CO12.2 - Développement de méthodes de sélection de variables incluant un terme de pénalisation en classification supervisée

N. Ngo, R. Giorgi

Développement de méthodes de sélection de variables incluant un terme de pénalisation en classification supervisée.

Mots-clés : Sélection de variables ; Pénalisation ; Classification ; Simulation

Introduction [1032 caractères]

Dans le contexte de la classification supervisée il est parfois nécessaire d'utiliser des méthodes de sélection de variables afin de n'inclure que les variables les plus pertinentes dans l'analyse, i.e. de ne conserver que les variables nous permettant de classer correctement une observation. Cette étape de sélection de variable est cruciale pour obtenir des modèles généralisables et des prédictions correctes. Ainsi, lors du développement de ces méthodes on cherche à atteindre au moins deux objectifs : l'élimination des variables non pertinentes et la minimisation de l'erreur de classification. Dans le cadre de nos travaux, nous avons été amenés à construire des méthodes de sélection de variables nous permettant d'identifier des variables pertinentes pour la classification. Cependant les méthodes retenaient à tort un certain nombre de variables. L'objectif de notre travail était de proposer une méthode de pénalisation afin de réduire la taille du sous-ensemble de variables tout en conservant un taux d'erreur faible.

Méthodes [703 caractères]

Notre pénalisation est basée sur le calcul de la gamma-metric auquel s'ajoute un terme dépendant de la dimension du sous-ensemble de variables. Pour vérifier l'apport de ce terme de pénalisation sur la méthode de sélection de variable, nous avons conduit une étude par simulation dans laquelle nous avons généré des variables informatives et non-informatives. Nous avons utilisé les méthodes de sélection de variables avec ou sans pénalisation sur un échantillon d'apprentissage et comparé i) le nombre et la nature des variables sélectionnées ainsi que ii) le taux d'observations correctement classées, sur un échantillon de validation, avec les modèles SVM construit à partir des variables sélectionnées.

Résultats [520 caractères]

Les résultats de la simulations sont donnés dans la Table 1. En terme de nombre de variables sélectionnées, les méthodes utilisant la pénalisation sélectionnaient moins de variables dans tous les cas. Parmi les variables sélectionnées, les méthodes avec pénalisation retenaient un peu moins souvent les variables dites informatives en moyenne mais, en terme de classification, le taux d'observations correctement classées augmentait avec les méthodes utilisant la pénalisation par rapport aux méthodes sans pénalisation.

Méthode	Nombre moyen de variables sélectionnées	Nombre moyen de variables informatives sélectionnées	Taux d'observations correctement classées (Apprentissage)	Taux d'observations correctement classées (Validation)
Sans sélection de variables	50	3	96,34	78,72
Backward sans pénalisation	39,94	2,88	95,66	79,86
Best first sans pénalisation	5,60	2,26	91,68	87,10
Forward sans pénalisation	5,60	2,26	92,18	87,12
Hill climbing sans pénalisation	29,96	2,72	95,60	80,96
Backward avec pénalisation	21,54	2,44	93,14	83,66
Best first avec pénalisation	2,06	2,06	90,28	87,94
Forward avec pénalisation	2,06	2,06	90,32	87,86
Hill climbing avec pénalisation	2,06	2,06	90,32	87,78

Table 1 : Nombre moyens de variables sélectionnées parmi les 50 variables initiales (47 non-informatives et 3 informatives) par les méthodes listées sur les 50 répétitions de la simulation. Le taux d'observations correctement classées est donnée en pourcentage. Chaque jeu de données est composée de 100 observations avec en moyenne 51,38% des observations classées en 1 sur l'ensemble des données d'apprentissage et 50,42% classées en 1 sur les données de validation.

Conclusion [217]

Les résultats obtenus sur cette simulation nous montrent que la pénalisation proposée nous a permis de conserver un taux d'erreur plutôt faible tout en supprimant des variables non-informatives quel que soit la méthode.

Total de caractères : 2472