Sélection de variables par optimisation multi-objective de l'information mutuelle

MOHAMED ABADI[1], ENGUERRAND GRANDCHAMP[2], OLIVIER ALATA[1], CHRISTIAN OLIVIER[1], MAJDI KHOUDEIR[1]

[1] Institut XLIM-SIC, UMR CNRS 6172
Université de Poitiers, BP 30179, 86962 Futuroscope-Chasseneuil, Cedex France
[2] Laboratoire LAMIA
Université des Antilles et de la Guyane, Campus de Fouillole, 97157 Pointe-à-Pitre Guadeloupe, France

[1]{abadi,alata,olivier,khoudeir}@sic.sp2mi.univ-poitiers.fr,
[2]egrandch@univ-ag.fr

Résumé - Ce travail propose une approche originale utilisant conjointement l'information mutuelle et la courbe de Pareto pour la sélection de variables. L'information mutuelle est utilisée pour estimer la dépendance des variables par rapport aux classes et la redondance entre les variables prises deux à deux. Contrairement à certains travaux, ces critères sont utilisés simultanément pour calculer la courbe de Pareto et ainsi déterminer l'ensemble de variables optimal. Des tests sur des données de référence sont réalisés. Ils montrent par le biais des taux de bonne classification, déterminés en appliquant plusieurs algorithmes de classification, l'importance d'un tel outil et sa capacité à choisir les ensembles de variables qui caractérisent mieux les classes vis-à-vis de données.

Abstract - This work proposes an original approach using mutual information and Pareto curve jointly for feature selection. Mutual information is used to estimate dependency between features and classes and redundancy between features taken two by two. Unlike some studies, these criteria are used simultaneously to compute Pareto curve and determine the optimal feature set. Our approach is tested on different reference data. Several clustering algorithms are used to compute classification accuracy. The obtained results show the importance of our tool and its ability to select the best feature sets that give the better description.

1 Introduction

La thématique sélection de variables est développée pour apporter des solutions aux problèmes de l'exploration et de l'interprétation de données. Ces données se caractérisent par un nombre de variables et/ou de réalisations très conséquent. En effet, son objectif principal est de trouver l'ensemble de variables le plus adapté pour décrire en un temps convenable les données. Contrairement aux algorithmes d'apprentissage qui prennent beaucoup de temps et dont l'utilisation est impossible pour un grand nombre d'applications (par exemple : caractérisation du texte [1], recherche d'images [2], bioinformatiques [3], classification d'images couleur [4], …).

Les processus de sélection, d'un ou plusieurs sous-ensembles optimaux à partir de l'ensemble formé par tous les attributs, se fait par trois types d'approches : *wrapper* [5], *filter* [6] et *embedded* [7].

Les méthodes *wrappers* utilisent les algorithmes de classification pour générer et ensuite évaluer la qualité des ensembles candidats. Cette approche est généralement pertinente mais elle dépend de la représentativité de l'ensemble d'apprentissage.

Les méthodes *filters* sont basées sur une fonction critère pour mesurer la pertinence de l'information contenue dans les ensembles candidats. Elles réduisent considérablement le temps de calcul alors que les méthodes *wrappers* permettent d'obtenir des meilleurs résultats.

Les méthodes *embedded* tentent de combiner les avantages des précédentes approches. Elles intègrent la phase de sélection des variables dans l'étape d'apprentissage. Néanmoins, le temps de calcul reste important. C'est probablement la raison principale pour laquelle les méthodes *filters* sont les plus populaires.

Afin d'avoir le meilleur compromis entre les contraintes temps et la qualité des résultats, nous nous intéressons donc, dans ce travail, aux méthodes *filters*. Ainsi différents critères ont été développés pour évaluer les sous-ensembles candidats. Ils explorent des mesures de distance [8], des mesures statistiques [9, 4] ou plus récemment des mesures probabilistes basées sur l'information mutuelle [10, 11, 12].

