Sélection topologique de variables dans un contexte de discrimination

Fatima-Zahra Aazi*, Rafik Abdesselam**

*Laboratoires ERIC & LAMSAD
Universités Lumière Lyon 2, France & Hassan 1^{er}, Settat, Maroc
5, avenue Pierre Mendès-France, 69676 Bron Cedex, France
Fatima-Zahra.Aazi@univ-lyon2.fr
**COACTIS-ISH, Université de Lyon, Lumière Lyon 2
14/16, avenue Berthelot, 69363 Lyon Cedex 07, France
rafik.abdesselam@univ-lyon2.fr
http://eric.univ-lyon2.fr/~rabdesselam/fr/

Résumé. En apprentissage automatique, la présence d'un grand nombre de variables explicatives conduit à une plus grande complexité des algorithmes et à une forte dégradation des performances des modèles de prédiction. Pour cela, une sélection d'un sous-ensemble optimal discriminant de ces variables s'avère nécessaire. Dans cet article, une approche topologique est proposée pour la sélection de ce sous-ensemble optimal. Elle utilise la notion de graphe de voisinage pour classer les variables par ordre de pertinence, ensuite, une méthode pas à pas de type ascendante "forward" est appliquée pour construire une suite de modèles dont le meilleur sous-ensemble est choisi selon son degré d'équivalence topologique de discrimination. Pour chaque sous-ensemble, le degré d'équivalence est mesuré en comparant la matrice d'adjacence induite par la mesure de proximité choisie à celle induite par la "meilleure" mesure de proximité discriminante dite de référence. Les performances de cette approche sont évaluées à l'aide de données simulées et réelles. Des comparaisons de sélection de variables en discrimination avec une approche métrique montrent une bien meilleure sélection à partir de l'approche topologique proposée.

1 Introduction

Le changement majeur de la nature des données, notamment l'accroissement de leur masse et de leur taille, nécessite de nouvelles approches de traitement statistique adaptées aux caractéristiques de ces données modernes et massives. En particulier, la grande dimension des données (nombre important de variables) pose un certain nombre de problèmes à la statistique multivariée, notamment aux techniques prédictives de classement. On peut citer les problèmes numériques, de collinéarité, d'inférence ou de biais des estimateurs. Il est nécessaire de développer des méthodes capables de pallier ces problèmes.

On s'intéresse ici au problème de sélection de variables en discrimination. En effet, vu que les variables pertinentes ne sont pas connues *a priori*, la sélection est justifiée, en présence

d'un grand nombre d'attributs, par la possibilité d'existence de variables corrélées entre elles et/ou de variables bruit ou redondantes qui donnent généralement des taux d'erreur importants.

La sélection de variables permet essentiellement d'améliorer les performances des modèles de classement en n'utilisant que les variables importantes dites discriminantes pour le problème étudié, de réduire le temps et le coût de calcul et de faciliter la compréhension du processus générateur d'information.

Plusieurs méthodes ont été proposées pour la sélection de variables dans un contexte de discrimination. Ces méthodes sont regroupées en trois catégories. La première qui regroupe les méthodes dites filtres, évalue la pertinence d'une variable ou d'un sous-ensemble de variables *a priori* avant l'estimation du modèle de prédiction en exploitant les caractéristiques intrinsèques des données d'apprentissage, Dudoit et al. (2002); Peng et al. (2005); Koller et Sahami (1996). La deuxième catégorie, constituée des méthodes appelées "wrapper", utilise la fonction de décision comme critère d'évaluation. Les méthodes de cette catégorie estiment d'abord la fonction de décision puis mesurent l'effet des sous-ensembles sur les performances du modèle ou de l'algorithme utilisé, Kohavi et John (1997); Guyon et al. (2002); Rakotomamonjy (2003). Enfin, la dernière catégorie représente les méthodes dites intégrées, comme leur nom l'indique, elles intègrent la sélection dans le processus de classification, Wang et Shen (2007); Zhang et al. (2008).

