Prétraitement de grands ensembles de données pour la fouille visuelle

Edwige Fangseu Badjio, François Poulet

ESIEA Pôle ECD,
Parc Universitaire de Laval-Changé,
38, Rue des Docteurs Calmette et Guérin,
53000 Laval France
fangseubadjio@esiea-ouest.fr
poulet@esiea-ouest.fr

Résumé. Nous présentons une nouvelle approche pour le traitement des ensembles de données de très grande taille en fouille visuelle de données. Les limites de l'approche visuelle concernant le nombre d'individus et le nombre de dimensions sont connues de tous. Pour pouvoir traiter des ensembles de données de grande taille, une solution possible est d'effectuer un prétraitement de l'ensemble de données avant d'appliquer l'algorithme interactif de fouille visuelle. Pour ce faire, nous utilisons la théorie du consensus (avec une affectation visuelle des poids). Nous évaluons les performances de notre nouvelle approche sur des ensembles de données de l'UCI et du Kent Ridge Bio Medical Dataset Repository.

1 Introduction

Nous nous intéressons au problème de prétraitement de grands ensembles de données. Notre but est de réduire les informations contenues dans les ensembles de données volumineux aux informations les plus significatives. Il existe des techniques expérimentalement validées pour ce faire. D'un point de vue applicatif, un problème majeur se pose quant au choix d'une de ses méthodes. Une solution qui constitue notre contribution dans ce travail serait d'utiliser une combinaison de techniques ou de stratégies. A cet effet, nous nous appuyons sur la théorie du consensus. L'utilisation de cette combinaison de stratégies ou d'expertises peut être justifiée par l'un des faits suivants :

- il n'est pas possible de déterminer a priori quelle méthode de sélection de sous-ensemble d'attributs est meilleure que toutes les autres (en tenant compte des différences entre le temps d'exécution et la complexité),
- un sous-ensemble optimal d'attributs n'est pas nécessairement unique,
- la décision d'un comité d'experts est généralement meilleure que la décision d'un seul expert.

Les résultats obtenus après des expérimentations permettent de conclure que l'approche proposée réduit de façon significative l'ensemble de données à traiter et permet de les traiter interactivement. Cette contribution commence par un état de l'art et la problématique du

sujet abordé, puis, l'algorithme de sélection d'attributs est présenté. Enfin, nous procédons à des expérimentations avant la conclusion.

2 Etat de l'art et problématique

Nous essayons de résoudre le problème suivant : comment sélectionner des attributs d'un ensemble de données pourvu de plusieurs attributs et rejeter les autres sans nuire à la qualité de l'algorithme de fouille visuelle utilisé ensuite ? Ceci tout en sachant que :

- la visualisation de plus de deux dizaines d'attributs rend souvent inutilisable la fouille visuelle de données,
- un sous-ensemble optimal d'attributs n'est pas nécessairement unique,
- il n'est pas possible de déterminer a priori quelle méthode de sélection de sous-ensemble d'attributs est meilleure que toute les autres,
- la décision d'un comité d'experts est généralement meilleure que la décision d'un seul expert.

Des techniques performantes (John et al., 1994), (Kira et Rendell, 1992), etc. de sélection de sous-ensembles d'attributs ont été développées mais, il n'existe pas une méthode qui soit meilleure que toutes les autres dans tous les cas.

Nous avons défini un nouvel algorithme de sélection d'attributs qui combine des décisions pondérées de plusieurs experts. Plus précisément, étant donné deux ou plusieurs méthodes de sélection de sous-ensembles pertinents d'attributs dans un ensemble de données, la question est de savoir comment l'on peut utiliser ces différentes méthodes pour fournir un résultat efficace. Afin de répondre à cette question, nous nous sommes appuyés sur la théorie du consensus qui peut être définie comme un procédé de prise de décision qui utilise entièrement les ressources d'un groupe. La théorie du consensus trouve l'une de ses justifications dans le fait qu'une décision prise par un groupe d'experts est meilleure en terme d'erreur quadratique moyenne que la décision d'un seul expert. Une telle démarche possède de nombreux avantages : statistiquement parlant, la consultation de plusieurs expertises lors de la résolution d'un problème est une façon subjective d'accroître la taille de l'échantillon dans une expérience, un ensemble d'experts permet d'obtenir plus d'information qu'un seul expert (Clemen et Winkler, 1999).

La section suivante présente l'algorithme de sélection d'attributs proposé.