Dans ce travail, nous proposons un algorithme de sélection de variables basé sur l'optimisation multi-objective [13] utilisant l'information mutuelle pour décrire simultanément la pertinence des variables par rapport aux classes et la redondance entre les variables prises deux à deux. L'utilisation de la pertinence est un moyen efficace qui permet de contourner la difficulté à mettre en œuvre la propriété de dépendance entre les variables et les classes [10]. La recherche du meilleur ensemble de variables se fait alors via la courbe de Pareto qui met en relation la pertinence et la redondance des variables. L'algorithme de sélection est testé sur plusieurs jeux de données de nature différente.

Notre contribution dans le domaine de la sélection de variables est introduite dans la section 2. Nous nous sommes intéressés, en premier lieu, à rappeler les deux concepts de base utilisés par l'information mutuelle, à savoir la dépendance et la redondance. Ensuite nous décrivons comment ces derniers sont combinés pour choisir les variables optimales. Enfin, l'évaluation de notre approche et la comparaison avec d'autres méthodes sont réalisées dans la section 4. Les conclusions et les perspectives sont présentées dans la section 5.

2 Algorithmes de sélection des variables

La sélection de variables permet de réduire le coût de calcul, le nombre de variables et préserve la qualité à des fins de reconnaissance et/ou de classification. Les m variables sélectionnées sont choisies parmi les M variables initiales ($m \ll M$). Elles doivent conserver leur dépendance avec les classes. L'approche la plus utilisée pour vérifier cette propriété est de calculer la pertinence entre les variables et les classes (c). Elle peut être exprimée par l'information mutuelle.

2.1 Information mutuelle et critères de sélection

Soient X et Y deux variables aléatoires, l'information mutuelle $I(X;Y)$ est définie par $P(x)$, $P(y)$ et $P(\mathrm{x},\mathrm{y})$ (lois de probabilité discrètes) :

$$I(X;Y) = \sum_{y\in Y}\sum_{x\in X} P(x,y)\cdot \log\frac{P(x,y)}{P(x)\cdot P(y)} \qquad (1)$$

Rappelons que $I(X;Y)$ est élevée si X et Y sont dépendantes et dans le cas où $I(X;Y)=0$ cela signifie que les variables sont indépendantes.

2.1.1 Estimation de l'information mutuelle

L'information mutuelle est considérée comme un bon indicateur de la pertinence et de la redondance entre les variables aléatoires. Son estimation est simple pour des variables discrètes car les probabilités conjointe et marginale peuvent être estimées en comptabilisant les échantillons représentatifs des variables [15]. Néanmoins, lorsqu'au moins une des variables est continue, l'information mutuelle demeure difficile à calculer, car elle dépend forcement de la technique utilisée pour approcher la densité de probabilité. Pour certaines applications, une étape de prétraitement consistant à discrétiser les données peut être envisagée. Cependant, cette discrétisation n'est pas toujours simple car elle dépend forcement des données. Kwak et al. [16] proposent d'utiliser la fenêtre de Parzen (eq. 4) comme moyen pour estimer la densité de probabilité et ainsi approcher $I(X;Y)$:

$$\hat{p}(x) = \frac{1}{N}\sum_{i=1}^{N}\delta(x-x_i,h) \qquad (4)$$

où $\delta(.)$ représente la fenêtre de Parzen. x_i est le ième échantillon, N est le nombre d'échantillons et h est la largeur de la fenêtre de Parzen. Parzen [17] a montré que si $\delta(.)$ et h sont bien choisies alors $\hat{p}(x)$ converge généralement vers $p(x)$ lorsque N tend vers l'infini.

2.1.2 Définition des critères de sélection

La pertinence D est choisie comme la valeur maximale de l'information mutuelle entre les variables X_i prises individuellement et les classes (c).

$$D = \max_{1,\cdots,M}\frac{1}{|S|}\sum_{X_i\in S} IM(X_i;c) \qquad (2)$$

où $|S|$ est la taille de l'ensemble sélectionné.