Dans cet article, on se propose de présenter une nouvelle approche de Sélection Topologique de variables dans un contexte de Discrimination (STVD). Cette méthode fait partie de la première catégorie d'approches, lesquelles dans la majorité des cas, calculent un score de pertinence des variables et éliminent celle(s) avec les plus faibles scores. Ces scores peuvent être calculés en utilisant des critères d'évaluation tels que les mesures d'information, de dépendance, de similitude ou encore de distance. Les notions de graphe de voisinage et d'équivalence topologique, Batagelj et Bren (1992, 1995); Malerba et al. (2001, 2002); Abdesselam et Zighed (2011); Zighed et al. (2012), sont utilisées ici pour calculer les scores de variables et sélectionner le sous-ensemble optimal discriminant.

Trois étapes sont décrites pour la sélection du sous-ensemble optimal de variables. La première consiste à choisir la "meilleure" mesure de proximité discriminante adaptée aux données, Abdesselam (2014); Aazi et Abdesselam (2015). Cette mesure de proximité est utilisée dans les deux étapes suivantes pour construire le graphe de voisinage de la structure topologique choisie. La deuxième étape est consacrée au classement des variables par ordre de pertinence, puis, en utilisant la méthode pas à pas ascendante "forward", on construit une suite de modèles en intégrant à chaque fois une variable dans l'ordre décroissant de pertinence. On calcule ensuite dans la dernière étape, le critère d'évaluation représenté par le degré d'équivalence topologique en discrimination. Le modèle qui donne la valeur optimale de ce critère d'évaluation, sur un ensemble de données de validation, sera choisi comme étant le meilleur modèle contenant le nombre optimal de variables.

L'article est organisé comme suit. Dans la section 2, on rappelle la notion de graphe de voisinage, la règle de construction de la matrice d'adjacence associée à une mesure de proximité et le calcul du degré d'équivalence topologique dans un contexte de discrimination. Une description des étapes de l'approche proposée est donnée dans la section 3. Les résultats des expérimentations menées sur des données simulées et réelles ainsi que les comparaisons effectuées sont présentés dans la section 4. Une conclusion ainsi que quelques perspectives pour la suite de ce travail sont données en section 5.

2 Structure topologique

La procédure de sélection de variables de l'approche proposée utilise le critère d'équivalence topologique basé sur la notion de graphe de voisinage. L'idée de base est assez simple : deux mesures de proximité sont équivalentes si les graphes topologiques correspondants induits sur l'ensemble des objets restent identiques.

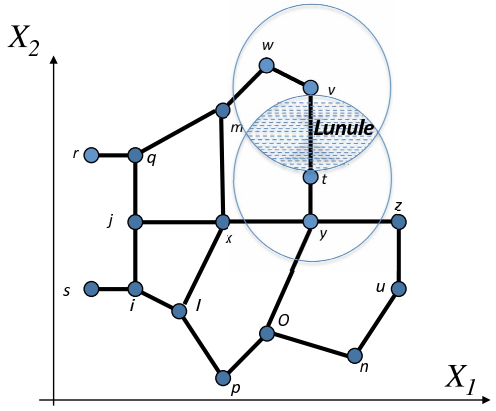
Nous allons tout d'abord rappeler brièvement qu'est-ce qu'un graphe topologique, comment le construire et établir la matrice d'adjacence binaire associée à une mesure de proximité choisie.

2.1 Graphe topologique

Etant donné un ensemble $E = \{x, y, z, \ldots\}$ de $n = |E|$ points-observations de R^p, on peut définir au moyen d'une mesure de proximité u une relation binaire de voisinage V_u sur $E \times E$. Sur cet ensemble d'observations, on peut construire un graphe de voisinage dont les sommets représentent les observations et les arêtes sont définies par une propriété de la relation de voisinage.

De nombreuses définitions sont possibles pour construire cette relation binaire de voisinage. On peut, par exemple, choisir l'Arbre de Longueur Minimale (ALM) Kim et Lee (2003), le Graphe de Gabriel (GG) Park et al. (2006), ou encore le Graphe des Voisins Relatifs (GVR) Toussaint (1980), dont tous les couples de points voisins vérifient la propriété suivante :

$$\begin{cases} V_u(x,y) = 1 & si \;\; u(x,y) \leq \max(u(x,z), u(y,z)) \;;\; \forall z \in E - \{x,y\} \\ V_u(x,y) = 0 & sinon \end{cases} \quad (1)$$



V_u	…	x	y	z	t	u	…
⋮	…	⋮	⋮	⋮	⋮	⋮	…
x	…	1	1	0	0	0	…
y	…	1	1	1	1	0	…
z	…	0	1	1	0	1	…
t	…	0	1	0	1	0	…
u	…	0	0	1	0	1	…
⋮	…	⋮	⋮	⋮	⋮	⋮	…

FIG. 1 – *Graphe topologique GVR - Matrice d'adjacence binaire associée.*

La figure 1 montre, dans R^2 muni de la distance euclidienne $u(x,y) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$, un exemple de graphe topologique GVR parfaitement défini par la matrice d'adjacence V_u associée, formée de 0 et de 1.