3 Algorithme de sélection d'attributs basé sur la théorie du consensus (CTBFS)

L'algorithme proposé « Consensus Theory Based Feature Selection » (CTBFS) reçoit en entrée des sous-ensembles d'attributs.

Le domaine considéré est constitué d'une valeur limite du nombre d'attributs susceptibles d'être correctement visualisés et traités da façon interactive (C_{cmd}), un ensemble m d'experts (algorithmes de sélection d'attributs) $E = \{E_1, \ldots, E_m\}$, chaque expert E_i dispose d'un sous-ensemble de L experts (qui représentent les différents critères ou paramètres importants des algorithmes de sélection d'attributs) $E_i = \{e_1, \ldots, e_l\}$. L'utilisation de ce sous-ensemble d'experts (E_i) peut être justifié par le fait que dans un algorithme de sélection d'attributs significatifs, il n'y a aucun critère qui permet d'obtenir de meilleurs résultats que tous les

autres. Chaque critère possède des attributs de qualité spécifiques. Il est nécessaire de prendre en considération tous les différents attributs de qualité.

Nous avons aussi un sous-ensemble d'attributs $DS = \{D_1, \ldots, D_L\}$, où $D_i = \{d_1, \ldots, d_K\}$ et K est variable. Les sous-ensembles d'attributs sont disponibles selon les paires expert/attributs (e_j, D_j), où $e_j \in E_i$ et $D_j \in DS$.

Chaque attribut sélectionné par un sous expert e_j a une fréquence d'apparition $freq = 1/nb$ dans la décision finale, où nb est le nombre d'attributs sélectionnés par le sous expert. Nous définissons un critère de préférence d'un attribut (règle de consensus) comme étant le produit des fréquences d'apparition de l'attribut dans les sous-ensembles d'attributs des experts. Nous utilisons la formule ci-dessous pour le calcul de la préférence d'un attribut d.

$$pref(X = d) = \prod_{i=1}^{N} P(X = d \mid D_i = b_i)^{w_i}$$

où : $P(X = d \mid D_i = b_i)$ est la probabilité a posteriori que l'attribut testé appartienne au sous-ensemble d'attributs à sélectionner lorsque la décision du $m^{ième}$ expert est b_i, w_i est le poids assigné à l'expert.

Des représentations graphiques de l'ensemble de données constituées uniquement des attributs sélectionnés sont utilisées pour la définition interactive de poids à affecter aux différents experts qui interviennent dans la sélection d'attributs. Il s'agit ici d'un problème d'optimisation de l'affectation de poids aux experts. Dans un problème d'optimisation, il y a un espace des solutions et une fonction d'évaluation afin d'accéder à la qualité de la solution.

Les poids affectés aux experts doivent à cet effet être proportionnels à leurs décisions. La méthode d'affectation de poids que nous proposons a pour fondements théoriques un principe de la théorie de Gestalt (une vue d'ensemble est meilleure que la somme des parties) et des propriétés pré-attentives de la vision humaine. En ce qui concerne le principe de Gestalt, en visualisant l'ensemble d'éléments intervenant dans une décision, un processus cognitif se met en place. L'application du principe de Gestalt se résume en une représentation graphique multi vues à base de coordonnées parallèles (Inselberg, 1985) des attributs sélectionnés. Les coordonnées parallèles permettent de représenter en 2D des données multidimensionnelles sans perte d'information. Chaque vue représente le point de vue de chaque expert (une des données d'entrée de CTBFS), comme l'indique la figure 1.

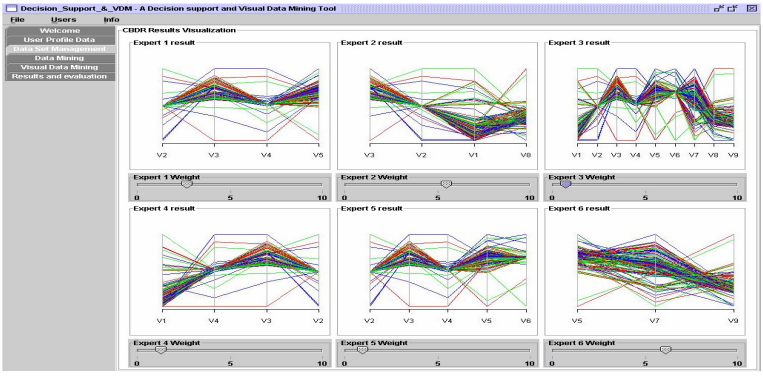

FIG. 1 – *Outil d'affectation visuelle de poids aux experts intervenant dans CTBFS.*

Six experts de type filtre ont servi à la sélection des attributs visualisés dans la figure 1. L'expert 1 représente le critère de sélection consistance, l'expert 2 représente l'entropie de Shannon, l'expert 3 quant à lui utilise la distance comme fonction d'évaluation. La fonction d'évaluation pour l'expert 4 est le gain d'information, le coefficient de Gini pour l'expert 5 et le coefficient de Cramer pour l'expert 6.