Il a été reconnu que la sélection des meilleures variables, en utilisant uniquement D ne retourne pas nécessairement la solution optimale car les variables peuvent être redondantes entre elles [11]. Différents moyens existent pour vérifier la redondance R entre les variables [10]. Celle-ci peut être définie aussi par l'information mutuelle entre deux variables $(X_i;X_j)$ avec $i,j=1,\cdots,|S|$ et $i\neq j$.

$$R = \min_{1,\cdots,M}\frac{1}{|s|^2}\sum_{X_i,X_j\in S} IM(X_i;X_j) \qquad (3)$$

La redondance minimale permet de sélectionner le couple des variables le moins redondant. Peng et Ding [10, 11] ont montré que les critères qui combinent ces propriétés sélectionnent les variables qui améliorent le taux de bonne classification. Ils ont ainsi proposé le critère *mRMR*[1] qui optimise (2) et (3), (par exemple : $\max_{1,\cdots,M}(D-R)$ ou $\max_{1,\cdots,M}\left(\frac{D}{R}\right)$). Il s'agit en réalité de critères mono-objectifs qui ne garantissent pas la solution optimale car la convergence simultanée de (2) et (3) est non assurée.

2.2 Méthode proposée pour la sélection de variables

Pour remédier à cet inconvénient, nous proposons un algorithme multi-objectif permettant d'optimiser conjointement D et R. Cette approche fournit un ensemble de solutions dites Pareto optimales c'est-à-dire les combinaisons ayant un couple (D,R) non dominé [18].

2.3 Courbe de Pareto

L'optimisation multi-objective consiste à trouver des solutions qui optimisent plusieurs critères. Dans ce contexte, on définit la courbe de Pareto comme étant l'ensemble des solutions qui ne sont pas dominées par d'autres solutions.

Soit $C_1,\cdots,C_p,\cdots,C_P$ un ensemble de P critères à maximiser. On dira qu'un élément ξ_1 domine un élément ξ_2 selon C_p si $C_p(\xi_1)\geq C_p(\xi_2)$.

Une solution ξ_1 est non dominée (appartient à la courbe de Pareto) si : $\nexists\ \xi_2 :\ \forall\, p\ C_p(\xi_2)\geq C_p(\xi_1)$. Dans notre contexte, nous cherchons la ou les combinaisons d'attributs S^* qui maximise la pertinence D et qui minimise la redondance R. La courbe de Pareto se définit donc comme suit :

$$\left\{S \mid \nexists\, S' :\ \left(\mathrm{D}(S') > \mathrm{D}(S)\right) \text{ et } \left(\mathrm{R}(S') \leq \mathrm{R}(S)\right)\right\} \qquad (5)$$

[1] maximal relevance minimal redundancy

3 Expérimentations

Les performances de notre approche sont évaluées et comparées sur sept jeux de données de références [2] (IRIS, TAE : Teaching Assistant Evaluation, WINE, STAT1 : image segmentation, STAT2 : Landsat image, SPAM, MFEAT : Multiple features). Ces données diffèrent principalement par leur nombre de réalisations, de variables et de classes. Le tableau (1) dresse les caractéristiques de ces jeux de données.

Tab 1 : Caractéristiques des jeux de données

	Variables	Réalisations	Classes	CV
IRIS	4	150	3	10-Fold
TAE	5	151	3	10-Fold
WINE	13	178	3	10-Fold
STAT1	18	2310	7	10-Fold
STAT2	32	6435	6	Holdout
SPAM	57	4601	2	Holdout
MFEAT	649	2000	10	Holdout

Nous avons varié le mode de la validation croisée (CV) pour construire les ensembles d'apprentissage et de test. Nous avons utilisé 10-Fold et Holdout à 50% (Tab 1). Le but de cette opération est d'obtenir des résultats statistiquement corrects et un jugement objectif.

Trois classifieurs sont utilisés pour calculer les pourcentages de bonne classification des ensembles de variables sélectionnés. A savoir kNN ($k = 1$, k-Neareast Neighbor), *NB* (Naive Bayes), *LDA* (Linear Discriminant Analysis). Ceci permet d'évaluer les ensembles sélectionnés, d'étudier et de comparer les critères de sélection de variables.