Sur le plan géométrique, si $V_u(x,y) = 1$ cela signifie que l'hyper-Lunule (intersection des deux hypersphères de rayon $u(x,y)$ centrées sur les points x et y) est vide.

Pour ce travail, nous avons choisi de travailler avec la structure topologique du GVR, considérée comme un super graphe de ALM et un sous-graphe du GG.

2.2 Mesure de l'équivalence topologique

Etant donné un modèle de classement à p variables explicatives $\{x^j; j = 1, ..., p\}$ et une variable cible à expliquer y, partition de $n = \sum_{k=1}^{q} n_k$ observations réparties en q classes $\{C_k; k = 1, ..., q\}$.

Pour toute mesure de proximité u, on construit selon la propriété (1), la matrice d'adjacence binaire globale V_u, juxtaposition de q matrices d'adjacence intra-classe $\{V_u^k; k = 1, ..., q\}$ symétriques et de $q(q-1)$ matrices d'adjacence inter-classes $\{V_u^{kl}; k \neq l; k, l = 1, ..., q\}$:

$$\begin{cases} V_u^k(x,y) = 1 & si\ u(x,y) \leq \max(u(x,z), u(y,z))\ ;\ \forall x, y, z \in C_k,\ z \neq x, \neq y \\ V_u^k(x,y) = 0 & sinon \end{cases}$$

$$\begin{cases} V_u^{kl}(x,y) = 1 & si\ u(x,y) \leq \max(u(x,z), u(y,z))\ ; \forall x \in C_k, y \in C_l,\ z \in C_l, z \neq y \\ V_u^{kl}(x,y) = 0 & sinon \end{cases}$$

$$V_u = \begin{pmatrix} \mathrm{V}_u^1 & \cdots & \mathrm{V}_u^{1k} & \cdots & \mathrm{V}_u^{1q} \\ & \cdots & & & \\ \mathrm{V}_u^{k1} & \cdots & \mathrm{V}_u^{k} & \cdots & \mathrm{V}_u^{kq} \\ & & & \cdots & \\ \mathrm{V}_u^{q1} & \cdots & \mathrm{V}_u^{qk} & \cdots & \mathrm{V}_u^{q} \end{pmatrix}$$

On note, $V_{u^*} = diag(1_{C_1}, \ldots, 1_{C_k}, \ldots, 1_{C_q})$ la matrice d'adjacence symétrique bloc-diagonale associée à la mesure de proximité inconnue u^* dite de référence ; qui discrimine parfaitement les q classes, où, 1_{n_k} désigne le vecteur d'ordre n_k dont toutes les composantes sont égales à 1 et $1_{C_k} = 1_{n_k}\, {}^t 1_{n_k}$, la matrice carrée d'ordre n_k dont tous les éléments sont égaux à 1.

$$V_{u^*} = \begin{pmatrix} 1_{C_1} & & & & \\ 0 & \cdots & & & \\ 0 & 0 & 1_{C_k} & & \\ 0 & 0 & 0 & \cdots & \\ 0 & 0 & 0 & 0 & 1_{C_q} \end{pmatrix}$$

Mesurer la ressemblance entre une mesure de proximité choisie u et la mesure de proximité de référence u^*, revient à comparer les graphes de voisinage induits par ces deux mesures.

Pour mesurer le degré d'équivalence topologique de discrimination entre les mesures u et u^*, on compare les matrices d'adjacence associées V_u et V_{u^*}, en calculant le critère de concordance ou de similarité suivant :

$$S(V_u, V_{u^*}) = \frac{\sum_{k=1}^{n}\sum_{l=1}^{n} \delta_{kl}}{n^2} \qquad \text{avec} \quad \delta_{kl} = \begin{cases} 1 \text{ si } V_u(k,l) = V_{u^*}(k,l) \\ 0 \text{ sinon.} \end{cases}$$

3 Sélection du sous-ensemble optimal de variables

Plusieurs méthodes ont été proposées pour la sélection de variables, mais le problème reste toujours d'actualité vu qu'aucune méthode n'est vraiment optimale ou meilleure et que les performances de chacune dépendent des données.

