Classification des données catégorielles via la maximisation spectrale de la modularité

Lazhar Labiod*, Younès Bennani**

* LIPADE, University of Paris Descartes,
45 rue des Saints Pères 75006 Paris, France.
email: Prénom.Nom@parisdescartes.fr,
** LIPN UMR 7030, Université Paris 13
99, avenue Jean-Baptiste Clément, 93430 Villetaneuse
Prénom.Nom@lipn.univ-paris13.fr,

Résumé. Ce papier présente un algorithme spectrale pour maximiser le critère de la modularité étendu à la classification des données catégorielles. Il met en evidence la connexion formelle entre la maximisation de la modularité et la classification spectrale, il présente en particulier le problème de maximisation de la modularité sous forme d'un problème algèbrique de maximisation de la trace. Nous développons ensuite un algorithme efficace pour trouver la partition optimale maximisant le critère de modularité. Les résultats expérimentaux montrent l'efficacité de notre approche.

1 Introduction

La classification automatique est une méthode d'apprentissage non supervisé permettant le partitionnement d'un ensemble d'observations en classes. les méthodes de classification automatique conduisent à une partition de la population initiale en groupes disjoints, tels que, selon un critère choisi a priori, deux individus d'un même groupe aient entre eux un maximum d'affinité et deux individus de deux groupes différents aient entre eux un minimum d'affinité. La classification automatique a été largement étudiée en apprentissage automatique, en bases de données et en statistique de divers points de vue.

La mesure de modularité a été utilisée récemment pour la classification de graphes [Agarwal and Kempe, 2008], [Newman and Girvan, 2004] et [White and Smyth, 2005]. Dans ce papier, nous montrons que le critère de modularité peut être formellement etendu pour la classification des données catégorielles. Nous développons ensuite une procédure spectrale efficace pour trouver la partition optimale maximisant le critère de modularité. Les résultats expérimentaux montrent l'efficacité de notre approche. La première contribution de ce papier est l'introduction d'une mesure de modularité étendue pour la classification des données catégorielles. La deuxième contribution est la présentation du problème de maximisation de la mesure de modularité étendue sous la forme d'un problème de maximisation de trace.
Le reste du papier est organisé comme suit: la section 2 introduit quelques notations et définitions. La section 3 présente la mesure de modularité étendue. Des discussions sur la connexion

spectrale du problème de maximization de la modularité et la procédure d'optimisation proposée sont décrites à la section 4. La section 5 montre nos résultats expérimentaux et enfin, la section 6 présente des conclusions et certains travaux futurs.

2 Définitions et notations

Soit I un ensemble de données avec N objets $\{O_1,O_2,...,O_N\}$ décrit par l'ensemble V de M attributs (ou variables catégorielles) $\{V^1,V^2,..,V^m,...,V^M\}$ chacun ayant $p_1,..,p_m,...,p_M$ catégories, respectivement, et soit $P = \sum_{m=1}^{M} p_m$, désigne le nombre total de catégories de toutes les variables. Chaque variable catégorielle peut être décomposée en une collection de variables indicatrices. Pour chaque variable V^m, considérons les p_m valeurs qui correspondent naturellement aux nombres de 1 à p_m et $V_1^m,V_2^m,...,V_{p_m}^m$ sont des variables binaires telles que, pour chaque j, $1 \leq j \leq p_m$, $V_j^m = 1$ si et seulement si V^m prend la $jieme$ valeur. Ainsi la matrice de données peut être exprimée comme une collection de M matrices K^m, $(m = 1,..,M)$ de terme général k_{ij}^m tel que : $k_{ij}^m = 1$ si l'objet i possède la catégorie j de V^m et 0 sinon. La matrcie disjonctive K de dimensions $N \times P$ s'écrit; $K = (K^1|K^2|...|K^m|...|K^M)$.

3 Extension de la mesure de modularité à la classification des données catégorielles

Cette section explique comment adapter la mesure de modularité à la classification des données catégorielles.