Les techniques de sélection de variables utilisent une fonction critère pour mesurer la qualité des ensembles afin de sélectionner le candidat optimal et une procédure de parcours pour la construction des ensembles candidats. Différents schémas de parcours de variables ont été développés [14]. Pour des raisons de simplicité et de représentation, nous avons choisi de ne présenter dans ce travail que les résultats issus du schéma de parcours SFS (Séquential Forward Selection).

Les tableaux 2, 3 et 4 présentent une comparaison quantitative entre les résultats obtenus par notre approche et ceux retournés par les algorithmes *mRMR* [10] et *FSDD* [8] appliqués sur les sept jeux de données du tableau 1. Il s'agit précisément de la moyenne du pourcentage de bonne classification (μ), de l'écart type (σ) et le nombre de variables (#) qui compose l'ensemble sélectionné. Nous signalons que (#) est déterminé en calculant la fréquence d'apparition de (#) pour toutes les itérations. Le nombre de variables optimales est celui qui possède le maximum de fréquence d'apparition. Ces tableaux montrent que les pourcentages de bonne classification (μ) de notre approche sont meilleurs pour la plupart des jeux de données. L'avantage est que ces pourcentages sont obtenus avec des ensembles de variables possédant généralement des (#) inférieurs aux autres méthodes.

[2] Base de données UCI Repository : http://archive.ics.uci.edu/ml/datasets.html

Le tableau 5 montre les pourcentages de bonne classification obtenus par les classifieurs *k-NN*, *LDA* et *NB* en utilisant notre critère, les critères *mRMR* [10] et *FSDD* [8] et suivant la dimension de l'ensemble optimal sélectionné. Nous constatons que les trois critères ont choisi la même variable pour un ensemble optimal à une dimension. Notre critère trouve la meilleure seconde variable formant ainsi le meilleur ensemble optimal à deux dimensions contrairement aux autres critères. En effet, la qualité des classifieurs utilisés explique la différence des pourcentages de bonne classification obtenus et non le choix de la variable puisque l'ensemble des variables obtenus ne change pas en appliquant notre approche.

Tab 2 : Pourcentages de bonne classification obtenus par le classifieur *K-NN*

	Notre approche			***mRMR***			***FSDD***		
	μ	σ	#	μ	σ	#	μ	σ	#
IRIS	**97.43**	0.56	**2**	95.92	0.22	4	96.61	0.69	2
TAE	**64.28**	1.81	**2**	60.37	2.04	5	60.37	2.04	5
WINE	**80.32**	1.26	12	78.33	1.37	**8**	77.70	1.19	11
STAT1	95.12	0.46	16	95.12	0.46	16	**95.21**	0.47	**12**
STAT2	**90.05**	--	**25**	**90.05**	--	28	89.85	--	27
SPAM	**90.82**	--	54	90.43	--	53	90.48	--	**43**
MFEAT	**89.90**	--	**47**	84.70	--	48	89.80	--	**47**

Tab 3 : Pourcentages de bonne classification obtenus par le classifieur *LDA*

	Notre Approche			***mRMR***			***FSDD***		
	μ	σ	#	μ	σ	#	μ	σ	#
IRIS	**98.53**	0.66	**2**	98.13	0.16	4	98.13	0.16	4
TAE	**54.77**	1.66	**3**	53.79	0.85	4	53.98	0.78	4
WINE	**99.05**	0.44	12	99.02	0.46	**6**	98.81	0.47	13
STAT1	**91.30**	0.24	**18**	90.25	0.24	18	90.61	0.47	18
STAT2	**84.05**	--	35	84.00	--	**34**	**84.05**	--	35
SPAM	89.83	--	**48**	**89.87**	--	53	89.48	--	57
MFEAT	**98.50**	--	**47**	**98.50**	--	**47**	98.30	--	48

Tab 4 : Pourcentages de bonne classification obtenus par le classifieur *NB*

	Notre Approche			***mRMR***			***FSDD***		
	μ	σ	#	μ	σ	#	μ	σ	#
IRIS	**97.40**	0.60	**2**	95.87	0.00	4	95.87	0.23	3
TAE	**55.15**	1.02	**4**	54.97	0.91	4	54.83	1.07	5
WINE	**98.96**	0.45	12	95.65	0.54	**6**	95.45	0.34	13
STAT1	**87.82**	0.41	**17**	87.39	0.57	18	87.43	0.58	**17**
STAT2	78.70	--	32	**79.20**	--	**27**	**79.20**	--	**27**
SPAM	**89.91**	--	**52**	89.30	--	56	89.35	--	53
MFEAT	**97.20**	--	**29**	**97.20**	--	**29**	95.70	--	32