L'idée principale de ce travail repose sur l'utilisation d'une structure topologique pour classer les variables par ordre de pertinence afin de déterminer le meilleur sous-ensemble garantissant une meilleure séparabilité des classes des objets. Les trois principales étapes de l'approche proposée ilustrées par la figure 2 sont décrites ci-après :

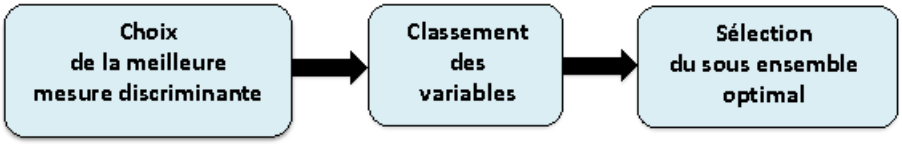


FIG. 2 – *Étapes de la sélection topologique des variables pour la discrimination.*

Étape 1 : Choix de la "meilleure" mesure discriminante

La mesure de proximité utilisée pour la construction du graphe de voisinage a un impact direct sur les résultats obtenus, c'est pour cela qu'il est nécessaire de choisir, parmi les nombreuses mesures de proximité possibles, celle qui est la plus adaptée aux données considérées. Dans un contexte de discrimination, l'approche proposée dans Aazi et Abdesselam (2015), permet d'aider à faire ce choix. Elle consiste à regrouper en classes des mesures de proximité selon leur degré d'équivalence topologique discriminant. La mesure (ou les mesures) appartenant à la classe où a été affectée la mesure de référence u^*, est choisie comme étant la "meilleure" mesure discriminante adaptée aux données.

Étape 2 : Classement des variables par ordre de pertinence

On utilise un score dit d'ordre zéro, Rakotomamonjy (2003), pour classer les variables par ordre de pertinence. Le score d'ordre zéro d'une variable est égal à la valeur d'un critère lorsque cette variable est éliminée ou retirée du modèle. Le critère utilisé ici est le taux de concordance qui correspond au degré d'équivalence topologique de discrimination mesuré entre les matrices d'adjacence V_u et V_{u^*} induites par les mesures de proximités u choisie et u^* de référence.

A partir des p variables de départ, on établit une série de p modèles supervisés à $(p-1)$ variables explicatives : le premier modèle ne contient pas la première variable, le deuxième ne contient pas la deuxième variable et ainsi de suite. Pour chacun de ces p modèles, le critère de concordance est calculé et est associé à la variable retirée du modèle. On obtient ainsi une série de p valeurs de ce critère, laquelle est ensuite ordonnée par ordre croissant, ce qui nous donne un classement des p variables par ordre de pertinence. Plus la valeur du critère de concordance correspondant à la variable éliminée est faible, plus cette variable contribue à l'augmentation de ce taux et donc plus elle est pertinente.

Étape 3 : Choix du sous-ensemble optimal de variables

Après avoir obtenu le classement par ordre de pertinence des variables, la dernière étape consiste à choisir le sous-ensemble optimal. Pour cela, on construit en utilisant la méthode pas à pas ascendante 'forward', une suite de p sous-ensembles de modèles dont le premier contient la première variable pertinente, le deuxième contient les deux premières variables pertinentes et ainsi de suite jusqu'à ce qu'on intègre toutes les variables une par une dans l'ordre décroissant de pertinence. Pour chaque sous-ensemble i, $\{i = 1, ..., p\}$, on détermine la matrice d'adjacence correspondante $\{V_u(i)\ ;\ i = 1, ..., p\}$. On calcule ensuite le taux de concordance entre la matrice d'adjacence induite par chaque sous-ensemble $V_u(i)$ avec la matrice d'adjacence de référence V_{u^*}. Le modèle dont le taux de concordance est le plus élevé, est choisi comme étant le meilleur modèle contenant le nombre optimal de variables.