3.1 Graphe et modularité

La modularité est une mesure récemment utilisée pour mesurer la qualité d'une classification de graphes, elle a immédiatement reçu une attention considérable comme en témoignent les articles [Newman and Girvan, 2004], [Agarwal and Kempe, 2008]. La maximisation de la mesure de modularité peut être exprimée sous la forme d'un problème de programmation linéaire en nombres entiers. Étant donné le graphe $G = (V,E)$, soit A une matrice binaire et symétrique où chaque entrée $a_{ii'} = 1$ s'il existe une arête entre les noeuds i et i', s'il n'y a pas de lien entre les noeuds i et i', $a_{ii'}$ est égal à zéro. A est une matrice contenant toutes les informations sur le graphe G, est souvent appelée matrice d'adjacence. Trouver une partition de l'ensemble des noeuds V en sous-ensembles homogènes conduit à la résolution du programme linéaire en variables bivalentes suivant :

$$\max_X Q(A,X) \text{ où } Q(A,X) = \frac{1}{2|E|} \sum_{i,i'=1}^{N} (a_{ii'} - \frac{a_{i.}a_{i'.}}{2|E|})x_{ii'} \quad (1)$$

où X est une matrice relation d'équivalence, avec, $2|E| = \sum_{i,i'} a_{ii'} = a_{..}$ est le nombre total d'arêtes (liens) et $a_{i.} = \sum_{i'} a_{ii'}$ le degré de l'objet i.
La modularité évalue la densité des arêtes dans les classes de façon relative à la densité attendue en cas d'indépendance entre les extrémités des arrêtes. Elle prend ses valeurs entre -1 et

1 et des valeurs positives, quand les classes ont plus d'arêtes observées que dans le cas d'indépendance des extrémités des arêtes. Ce critère vaut 0 dans les deux cas d'une partition triviale; le cas d'une seule classe et le cas où chaque noeud est isolé dans une classe.

3.2 Extension par intégration a priori

L'intégration a priori consiste en une combinaison directe de graphes obtenus à partir de toutes les variables dans un seul ensemble de données (graphe) avant d'appliquer l'algorithme d'apprentissage. Prenons la matrice $S = KK^t$, (où chaque entrée $s_{ii'} = \sum_{m=1}^{M} s_{ii'}^m$), qui peut être considérée comme une matrice de poids associée au graphe $G = (V,E)$, où chaque arête $e_{ii'}$ a le poids $s_{ii'}$. Par analogie avec la mesure de modularité classique, nous définissons l'extension $Q_1(S,X)$ comme suit :

$$Q_1(S,X) = \frac{1}{2|E|} \sum_{i,i'=1}^{N} (s_{ii'} - \frac{s_{i.}s_{i'.}}{2|E|})x_{ii'} = \frac{1}{2|E|} Tr[(S-\delta)X] \tag{2}$$

où $2|E| = \sum_{i,i'} s_{ii'} = s_{..}$ est le poids total et $s_{i.} = \sum_{i'} s_{ii'}$ le degré de l'objet i, et $\delta_{ii'} = \frac{s_{i.}s_{i'.}}{2|E|}$

4 Maximisation spectrale de la modularité normalisée

Le critère de modularité n'est pas pondéré par les cardinalités des classes. Ce qui signifie qu'une classe peut se faire petite quand elle est touchée par des valeurs aberrantes. Ainsi, nous définissons le nouveau critère de modularité normalisée comme suit :

$$\tilde{\mathcal{Q}}_1(S,X) = \frac{1}{2|E|} Tr[(S-\delta)Z(Z^TZ)^{-1}Z] \tag{3}$$

Où $X = ZZ^T$, Z est une matrice binaire de partition de l'ensemble I en $\mathcal{K}$ classes. La matrice Z^TZ est diagonale, chaque élément diagonal correspond à la cardinalité d'une classe de la partition Z.

4.1 Connexion spectrale

Dans cette section nous donnerons une interprétation spectrale au problème de maximisation du critère de modularité en exploitant quelques propriétés algèbriques de la relation d'équivalence X. Nous appliquons en suite une relaxation spectrale du problème pour maximiser le critère de modularité normalisée.
D'un côté, il est bien connu que la plus grande valeur propre de la matrice pondérée $\tilde{S} = D^{-\frac{1}{2}}SD^{-\frac{1}{2}}$ (où $D = diag(Se)$ et e est un vecteur de dimension appropriée dont toutes ses valeurs valent 1.) et son vecteur associé sont $\lambda_0 = 1$, $U_0 = \frac{D^{-\frac{1}{2}}e}{S_{..}}$ [Ding and Simon, 2011], [Bach and Jordan], où $S_{..} = \sqrt{e^tSe}$.
Applicant maintenant la décomposition spectrale de la matrice $\tilde{S}$ au lieu de S, on obtient ainsi $S = D^{\frac{1}{2}} \sum_{k=0} U_k\lambda_kU_k^tD^{\frac{1}{2}}$. Soustraire le vecteur propre trivial correspondant à la plus grande valeur propre $\lambda_0 = 1$ donne $S = \frac{Dee^tD}{e^tDe} + D^{\frac{1}{2}} \sum_{k=1} U_k\lambda_kU_k^tD^{\frac{1}{2}}$. Considérons maintenant

