{2}$$

La probabilité exprimée en (2) se calcule aisément si l'on dispose de la loi conjointe $p_{\boldsymbol{Y}}(\boldsymbol{y}_k) = P[\boldsymbol{Y} = \boldsymbol{y}_k]$. Comme nous l'avons déjà évoqué en introduction, nous disposons de deux approches pour évaluer cette distribution conjointe.

Modèle de Bahadur

Sous l'hypothèse que les classifieurs sont indépendants, les probabilités conjointes pour une réalisation $\boldsymbol{y} = [\, y_1, y_2, \ldots y_N \,]^t$ donnée s'écriront :

$$\begin{cases} P_{indep}[\boldsymbol{Y} = \boldsymbol{y} \mid H_0] &= \prod_{\ell=1}^{N} \alpha_\ell^{y_\ell}(1-\alpha_\ell)^{1-y_\ell} \\ P_{indep}[\boldsymbol{Y} = \boldsymbol{y} \mid H_1] &= \prod_{\ell=1}^{N} \beta_\ell^{1-y_\ell}(1-\beta_\ell)^{y_\ell} \end{cases} \tag{3}$$

Le modèle proposé par Bahadur [1] s'écrit alors

$$P[\boldsymbol{Y} = \boldsymbol{y}\,] = f(\boldsymbol{y}) \times P_{indep}[\boldsymbol{Y} = \boldsymbol{y}\,] \tag{4}$$

$$\text{avec} \;\; f(\,\boldsymbol{y}\,) = 1 + \sum_{i<j} r_{ij} z_i z_j + \ldots + r_{12\ldots N} z_1 \ldots z_N \tag{5}$$

Dans l'équation (5) les z_i représentent les réalisations des variables Z_i qui sont les versions centrées réduites des variables Y_i et les coefficients de corrélations r s'expriment comme suit

$$r_{i_1 i_2 \ldots i_n} = E[Z_{i_1} Z_{i_2} \ldots Z_{i_n}] \tag{6}$$

Il y a au total $2^N - N - 1$ coefficients à fixer pour évaluer le modèle complet, ce qui pose un problème de taille lorsque le nombre de classifieurs est trop important. Par ailleurs, le choix d'un modèle tronqué ne prenant en compte qu'une partie des coefficients intervenant dans la fonction correctrice f, engendre des solutions qui ne vérifient pas les axiomes des probabilités.

Maximisation de l'entropie

Le principe proposé par Van der Geest [5] consiste à maximiser l'entropie de la distribution conjointe sous contraintes prenant en compte les probabilités marginales et les moments d'ordre 2 (Cf. système d'équations (7)).

$$\text{Maximiser} \quad -\sum_{i=1}^{2^N} p_{\boldsymbol{Y}}(\boldsymbol{y}_i)\log(p_{\boldsymbol{Y}}(\boldsymbol{y}_i))$$

$$\text{s.c.} \begin{cases} \sum_{k=1}^{2^N} p_{\boldsymbol{Y}}(\boldsymbol{y}_k) = 1 \\ \sum_{k=1}^{2^N} A_{rk} p_{\boldsymbol{Y}}(\boldsymbol{y}_k) = b_r \;\; (r = 1, \ldots, N(N+1)/2) \end{cases} \tag{7}$$

Les coefficients A_{rk} forment les lignes d'une matrice $\boldsymbol{A}$ $(N(N+1)/2) \times 2^N$. les N premières valeurs de b_r contiennent les espérances marginales $E[Y_i] = p_i = P[Y_i = 1]$ et les dernières contiennent les $N(N-1)/2$ moments d'ordre 2, $E[Y_i Y_j]$. En notant que $E[Y_i Y_j] = P[Y_i = 1, Y_j = 1]$, chaque ligne de la matrice est constituée de "0" et de "1", afin de sélectionner les termes de la distribution conjointe dont la somme donne les valeurs b_r