4 Expérimentations

Dans cette section sont présentés les résultats des tests effectués, sur des données simulées et réelles, pour évaluer les performances de l'approche STVD en utilisant la structure topologique du GVR. Les caractéristiques des données utilisées sont consignées dans le tableau 1. Les simulations et les résultats de ces expérimentations sont obtenus en utilisant le langage de programmation Matlab.

Nom	Type	$X_{(n\times p)}$	$Y_{(q)}$
Exemple 1	Simulées	201 × 20	3
Exemple 2.1	Données simulées	201 × 500	3
Exemple 2.2	Données simulées	201 × 1000	3
Exemple 3	Données simulées	201 × 20	3
Exemple 4	Données réelles - Sonar	208 × 60	2

TAB. 1 – *Jeux de données simulées et réelles.*

4.1 Données simulées

Les ensembles de données simulées des exemples 1 et 2 sont générés de telle sorte que les deux premières variables soient pertinentes pour la discrimination et les autres soient considérées comme des variables bruit. Pour cela, comme décrit dans Wang et Shen (2006), on génère un ensemble $\{v^j\ ;\ j = 1, ..., p\}$ de p variables normalement distribuées, centrées et réduites, puis on affecte les observations dans $q = 3$ classes à discriminer, équipondérées (chaque classe avec le même nombre d'observations). On procède ensuite à la transformation linéaire suivante :

$$\begin{cases} x^j = v^j + a_j \ \ pour \ \ j = 1, 2 \\ x^j = v^j \ \ pour \ \ j = 3, ..., p. \end{cases}$$

avec, $(a_1, a_2) = (\sqrt{2}, \sqrt{2})$, $(-\sqrt{2}, -\sqrt{2})$ et $(\sqrt{2}, -\sqrt{2})$ pour respectivement les classes 1,2 et 3. Ainsi, avec cette transformation, les deux premières variables x^1 et x^2 sont les seules qui sont pertinentes pour la discrimination.

A noter que la mesure de proximité de Tchebychev a été choisie comme étant la meilleure mesure discriminante pour les 4 ensembles de données simulées.

• **Exemple 1**

Pour ce premier ensemble de données, on a généré $n = 201$ observations et $p = 20$ variables dont seules les deux premières sont pertinentes, les 18 autres sont considérées comme variables bruit.

L'objectif est de vérifier si le score calculé pour chaque variable en utilisant le GVR, qui correspond au taux de concordance quand cette variable est éliminée ou retirée du modèle, permet de retrouver l'ordre correct de pertinence des variables, autrement dit, d'affecter des scores minimums aux deux premières variables. Les principaux résultats sont présentés dans le tableau 2

Variable retirée	Score	Classement
x^1	**26268**	**2**
x^2	**26234**	**1**
x^3	26574	18
x^4	26547	11
x^5	26501	3
x^6	26555	13
x^7	26572	17
x^8	26524	7
x^9	26593	20
x^{10}	26540	9
x^{11}	26540	10
x^{12}	26558	14
x^{13}	26586	19
x^{14}	26558	15
x^{15}	26548	12
x^16	26512	5
x^{17}	26509	4
x^{18}	26514	6
x^{19}	26571	16
x^{20}	26538	8

TAB. 2 – *Scores et classements des données simulées 1.*

La variable la plus importante est celle dont la valeur du taux de concordance calculé est minimale quand elle est retirée du modèle. Au vu des résultats des scores et de classement obtenus, la procédure proposée a donc bien réussi à classer les deux premières variables, qui sont les plus importantes, au deux premiers rangs. Quant au sous-ensemble optimal recherché, il est dans ce cas constitué des trois premières variables.

• **Exemples 2.1 & 2.2**

Pour ces deux exemples, nous avons ajouté des variables bruit au premier ensemble de données pour mesurer l'effet grande dimension des données (nombre important de variables) sur les performances du score.

Deux ensembles de données sont donc considérés, le premier contient $p = 500$ variables et le second $p = 1000$ variables. Dans les deux cas, seules les deux premières variables x^1 et x^2 sont considérées comme pertinentes. Les taux de concordance de classement des variables

par ordre de pertinence sont représentés sur les deux graphiques de la figure 3 avec des zoom sur les premières variables.

Ces résultats confirment l'efficacité de la procédure topologique de classement des variables même en grande dimension. En effet, pour ces deux cas, les deux variables pertinentes occupent les deux premiers rangs.