que les $\mathcal{K}-1$ vecteurs propres principaux de $S-\frac{Dee^tD}{e^tDe} = D^{\frac{1}{2}}\sum_{k=1}U_k\lambda_kU_k^tD^{\frac{1}{2}}$. Cette matrice multipliée par la constante $\frac{1}{e^tDe}$ (qui n'a pas d'impact sur la position du maximum du critère de modularité) est exactement la matrice $(S-\delta)$ présente dans la partie donnée dans le critère de modularité (voir equation (3)).
De l'autre côté le probléme de maximisation de la modularité étendue peut être formaliser sous la forme d'un probléme algèbrique de maximisation de trace sous la contrainte d'orthogonalité de la matrice de partition. $\max_{\tilde{Z}^t\tilde{Z}=I_K} Tr[\tilde{Z}^t(S-\delta)\tilde{Z}]$, avec $\tilde{Z}=Z(Z^tZ)^{-1/2}$, on peut facilement vérifier que $\tilde{Z}$ est une matrice orthogonale, c.a.d $\tilde{Z}^t\tilde{Z}=I_{\mathcal{K}}$, où $I_{\mathcal{K}}$ la matrice identité d'ordre $\mathcal{K}$.

La matrice $S-\delta$ utilisée dans le critère de modularité est exprimé en termes des $\mathcal{K}-1$ plus grands vecteurs propores de la matrice de similarité $\tilde{S}$. Après résolution de la décomposition spectrale de $\tilde{S}$, on obtient les $(\mathcal{K}-1)$ vecteurs propres associés aus plus grandes valeurs propres. Ainsi, on peur définir la matrice $U=[U_1,...,U_{\mathcal{K}-1}]$ de dimensions $N\times(\mathcal{K}-1)$, avec U_k le $kieme$ vecteur propre des $\mathcal{K}-1$ vecteurs retenus. On normalise chaque colonne de cette matrice telle que $\tilde{U}_k=\frac{D^{\frac{1}{2}}U_k}{||D^{\frac{1}{2}}U_k||}$.
Le programme ci-dessus est l'équivalent spectral du problème de maximisation du critère de modularité normalisée. L'algorithme proposé appelé SpectCat commence par le calcul des $(\mathcal{K}-1)$ premiers vecteurs propres en ignorant le vecteur trivial. L'algorithme intégre les données d'entrée dans un espace Euclidien par une décomposition spectrale de la matrice de similarité, puis on applique un algorithme de clustering géométrique sur $\tilde{U}$. Les principales étapes de l'algorithme spectral utilisé sont décrites ci-après (**Algorithme1**).

Algorithm 1 SpectCAT

Input: Matrice de similarité S, nombre de classes $\mathcal{K}$

Output: Matrice de partitions Z

2. Définir D comme étant la matrice diagonale $D=diag(S\mathbb{1})$

3. Trouver U $(\mathcal{K}-1)$ vecteurs propres de $\tilde{S}=D^{-\frac{1}{2}}SD^{-\frac{1}{2}}$

4. Définir la matrice $\tilde{U}$ à partir de U par $\tilde{U}_k=\frac{D^{\frac{1}{2}}U_k}{||D^{\frac{1}{2}}U_k||}$, $\forall k=1,...,\mathcal{K}-1$

5. Partitionner les lignes de $\tilde{U}$ en $\mathcal{K}$ classes en utilisant par exemple les $k$means

6. Affecter l'object i à la classe Z_k si et seulement si la ligne correspondante $\tilde{U}_i$ de la matrice U a été affectée à la classe Z_k

5 Expérimentation et validation

Une étude de performance a été réalisée pour évaluer notre méthode. Dans cette section, nous décrivons les expériences et les résultats. Nous avons testé notre algorithme sur des données réelles obtenues à partir du référentiel UCI (Machine Learning Repository) et comparer ses performances avec d'autres algorithmes de clustering en utilisant la pureté pour mesurer la qualité d'un résultat de clustering. Nous effectuons des comparaisons sur sept bases de données UCI (Asuncion and Newman, 1997. Soybean small, Mushroom, Congressional votes,

Zoo, hayes-roth, Balance scale, Car evaluation et Audiology. La description synthétique des bases de données utilisées est donnée dans le tableau 1:

