Sélection de mesures de similarité pour la classification de données catégorielles

Guilherme Alves, Miguel Couceiro, Amedeo Napoli

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy
{guilherme.alves-da-silva,miguel.couceiro,amedeo.napoli}@loria.fr

Résumé. Le partitionnement de données est une opération très utilisée dans l'exploration et l'analyse de données, en particulier pour traiter des tableaux de données qui comprennent des attributs catégoriels. Une telle opération repose sur des mesures de similarité, qui sont proposées en nombre dans la littérature. Cependant, le choix d'une mesure est complexe et dépend du contexte et des données en cours d'étude. Dans cet article, nous cherchons à caractériser de façon automatique la "meilleure" mesure de similarité pour partitionner un jeu de données particulier. Nous présentons les bases de notre approche et une étude empirique qui porte sur des données catégorielles ainsi qu'une évaluation de cette approche.

1 Introduction

Beaucoup d'opérations du monde réel engendrent de très grandes masses de données, relatives par exemple à la consommation, la santé, le temps, l'espace, les réseaux sociaux ... Ces données sont généralement hétérogènes, signifiant entre autres qu'elles sont décrites par des attributs de types différents, numériques ou catégoriels (Šulc et Řezanková, 2019).

Diverses méthodes de fouille de données permettent de tenir compte de cette hétérogénéité. Ainsi, les méthodes de partitionnement [1] sont souvent employées pour analyser de telles données, découvrir des classes ou des profils, ou résumer et organiser des données (Barioni et al., 2014). Un processus de partitionnement repose sur des mesures de similarité, entre les objets et entre les classes, qui jouent un rôle fondamental. Il existe de nombreuses mesures pour les données catégorielles (Boriah et al., 2008), sans qu'il n'en existe une qui soit universelle et meilleure qu'une autre dans toutes circonstances. De fait, le choix d'une mesure de similarité dépend le plus souvent des caractéristiques des données et du contexte de l'étude.

Plusieurs approches ont été proposées pour automatiser le choix d'un algorithme de partitionnement (Pimentel et de Carvalho, 2019; Abdulrahman et al., 2018). Dans ce cadre, le "méta-apprentissage" ("meta-learning") peut guider la construction de classifieurs en fonction des caractéristiques des données et de celles des algorithmes (Brazdil et al., 2008). Par exemple, il est possible, à l'image du raisonnement à partir de cas, de s'appuyer sur une base d'épisodes de résolution de problèmes et d'adapter un des épisodes (le plus proche) au traitement des données courantes. Dans notre cas, c'est plutôt le choix d'une mesure de similarité

1. Nous parlons ici de "classification non supervisée" ou de "clustering".

pour le partitionnement qui retient notre attention, un problème qui reste peu abordé. Ainsi, dans cet article nous proposons une stratégie pour choisir automatiquement une mesure de similarité pour le partitionnement de données catégorielles et répondre à la question : "étant donnés un ensemble de mesures de similarité et un tableau de données [2], quelle mesure choisir qui soit bien adaptée au partitionnement des données à traiter".

L'article est organisé comme suit. Dans la section 2, nous rappelons les principales notions nécessaires à la compréhension de notre travail de recherche. Ensuite, dans la section 3 nous proposons une approche pour la sélection de mesures de similarité. Puis les expérimentations associées et une discussion suivent en section 4, ainsi qu'une conclusion et des perspectives en section 5.

2 Quelques mots sur l'état de l'art

Le méta-apprentissage.

Un système de "méta-apprentissage" peut tirer parti d'expériences passées pour traiter des problèmes courants (Brazdil et al., 2008). La représentation d'expériences peut passer par la création d'un "méta-tableau de données" où chaque ligne correspond à un ensemble de donnés et chaque colonne à une caractéristique globale d'un tel ensemble de données. Un attribut "cible" peut aussi être associé à chaque ligne du méta-tableau en lien avec un objectif donné, comme par exemple la mesure de similarité qui produit les meilleurs résultats au sens d'une métrique choisie pour le mesurer.

Ainsi, dans Pimentel et de Carvalho (2019), la sélection (ou recommandation) d'algorithmes de partitionnement s'appuie sur des caractéristiques combinant les mesures de similarités et le coefficient de corrélation de Spearman (ici les données sont numériques). Des histogrammes représentant les variations coefficient de corrélation servent de données d'entraînement pour construire un classifieur qui est ensuite utilisé pour prédire le meilleur algorithme de partitionnement pour un nouveau tableau de données.

Partitionnement de données catégorielles.

Il s'applique à des données non numériques, où les attributs sont à valeurs dans des domaines non totalement ordonnés et/ou non mesurables (Andritsos et Tsaparas, 2017). Il est possible de se ramener à un tableau de données binaire via un "échelonnage" ("scaling") et à transformer un attribut catégoriel en un ensemble d'attributs binaires (un par valeur). Mais cela peut conduire à une perte d'information, à la création d'un nombre considérable d'attributs binaires pour des tableaux de données quelquefois très creux, et enfin affecter les interprétations et les explications. Une autre façon de faire consiste à passer à des données numériques. Ainsi, l'algorithme "K-means" a été étendu dans (Ahmad et Dey, 2007) à des données mixtes, catégorielles et numériques, en adaptant la notion de "centre de gravité" d'une classe. Un état de l'art récent sur les algorithmes de partitionnement s'appliquant à des données mixtes est donné dans Ahmad et Khan (2019). L'algorithme proposé dans (Guha et al., 2000) traite quant à lui des données purement catégorielles en utilisant une méthode de classification hiérarchique.