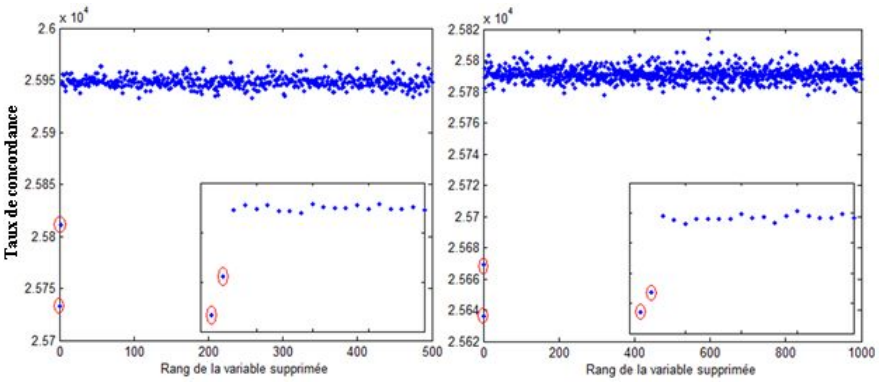


FIG. 3 – *Scores des variables en grande dimension.*

• **Exemple 3**

Pour les deux exemples précédents, les deux variables pertinentes générées le sont pour toutes les $q = 3$ classes considérées.

Variable retirée	Score	Classement
x^1	**26035**	**1**
x^2	26347	5
x^3	**26177**	**2**
x^4	26377	17
x^5	26365	15
x^6	26331	3
x^7	26365	16
x^8	26343	4
x^9	26381	18
x^{10}	26355	9
x^{11}	26357	10
x^{12}	26353	7
x^{13}	26360	12
x^{14}	26350	6
x^{15}	26393	20
x^{16}	26360	13
x^{17}	26353	8
x^{18}	26358	11
x^{19}	26360	14
x^{20}	26385	19

TAB. 3 – *Scores et classements des données simulées.*

D'une façon générale en discrimination, certaines variables peuvent être plus pertinentes pour différencier et séparer au mieux certaines classes et non pertinentes pour les autres classes. Un nouvel ensemble de données est généré pour étudier ce cas et pour vérifier si l'approche topologique proposée pour le classement des variables réussit à identifier les variables pertinentes et les classer selon l'ordre connu *a priori*.

L'exemple 3 est constitué d'un ensemble de $n = 201$ observations réparties en $q = 3$ classes équipondérées et $p = 20$ variables dont la variable x^1 comme variable pertinente pour différencier les classes 2 et 3 et la variable x^3 pertinente pour distinguer la classe 1 uniquement.

Les résultats obtenus présentés dans le tableau 3, confirment là aussi, le bon ordre de classement des variables, à savoir, la première variable en première position, suivie de la troisième variable puis des dix-huit autres variables bruit.

4.2 Données réelles

Pour illustrer les performances de l'approche proposée à partir de données réelles, nous avons considéré les données de la base Sonar, UCI (2013), constituée de $p = 60$ variables et $n = 208$ observations réparties dans $q = 2$ classes.

Des comparaisons sont aussi faites avec une méthode de classement de variables par ordre de pertinence, largement utilisée dans le contexte de la discrimination, Dudoit et al. (2002), en termes du choix du sous-ensemble optimal ainsi qu'en termes des taux de concordance obtenus pour chaque sous-ensemble de variables construit.

Les 208 observations ont été réparties en 150 observations d'apprentissage (permettant de choisir la meilleure mesure de proximité et d'obtenir le classement des variables par ordre de pertinence) et 58 observations de validation pour le choix du sous-ensemble optimal de variables.

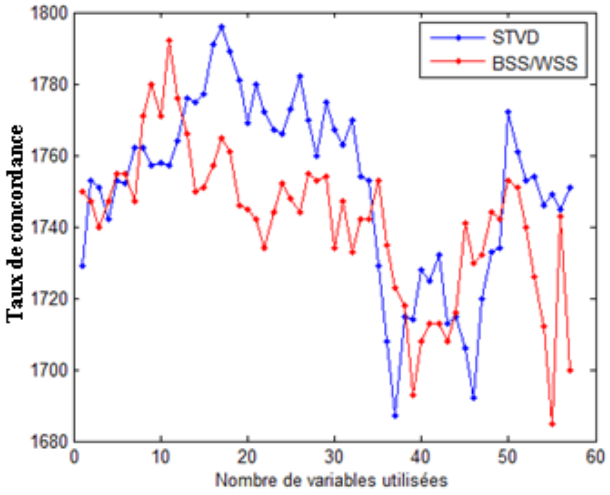


FIG. 4 – *Comparaison des approches.*

La mesure de proximité de Mahalanobis a été choisie comme étant la "meilleure" mesure discriminante pour ces données.