Tab 5 : Pourcentages de bonne classification obtenus pour chaque dimension sur le jeu de données IRIS

Classifieurs	**Critères**	**Dimension des ensembles**			
		1	**2**	**3**	**4**
k-NN	**Notre approche**	91.67	**97.43**	96.8	95.92
	mRMR	91.67	92.47	94.93	95.92
	FSDD	91.67	96.61	95.06	95.92
LDA	**Notre approche**	94.47	**98.80**	**98.80**	98.13
	mRMR	94.47	94.87	96.67	98.13
	FSDD	94.47	95.80	97.40	98.13
NB	**Notre approche**	95.33	**97.40**	96.60	95.87
	mRMR	95.33	94.13	95.67	95.87
	FSDD	95.33	95.73	95.87	95.87

Les figures 1.a, 1.b et 1.c présentent respectivement pour les classifieurs *k-NN*, *LDA* et *NB* le pourcentage de bonne classification pour chaque dimension représentant l'ensemble optimal sélectionné à partir du jeu de

données STAT1. Les courbes montrent une supériorité de notre approche par rapport aux autres critères pour les différents classifieurs testés. Nous constatons aussi que notre critère converge plus rapidement vers les solutions optimales avant de se stabiliser. Cette solution est obtenue avec une dimension moyennement faible. Cette convergence rapide ainsi que la stabilité est garantie par l'utilisation simultanée de la pertinence et de la redondance des variables.