Exploitation du modèle

Une fois déterminée la loi conjointe $p_{\boldsymbol{Y}}(\boldsymbol{y}_k) = P[\boldsymbol{Y} = \boldsymbol{y}_k]$, on peut en déduire les lois conjointes de tout sous-ensemble de variables extraites de $\boldsymbol{Y}$. En considérant par exemple un vecteur $\boldsymbol{Y}_i^{(L)}$ constitué d'un ensemble de $L < N$ composantes extraites de $\boldsymbol{Y}$, la règle de décision du vote majoritaire bâtie sur l'ensemble des classifieurs correspondant à $\boldsymbol{Y}_i^{(L)}$ s'écrit alors

$$D_1 \quad (\text{décision } \omega_1) \quad \text{si} \quad \sum_{\ell=1}^{L} Y_{i,\ell} > L/2 \tag{8}$$

et les caractéristiques de cette règle se calculent aisément comme suit selon les expressions (9)

$$\begin{cases} \alpha_L = P[D_1 \mid \omega_0] &= P\left[\sum_{\ell=1}^{L} Y_\ell > \frac{L}{2} \mid \omega_0\right] \\ \beta_L = P[D_0 \mid \omega_1] &= P\left[\sum_{\ell=1}^{L} Y_\ell \leq \frac{L}{2} \mid \omega_1\right] \end{cases} \tag{9}$$

3 Expérimentations et résultats

Protocole expérimental

Nous considérons les sorties X des capteurs comme des variables gaussiennes de même variance $\sigma^2 = 1$ dans les deux classes. Les moyennes $m_0 = E[X \mid \omega_0]$ et $m_1 = E[X \mid \omega_1]$ ont été fixées en fonction de $\alpha_\ell = \alpha$ et $\beta_\ell = \beta \;\forall\; \ell$ et de telle sorte que la règle adoptée pour chaque classifieur défini sur d capteurs soit de la forme D_1 si $\sum_{i=1}^{d} X_i < 0$. Dans ce contexte, nous avons effectué une simulation en considérant $K = 5$ capteurs et $N = 10$ sous-ensembles de dimension $d = 2$ et nous avons fixé $\alpha = 0.15$ et $\beta = 0.05$ pour l'ensemble des N classifieurs. Nous avons alors déterminé la loi

conjointe du vecteur décisionnel $\boldsymbol{Y}$ par le modèle de Bahadur et par maximisation de l'entropie. Les coefficients r (Cf. équation (5)) ainsi que les moments intervenant dans les contraintes du système (7) ont été estimés sur deux ensembles d'apprentissage de tailles $N_{app} = 1000$ et 100000 exemples pour chaque classe.

Nous avons ensuite évalué les performances α_N et β_N de la règle du vote majoritaire à l'aide de la loi conjointe du vecteur décisionnel complet $\boldsymbol{Y}$ (de dimension N) au moyen de la relation (9), ainsi que celles α_L et β_L des règles reposant sur 5 vecteurs décisionnels $\boldsymbol{Y}_i^{(L)}$ constitués de $L = 3, 5$ et 7 variables décisionnelles tirées au hasard parmi les $N = 10$ composantes de $\boldsymbol{Y}$. Afin de valider le calcul des performances de la règle globale à l'aide des modèles conjoints des vecteurs décisionnels, nous présentons en parallèle les performances moyennes estimées ($\hat{\alpha}_L$, $\hat{\beta}_L$, $\hat{\alpha}_N$ et $\hat{\beta}_N$ sur plusieurs ensembles de test de tailles $N_{test} = 500, 1000$ et 100000 exemples pour chaque classe. Enfin, pour comparaison, les performances (α_{CU} et β_{CU}) d'un classifieur unique reposant sur les K capteurs avec la règle D_1 si $\sum_{i=1}^{K} X_i < 0$ sont également présentés.

Analyse des résultats

Tout d'abord nous avons estimé la distribution du vecteur décisonnel sous les deux hypothèses au moyen de deux échantillons de 1000000 exemples afin d'avoir une référence pour évaluer la pertinence des distributions obtenues par le modèle de Bahadur et par le maximum d'entropie. Le tableau 2 présente les distances quadratiques moyennes entre chaque modèle et la loi estimée qui sert de référence. On peut observer que le modèle de Bahadur comme la méthode du maximum d'entropie proposent une bonne approximation du modèle conjoint de $\boldsymbol{Y}$.

TABLE 2 – Distance quadratiques moyennes entre les modèles proposés et la distribution estimée

	Bahadur $N_{app} = 10^3$	Bahadur $N_{app} = 10^5$	Bahadur $N_{app} = 10^6$	Entropie $N_{app} = 10^3$	Entropie $N_{app} = 10^5$
H_0	$7,02\ 10^{-4}$	$6,92\ 10^{-6}$	$1,01\ 10^{-6}$	$9,09\ 10^{-4}$	$2,19\ 10^{-4}$
H_1	$1,72\ 10^{-4}$	$4,91\ 10^{-6}$	$8,65\ 10^{-7}$	$6,50\ 10^{-3}$	$7,50\ 10^{-5}$

Sur le graphique de la figure 1, les points représentent les estimations de α_N et β_N obtenues sur 5 échantillons test de tailles allant de 500 à 100000. Les lignes horizontales représentent quant à elles, les valeurs de α_N et β_N calculées avec les distributions obtenues par le modèle de Bahadur et par la méthode de l'entropie pour une taille d'ensemble d'apprentissage de 100000 exemples.