Après avoir classé les variables par ordre de pertinence par l'approche STVD et par la méthode de classement selon le critère BSS/WSS [1], Chia et al. (2005); Sehgal et al. (2006), les taux de concordance obtenus par ces deux approches en fonction du nombre de variables sont représentés sur la figure 4.

Le meilleur taux de concordance est obtenu par l'approche STVD en utilisant les 17 premières variables par ordre décroissant de pertinence. Ainsi, en termes des taux de concordance obtenus en fonction du nombre de variables utilisées, ceux de l'approche STVD sont pour la majorité des sous-ensembles, meilleurs que ceux de l'approche BSS/WSS.

5 Conclusion et perspectives

Dans ce travail, nous avons proposé une nouvelle approche de sélection de variables dans un contexte de discrimination en utilisant une structure topologique. Testée sur des données simulées et réelles, la méthode STVD a réussi à bien classer les variables pertinentes de tous les exemples de données simulées considérés. Sur des données réelles, elle donne un meilleur sous-ensemble que celui obtenu par la méthode BSS/WSS.

Pour le choix du sous-ensemble optimal de variables, nous avons utilisé le critère du taux de concordance, cependant à partir de l'ordre des variables obtenus, on peut aussi utiliser un modèle de discrimination tel que les SVM pour choisir le sous-ensemble optimal en utilisant cette fois le taux de reconnaissance comme critère de choix.

D'un point de vue pratique, nous avons utilisé ici des exemples de données explicatives continues, mais ce travail peut parfaitement s'étendre à d'autres types de données (binaires, qualitatives, floues, etc.) avec la bonne mesure de proximité adaptée aux données.

L'approche topologique proposée pour la sélection de variables discriminantes est une technique qui peut être très utile en modélisation prédictive - Science des données (Data science) en présence de données massives - Datamasse (Big Data).

Références

Aazi, F.-Z. et R. Abdesselam (2015). Choix d'une mesure de proximité discriminante dans un contexte topologique. *RNTI - Revue des nouvelles Technologies de l'Information, E.28, Hermann Editions, EGC-2015*, 101–112.

Abdesselam, R. (2014). Proximity measures in topological structure for discrimination. *In a Book Series SMTDA-2014, 3nd Stochastic Modeling Techniques and Data Analysis, International Conference, Lisbon, Portugal, C.H. Skiadas (Ed), ISAST*, 599–606.

Abdesselam, R. et D. Zighed (2011). Comparaison topologique de mesures de proximité. *Actes des XVIIIème Rencontres de la Société Francophone de Classification*, 79–82.

Batagelj, V. et M. Bren (1992). Comparing resemblance measures. Technical report, Proc. International Meeting on Distance Analysis (DISTANCIA'92).

1. Modification du F-ratio de Fisher (ANOVA). Rapport de 2 variances, l'Inter-classe sur l'Intra-classe. Indique la capacité à discriminer entre plusieurs classes. BSS : "Between Sum of Squares", WSS : "Within Sum of Squares".

Batagelj, V. et M. Bren (1995). Comparing resemblance measures. *Journal of classification 12*, 73–90.

Chia, H., C. Madhu, et W. Shy (2005). Redevance, redundancy and differential prioritization in feature selection. *Biological and Medical Data Analysis: 6th International Symposium, ISBMDA LNBI 3745, Springer-Verlag Berlin Heidelberg*, 367–378.

Dudoit, S., J. Fridlyand, et T. Speed (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of American Statistical Association 97(457)*, 77–87.

Guyon, I., J. Weston, S. Barnhill, et V. Vapnik (2002). Gene selection for cancer classification using support vector machines. *Machine Learning 46(1-3)*, 389–422.

Kim, J. et S. Lee (2003). Tail bound for the minimal spanning tree of a complete graph,. *Statistics Probability Letters, 64*(4), 425–430.