TAB. 1 – – *Description des bases de données*

Bases de données	# d'objects	# d'attributes	# de classes
Soybean small	47	21	4
Mushroom	8124	22	2
Congressional votes	435	16	2
Zoo	101	16	7
Hayes-roth	132	4	3
Balance Scale	625	4	3
Car evaluation	1728	6	4
Audiology	200	69	24

5.1 Analyse des résultats

Nous avons étudié la clustering trouvé par quatre algorithmes, notre algorithme SpectCat, k-modes standard, l'algorithme K-representative et l'algorithme Wk-modes. Nous avons observé que la plupart des algorithmes de clustering nécessitent le nombre de clusters comme paramètre d'entrée, alors, dans nos expériences, nous avons fait varier pour chaque base de données le nombre de clusters, allant du nombre réel de classes pour chaque base de données à 10. Pour chaque nombre de clusters, la pureté du clustering des différents algorithmes a été comparée.

Comme la méthode proposée est une approche spectrale adapté aux données catégorielles, nous avons comparé les performances de l'algorithme proposé avec d'autres algorithmes de classification de données catégorielles. Du tableau 2, il est clair que la performance de la méthode proposée qui repose sur le principe de la classification spectrale donne des résultats meilleurs ou semblables à ceux d'autres approches. Cela signifie que l'approche proposée améliore la pureté du clustering. les disparités des performances de SpectCat par rapport aux différents algorithmes peut s'expliquer par la structure interne aux données, à savoir le nombre de modalités par variable, l'effectif de chaque modalité et la cavité de la matrice de données.

TAB. 2 – – *mesure de Pureté pour K-modes, K-representatives, weighted k-modes et SpectCat*

Bases de données	K-Modes	K-Representatives	WK-Modes	SpectCat
Soybean small	66	89	89	100
Mushroom	59	61	61	61
Congressional votes	62	87	88	88
Zoo	88	89	90	90
Hayes-roth	41	42	42	54
Balance Scale	50	52	52	56
Car evaluation	70	70	71	70
Audiology	62	61	62	61

Pour la base de données Balance scale, la méthode proposée est efficace pour la plupart des valeurs de K, la méthode proposée donne de moins bons résultats que l'algorithme K-representative dans les cas où $K = 7$. Pour les autres cas, SpectCat produit des clusters de haute pureté comme indiqué dans la figure 2 (à gauche). Pour la base de données Zoo, SpectCat donne des résultats comparables ou moins bons par rapport aux autres algorithmes, les résultats sont donnés dans la figure 2 à droite.

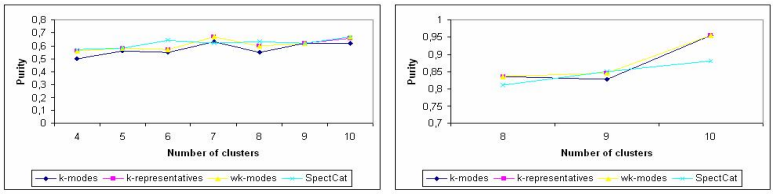

FIG. 1 – – *Mesure de Pureté vs. différent nombre de clusters (- gauche: Balance scale data set - droite: Zoo data set)*

6 Conclusions et perspectives

Dans ce papier, nous avons étudié la maximisation spectrale de la modularité pour la classification des données catégorielles. Nous avons proposé une approximation spectrale de la matrice de modularité et un équivalent spectral du problème de maximisation de la modularité. Une procédure efficace pour l'optimisation est présentée. les résultats expérimentaux obtenus en utilisant différentes bases de données réelles montrent que notre méthode fonctionne efficacement. Notre méthode peut être facilement étendue à un cadre spectral plus général permettant de combiner de multiples ensembles de données hétérogènes.

Références

G. Agarwal and D Kempe. Modularity-maximizing graph communities via mathematical programming. *Journal of Machine Learning Research*, B 66:33:409–418, 2008.

Francis R. Bach and Michael I. Jordan. Learning spectral clustering, with application to speech separation. *The European Physical Journal*.

Xiaofeng Zha Hongyuan Ding, Chris H.Q. He and Horst D Simon. Self-aggregation in scaled principal component space. Technical report, Ernest Orlando Lawrence Berkeley National Laboratory, Berkeley, CA (US), 2011.

M. Newman and M Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69, 2004.

S. White and P. Smyth. A spectral clustering approach to finding communities in graphs. In *SDM*, pages 76–84, 2005.

Summary

In this paper we propose a spectral based clustering algorithm to maximize an extended Modularity measure for categorical data. The maximization of the extended modularity is shown as a trace maximization problem. A spectral based algorithm is then presented to search for the partitions maximizing the extended Modularity criterion.