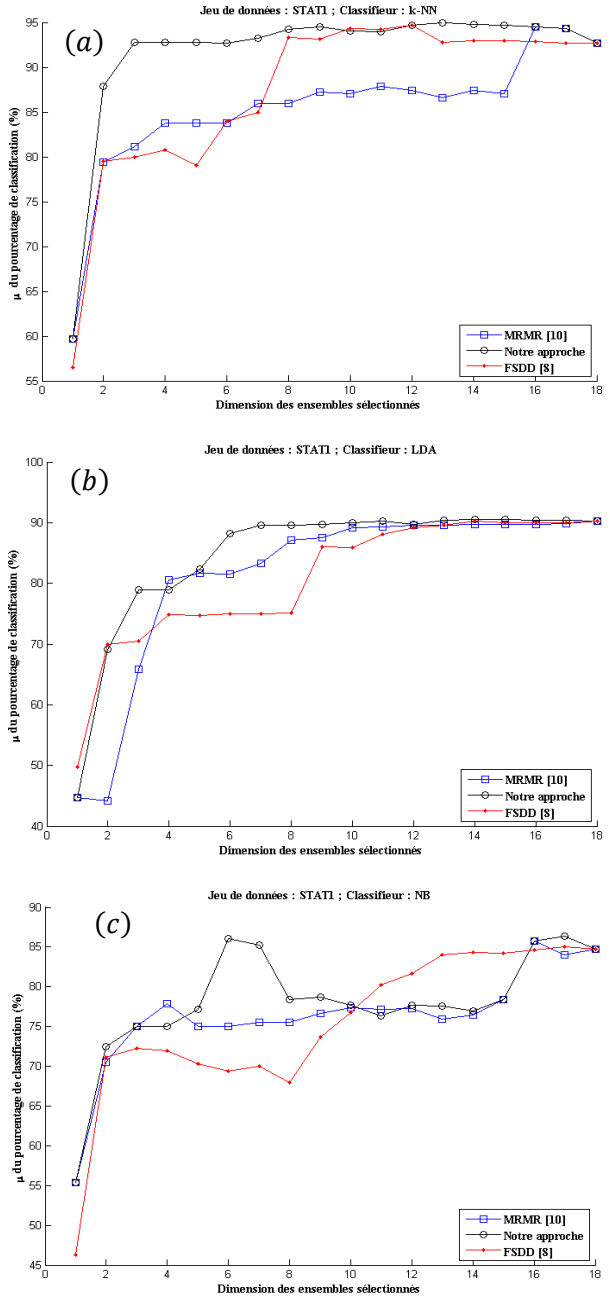

Figure 1 : moyennes du pourcentage de bonne classification

4 Conclusions

Nous avons présenté une approche originale basée sur l'optimisation multi-objective de l'information mutuelle pour sélectionner la combinaison optimale de variables. Il s'agit d'une approche multi-objective qui conserve les solutions qui maximisent la pertinence entre les variables prises individuellement et les classes et qui minimisent la redondance entre les variables prises deux à deux. Les solutions non dominées, c'est-à-dire celles qui se trouvent sur la courbe de Pareto, représentent l'ensemble des combinaisons optimales. Les tests réalisés sur des données références et les comparaisons avec d'autres algorithmes ont montré la bonne tenue de notre approche.

Références

[1] G. Forman, An extensive empirical study of feature selection metrics for text classification, Journal of Machine Learning Research 3 (2003) 1289–1305.

[2] J.G. Dy, C.E. Brodley, A.C. Kak, L.S. Broderick, A.M. Asien, Unsupervised feature selection applied to content-based retrieval of lung images, IEEE TPAMI 25 (3) (2003) 373–378.

[3] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (19) (2007) 2507–2517.

[4] A. Porebski, N. Vandenbroucke, L. Macaire, Comparison of feature selection schemes for color texture classification, IPTA Paris, July 2010.

[5] R. Kohavi and G. John, Wrapper for Feature Subset Selection, Artificial Intelligence 97 (1-2) (1997) 273-324.

[6] P. Pudil and J. Novovicova, Novel methods for subset selection with respect to problem knowledge, IEEE Intell. Syst. 13 (2) (1998) 66–74.

[7] Y. Sun, S. Todorovic, and S. Goodison, Local-Learning-Based Feature Selection for High-Dimensional Data Analysis, TPAMI 32 (9) (2010) 1610–1626.

[8] Jianning Liang, Su Yang, Adam C. Winstanley, Invariant optimal feature selection: A distance discriminant and feature ranking based solution. Pattern Recognition 41 (5) (2008) 1429-1439.

[9] G. Qu, S. Hariri, M. Yousif, A new dependency and correlation analysis for features, IEEE Transactions on Knowledge and Data Engineering 17 (9) (2005) 1199–1207.

[10] H. Peng, F. Long, C. Ding, Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, TPAMI 27 (8) (2005) 1226-1238.

[11] A. Al-Ani, M. Deriche, J. Chebil, A new mutual information based measure for feature selection, Intelligent Data Analysis 7 (1) (2003) 43–57.

[12] H. Liu, J. Sun, L. Liuand, H. Zhang, Feature selection with dynamic mutual information, Pattern Recognition 42 (7) (2009) 1330-1339.

[13] J. Handl and J. Knowles, Feature Subset Selection in Unsupervised Learning via Multiobjective Optimization, IJCIR 2 (3) (2006) 217–238.

[14] H. Liu and L. Yu, Toward integrating feature selection algorithms for classification and clustering, IEEE Transactions on Knowledge and Data Engineering, 17 (3) (2005) 491–502.

[15] Yao, Y. Y. Information-theoretic measures for knowledge discovery and data mining, Entropy Measures, Maximum Entropy Principle and Emerging Applications, Karmeshu (ed.), Springer, pp. 115-136, 2003.

[16] N. Kwak and C.H. Choi, "Input Feature Selection by Mutual Information Based on Parzen Window," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 12, pp. 1667-1671, Dec. 2002.

[17] E. Parzen, "On Estimation of a Probability Density Function and Mode," Annals of Math. Statistics, vol. 33, pp. 1065-1076, 1962.

[18] R.T. Marler and J.S. Arora, Survey of multi-objective optimization methods for engineering, Struct Multidisc Optim 26, 369–395 (2004)