Kohavi, R. et G.-H. John (1997). Wrappers for feature subset selection. *Artificial intelligence 97(1)*, 273–324.

Koller, D. et M. Sahami (1996). Toward optimal feature selection. *in ICML'96*, 284–292.

Lesot, M.-J., M. Rifqi, et H. Benhadda (2009). Similarity measures for binary and numerical data: a survey. *IJKESDP 1*(1), 63–84.

Liu, H., D. Song, S. Ruger, R. Hu, et V. Uren (2008). Comparing dissimilarity measures for content-based image retrieval. *Information Retrieval Technology*, 44–50.

Malerba, D., F. Esposito, V. Gioviale, et V. Tamma (2001). Comparing dissimilarity measures for symbolic data analysis. *Proceedings of Exchange of Technology and Know-how and New Techniques and Technologies for Statistics 1*, 473–481.

Malerba, D., F. Esposito, et M. Monopoli (2002). Comparing dissimilarity measures for probabilistic symbolic objects. *Series Management Information Systems 6*, 31–40.

Park, J., H. Shin, et B. Choi (2006). Elliptic Gabriel graph for finding neighbors in a point set and its application to normal vector estimation. *Computer-Aided Design 38*(6), 619–626.

Peng, H., F. Long, et C. Ding (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 27(8)*, 1226–1238.

Rakotomamonjy, A. (2003). Variable selection using SVM-based criteria. *The Journal of Machine Learning Research 3*, 1357–1370.

Richter, M. (1992). Classification and learning of similarity measures. *Proceedings der Jahrestagung der Gesellschaft fur Klassifikation, Studies in Classification, Data Analysis and Knowledge Organisation. Springer Verlag*.

Rifqi, M., M. Detyniecki, et B. Bouchon-Meunier (2003). Discrimination power of measures of resemblance. *IFSA'03*.

Sehgal, M., I. Gondal, et L. Dooley (2006). Missing value imputation framework for microarray significant gene selection. *Data Mining for Biomedical Applications: PAKDD 2006 Workshop, BioDM 2006, Singapore Proceedings LNBI 3916*, 131–142.

Spertus, E., M. Sahami, et O. Buyukkokten (2005). Evaluating similarity measures: a large-scale study in the orkut social network. In *Proceedings of the eleventh ACM SIGKDD*

international conference on Knowledge discovery in data mining, pp. 684. ACM.

Toussaint, G. (1980). The relative neighbourhood graph of a finite planar set. *Pattern recognition 12*(4), 261–268.

UCI (2013). UCI machine learning repository, [http://archive.ics.uci.edu/ml]. irvine, CA: University of California, school of information and computer science.

Wang, L. et X. Shen (2006). Multi-category support vector machines, feature selection and solution path. *Statistica Sinica 16*, 617–633.

Wang, L. et X. Shen (2007). On l1-norm multi-class support vector machines: methodology and theory. *Journal of the American Statistical Association 102*, 583–594.

Warrens, M. (2008). Bounds of resemblance measures for binary (presence/absence) variables. *Journal of Classification 25*(2), 195–208.

Zhang, H.-H., Y. Liu, Y. Wu, et J. Zhu (2008). Variable selection for the multicategory svm via adaptive sup-norm regularization. *Electronic Journal of Statistics 2*, 149–167.

Zighed, D., R. Abdesselam, et A. Hadgu (2012). Topological comparisons of proximity measures. *The 16th PAKDD 2012 Conference. In P.-N. Tan et al. (Eds.) Part I, LNAI 7301, Springer-Verlag Berlin Heidelberg*, 379–391.

Summary

In machine learning, the presence of a large number of explanatory variables leads to a high complexity of algorithms and a strong degradation of the performance of prediction models. In this case, a selection of an optimal discriminant subset of these variables is necessary. In this article, a topological approach is proposed for the selection of this optimal subset. It uses the concept of neighborhood graph for ranking variables in order of relevance, then the forward method is applied to construct a series of models among which the best subset is selected based on the degree of topological equivalence in discrimination. For each subset, the degree of equivalence is measured by comparing the adjacency matrix induced by the proximity measure selected to that induced by the "best" discriminant proximity measure called of reference. The performance of this approach is evaluated using simulated and real data. Comparisons of the results of variables selection in discrimination with those of a metric approach show a much better selection using the proposed topological approach.