Il est à noter que les outils usuels d'affectation de poids sont des « boîtes noires ». L'avantage principal de l'approche ainsi proposée tient au fait que l'utilisateur est impliqué et participe dans le processus de prise de décision. Il existe un ensemble de propriétés visuelles qui sont traitées de manière pré attentive très rapidement, avec précision et sans effort particulier, ce qui permet aux utilisateurs d'affecter des poids convenables aux différents experts.

De plus, les techniques de visualisation permettent d'améliorer la résolution de problèmes. La visualisation permet de découvrir plus aisément des motifs dans les données, de réduire l'espace de recherche d'information par rapport aux méthodes automatiques, de procéder à des opérations perceptuelles d'inférence et d'augmenter la mémoire et les ressources de traitement de l'utilisateur (Card et al., 1999).

4 Expérimentations

Pour les besoins d'expérimentation de la technique proposée qui a été développée sous Windows avec Java et le langage R, nous utilisons un pentium IV, 1.7 GHz. Les ensembles de données que nous utilisons ont été référencées par (Blake et Merz, 1998) et (Jinyan et Huiqing, 2002). Pour les besoins de ces expérimentations, les poids affectés aux différents experts ont pour valeur 1.

Le domaine considéré dans le cadre de cette première expérimentation est constitué d'un ensemble M constitué de 3 experts de type filtre et de 3 experts de type enveloppe E = *{consistence, entropie de Shannon, distance, (LDA, QDA, Kppv)* (Ripley, 1996)*}*, le nombre d'attributs susceptibles d'être traités convenablement est $C_{cmd} = 20$.

Les résultats de l'algorithme proposé (CTBFS) sont comparés à ceux de Las Vegas Filter (Liu et Setiono, 1996), un algorithme de type filtre et StepClass du package KlaR (langage de programmation R), un algorithme de type enveloppe. A cet effet, nous évaluons les performances des ensembles de données pourvus des attributs sélectionnés par ces trois méthodes avec l'algorithme des k plus proches voisins kppv (implémentation de WEKA (Witten et Eibe, 2005)). Nous avons fixé le paramètre K de l'algorithme des kppv à 1.

Les ensembles de données à traiter dans le cadre de cette première expérimentation sont pourvus de nombreux attributs (colonne 2 du tableau 1) qu'il serait impossible de visualiser en une seule fois à l'écran quelque soit la méthode de représentation graphique.

Les résultats exposés dans le tableau 1 permettent d'observer que l'algorithme CTBFS permet de réduire considérablement le nombre d'attributs des ensembles de données comme le montre les résultats de la colonne 3. La colonne 5 de ce tableau quant à elle fait observer que la précision de l'algorithme de kppv est améliorée pour 4 ensembles de données sur 7. Pour les trois autres ensembles de données, on assiste certes à une perte de précision avec un écart maximal de 16.97% avec un minimum de précision de 68.87% mais l'ensemble de données final peut être visualisé et traité de manière interactive, ce qui n'est pas le cas des ensembles de données initiaux comme nous l'avons souligné.

Nom	NbAt_Initial	NbAt_CTBFS	Précision_init	Précision_CTBFS
Lung-Cancer	57	**4**	37.5%	**75%**
Promoter	59	**9**	**85.84%**	68.87%
Sonar	60	**8**	**86.54%**	71.15%
Arrhythmia	280	**4**	53.44%	**59.96%**
Isolet	618	**14**	**85.57%**	70.24%
ColonTumor	2000	**19**	77.42%	**79.03%**
CentralNervSyst	7129	**20**	56.67%	**60%**

TAB. 1 – *Comparaison du nombre d'attributs et de la précision obtenus avec l'algorithme des kppv avant et après la sélection d'attributs par l'algorithme CTBFS*