On peut noter que le modèle de Bahadur s'avère plus précis que celui fourni par le maximum d'entropie. Par ailleurs on observe que l'évaluation des performances globales par estimation nécessite une taille d'ensemble test élevée pour parvenir à des résultats comparables à ceux donnés par le modèle. Cette remarque confirme l'utilité de disposer de la loi du vecteur décisionnel dès la phase d'apprentissage. Sur la figure 2 on peut observer que le modèle de Bahadur reposant sur une taille d'apprentissage de 1000 exemples, permet d'évaluer les performances de la règle de décision globale de façon plus précise que par estimation sur un ensemble test de taille comparable.

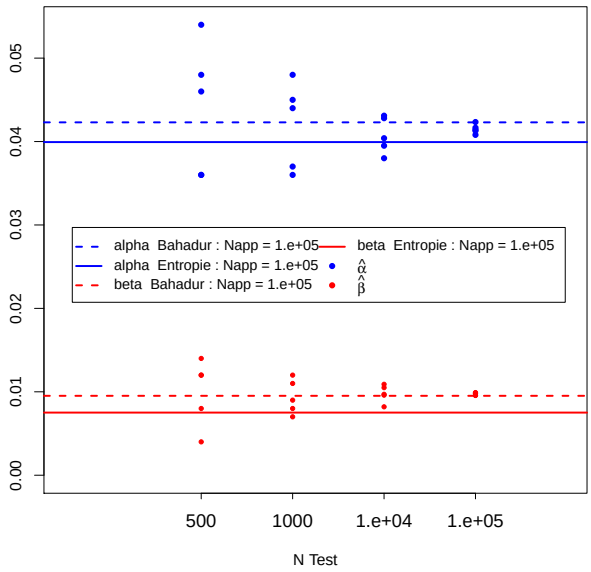


FIGURE 1 – Comparaison des caractéristiques de la règle globale selon leur mode de calcul.

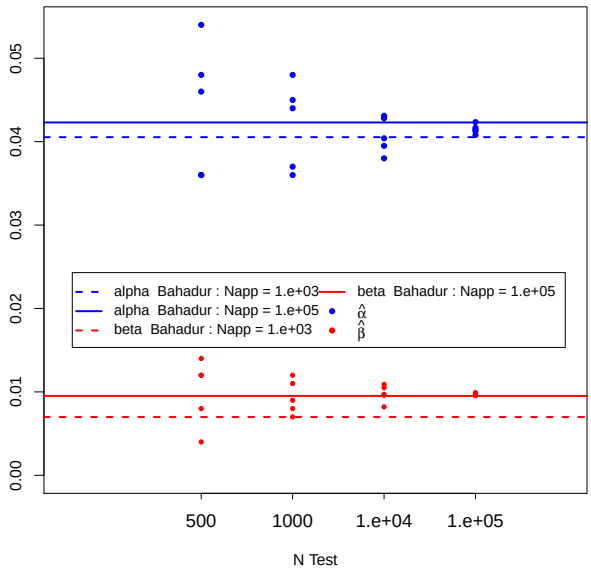
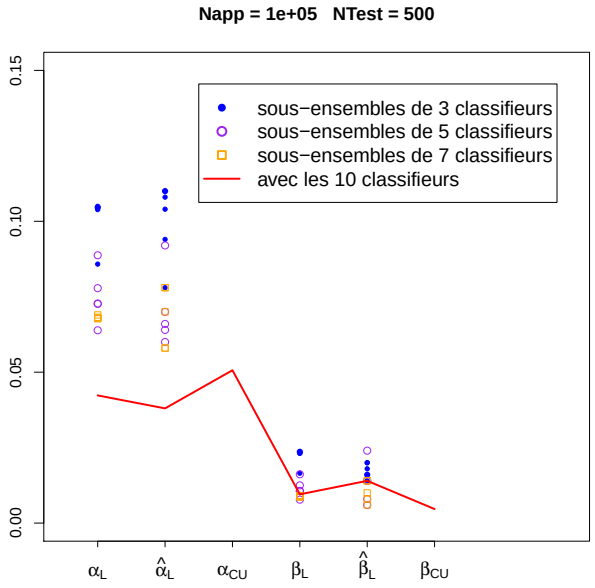


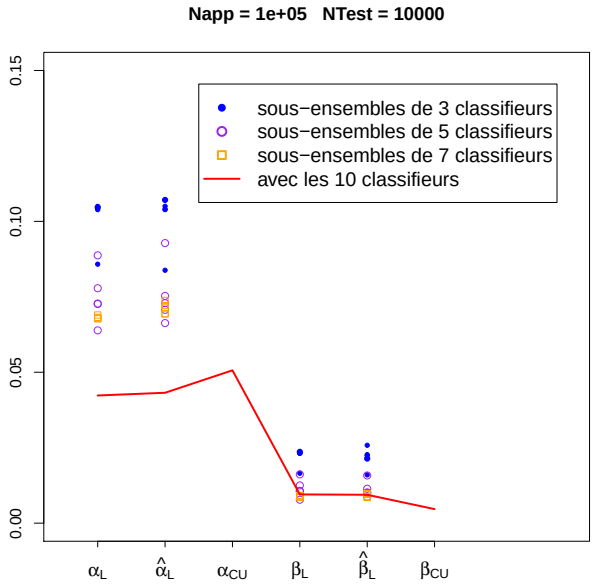
FIGURE 2 – Comparaison des caractéristiques de la règle globale selon l'effectif d'apprentissage.

Enfin, les graphiques des figures 3 (a) et (b) illustrent les ré-

sultats obtenus par estimation et par évaluation avec le modèle de Bahadur sur 5 sous-ensembles de 3, 5 et 7 classifieurs. Là encore on peut observer une bonne concordance entre les caractéristiques estimées au moyen de données test et celles obtenues par calcul à l'aide de la distribution du vecteur décisionnel. On peut également noter l'effet du nombre de classifieurs sur les caractéristiques globales de la règle du vote majoritaire.



(a)



(b)

FIGURE 3 – Caractéristiques de la règle globale selon le nombre de classifieurs et avec le modèle de Bahadur

4 Conclusion

Les résultats présentés dans cet article montrent que la connaissance de la distribution conjointe du vecteur décisionnel associé à l'ensemble de tous les classifieurs possibles, permet d'évaluer à tout moment les performances de toute règle reposant sur un sous-ensemble quelconque de classifieurs extrait de l'ensemble complet prédéfini et ceci, sans avoir à procéder à de nouveaux « réglages ». Bien que l'exemple que nous avons choisi corresponde à une situation relativement simple ces résultats montrent aussi que cette approche permettra peut-être d'étudier les relations entre les performances des classifieurs individuels et celles de la règle globale qui fusionne les décisions. À terme, l'exploitation de ces relations permettrait de maîtriser plus facilement la mise en œuvre de méthodes d'ensembles, notamment en permettant de guider le concepteur dans la phase de réglage des classifieurs individuels pour optimiser la règle de décision finale. Un aspect crucial réside cependant dans le nombre de paramètres à estimer pour le modèle de Bahadur qui peut constituer un point de blocage dnas le cas d'un grand nombre de capteurs et donc de classifieurs. La méthode du maximum d'entropie semble a priori moins problématique de ce point de vue, mais il faut néanmoins réfléchir à une formulation minimale en termes de contraintes si l'on souhaite par exemple intégrer des moments d'ordre supérieur à 2 dans celles-ci. Enfin, dans ce contexte, nous comptons également nous intéresser au cas où les classifieurs individuels ne partagent plus les mêmes caractéristiques et pour lequel la règle du vote majoritaire n'est plus pertinente.

Références

[1] R.R. Bahadur. *A representation of the joint distribution of responses to n dichotomous items*. Studies in item Analysis and Prediction, pages 158-168. Stanford University Press, H. Solomon edition, 1961.

[2] P. Beauseroy, A. Smolarz, and Y. Dong. Classification robuste via la sélection d'un ensemble de sous-espaces de représentation. Bruxelles, 21 - 25 Mai 2012. 44èmes Journées de statistique.

[3] L. Breiman. Random forests. *Machine learning*, Vol. 45(1) :5–32, 2001.

[4] T.K. Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20 :832–844, 1998.

[5] P.A.G. Van der Geest. The binomial distribution with dependent Bernoulli trials. *Journal of Statistical Computation and Simulation*, Vol. 75 :141–154, 2005.

[6] Alexander Zaigraev and Serguei Kaniovski. Exact bounds on the probability of at least successes in exchangeable Bernoulli trials as a function of correlation coefficients. *Statistics & Probability Letters*, Vol. 80(13-14) :1079 – 1084, 2010.