Nom	NbAttr CTBFS	NbAttr LVF	NbAttr Stepclass	CTBFS précision	LVF précision	Stepclass précision
Lung-Cancer	**4**	17	**4**	**75%**	62.5%	71.87%
Promoter	**9**	16	59	68.87%	80.19%	**85.85%**
Sonar	**8**	18	**4**	***71.15%***	82.21%	***71.63%***
Arrhythmia	**4**	109	**4**	**59.96%**	54.65%	**60.84%**
Isolet	**14**	268	**8**	***70.24%***	83%	***57.98%***
ColonTumor	**19**	918	**5**	**79.03%**	77.42%	**79.03%**
CentralNervSyst	**20**	3431	**8**	***60%***	***58.33%***	**71.67%**

TAB. 2 – *Comparaison du nombre d'attributs et de la précision obtenus avec l'algorithme des kppv avant et après la sélection d'attributs par les algorithmes CTBFS, LVF et Stepclass.*

On observe à travers la colonne 3 du tableau 2 que la méthode LVF permet de sélectionner un nombre très important d'attributs, qu'il serait impossible de visualiser (par exemple pour les ensembles de données Arrhythmia, Isolet, ColonTumor et CentralNervSyst). Par rapport à la méthode proposée, la précision obtenue pour ces ensembles de données est équivalente voire supérieure par exemple pour l'ensemble de données Isolet, sachant que l'algorithme CTBFS renvoie au maximum 20 attributs. En ce qui concerne l'algorithme Stepclass, l'ensemble de données Promoter possède aussi un nombre important d'attributs.

En terme de précision, en dehors de l'ensemble de données Promoter pour lequel CTBFS a une précision inférieure à celle de Stepclass et de LVF, la précision obtenue pour les autres ensembles de données avec l'algorithme proposé est au moins égale suivant les cas à celle de LVF ou à celle de Stepclass mais avec un nombre d'attributs qui convient à la fouille visuelle de données.

5 Conclusion

Nous avons présenté un algorithme basé sur la théorie du consensus et l'affectation visuelle de poids pour la sélection d'attributs significatifs en FVD. En effet, lorsque le nombre d'attributs et/ou le nombre d'observations d'un ensemble de données est important, il s'avère impossible ou alors pénible de représenter graphiquement l'ensemble de données et d'observer des corrélations dans cet ensemble de données.

La technique présentée permet de définir un nombre maximum d'attributs à sélectionner dans l'ensemble de données à traiter, rendant possible la visualisation de ces données. La

première nécessité pour nous est de pouvoir représenter visuellement l'ensemble de données à traiter. Les expérimentations effectuées à cet effet ont été concluantes.

Nos investigations en ce qui concerne la réduction des observations dans un ensemble de données consistent à agréger l'information contenue dans cet ensemble de données. A la suite de la sélection des attributs, l'utilisation des algorithmes de clustering nous permet de réduire le nombre d'individus des ensembles de données de 50 à 75% avec un maximum de 200 clusters par application de l'algorithme K-Means.

Comme perspectives à ces travaux, nous comptons étendre l'application de la théorie du consensus au choix de la meilleure méthode de classification supervisée ou non supervisée pour un ensemble de données à traiter.

Références

Blake, C. et C. Merz (1998). UCI Repository of machine learning databases. Irvine, University of California, Department of Information and Computer Science, from www.ics.uci.edu/~ mlearn/MLRepository.html.

Card, S.K., J. D. Mackinlay, et B. Shneiderman (1999). Information Visualization: Using Vision to Think. Academic Press.

Clemen, R. T. et R.L. Winkler (1999). Combining probability distributions from experts in risk analysis. Risk Analysis, 19(2):187–203, 1999.

A. Inselberg (1985). The plane with parallel coordinates. The Visual Computer, 1, pages 69--91.

Jinyan, L. et L. Huiqing (2002). Kent Ridge Bio-medical Data Set Repository. http://sdmc.lit.org.sg/GEDatasets.

John, G. H., R. Kohavi et K. Pfleger (1994). Irrelevant features and the subset selection problem, in International Conference on Machine Learning, pp. 121-129.

Kira, K. et L. A. Rendell (1992). A practical approach to feature selection. In Proc. of the Tenth Int'l Conf. on Machine Learning, pages 500–512.

Liu, H. et R. Setiono (1996). A probabilistic approach to feature selection: a filter solution. In Proc, The 13th International Conference on Machine Learning, pages 319-327.

Ripley, B. D. (1996). Pattern Recognition and Neural Networks. Cambridge Univ.Press.

Witten I.H. et F. Eibe (2005). Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

Summary

Visualization methods do not scale well with high number of features. We present an approach using a consensus theory based feature selection (CTBFS) algorithm, clustering for sampling and visualization for weight assignment in order to aggregate multivariate and multidimensional datasets.