Mesures de similarité pour les données catégorielles.

De nombreuses mesures de similarité permettent de quantifier la similarité entre deux objets. Une classification de mesures de similarité et de méthodes de partitionnement pour don-

2. Les données considérées sont représentées par des tableaux de données objets $\times$ attributs.

nées catégorielles associées est proposée dans (Alamuri et al., 2014). De même, une analyse et une comparaison de mesures de similarité sont réalisées dans (dos Santos et Zárate, 2015). Dans (Nguyen et al., 2019), deux mesures de similarité sont introduites, qui s'appuient sur la variabilité des données et l'entropie, et sont ensuite comparées à 11 autres mesures dans le cadre du partitionnement. À l'issue de l'analyse, les auteurs concluent que le comportement des algorithmes de partitionnement en fonction des mesures de similarité dépend très fortement des caractéristiques des données. Enfin, différentes mesures sont comparées dans (Boriah et al., 2008) pour détecter des données aberrantes ("outliers").

De façon à formaliser les mesures utilisées ici, considérons un tableau ou matrice de données $X = [x_{ij}]_{i=1,n}^{j=1,m}$, où n est le nombre de lignes et m le nombre de colonnes. Chaque objet (en ligne) x_i se décrit par m attributs (en colonne), où x_{ik} dénote la valeur du $k$ième attribut (la colonne A_k) pour le $i$ème objet. La similarité entre deux objets x_i et x_j, notée $S(x_i, x_j)$, se calcule en fonction de la similarité entre chaque attribut $S_k(x_{ik}, x_{jk})$ selon qu'ils s'apparient $(x_{ik} = x_{jk})$ ou pas $(x_{ik} \neq x_{jk})$. L'agrégation de ces tests d'appariement s'écrit :

$$S'(x_i, x_j) = \frac{1}{m} \sum_{k=1}^{m} S_k(x_{ik}, x_{jk}) \tag{1}$$

$$S''(x_i, x_j) = \frac{\sum_{k=1}^{m} S_k(x_{ik}, x_{jk})}{\sum_{k=1}^{m} \log p(x_{ik}) + \log p(x_{jk})} \tag{2}$$

La quatrième colonne de la table 1 indique la formule utilisée (S' ou S'') pour l'agrégation de la similarité de tous les attributs. Dans un partitionnement, il est possible d'utiliser aussi des mesures de dissimilarité D' et D'' qui sont associées aux mesures de similarité S' et S'' de la façon suivante :

$$D'(x_i, x_j) = 1 - S(x_i, x_j) \tag{3}$$

$$D''(x_i, x_j) = \frac{1}{S(x_i, x_j)} - 1 \tag{4}$$

La cinquième colonne de la table 1 indique la formule utilisée (D' ou D'') pour construire une dissimilarité globale par agrégation des dissimilarités. Les combinaisons données dans les quatrième et cinquième colonnes de la table 1 sont prédéfinies pour chaque mesure de la première colonne. Elles dépendent des domaines des attributs et des poids associés. Ainsi, l'équation 1 attribue le même poids à chaque élément de similarité tandis que l'équation 2 s'appuie sur une distribution de probabilités pour attribuer un poids. L'équation 3 s'applique aux mesures qui prennent leur valeurs entre 0 et 1, tandis que l'équation 4 s'applique à des mesures dont la valeur peut être plus grande que 1.

Nous employons également une notation de (Boriah et al., 2008) pour définir une mesure de similarité : $f_k(x)$ compte le nombre de fois que la valeur x apparaît pour l'attribut A_k et n_k est le nombre de valeurs distinctes que prend l'attribut A_k ; $\hat{p}_k(x)$ est la probabilité que l'attribut A_k prenne la valeur x et se calcule selon la formule $\hat{p}_k(x) = \frac{f_k(x)}{n}$. L'expression $p_k^2(x) = \frac{f(x)(f(x)-1)}{n(n-1)}$ propose une autre forme de calcul de la probabilité que l'attribut A_k prenne la valeur x.

	$S_k(x_{ik}, x_{jk})$			
Mesure	$x_{ik} = x_{jk}$	$x_{ik} \neq x_{jk}$	$S(x_i, x_j)$	$D(x_i, xj)$
Overlap	1	0	Eq. 1	Eq. 3
ES	1	$\frac{n_k^2}{n_k^2+2}$	Eq. 1	Eq. 4
G1	$1 - \sum_{q \in Q} p_k^2(q)$	0	Eq. 1	Eq. 3
G2	$1 - \sum_{q \in Q} p_k^2(q)$	0	Eq. 1	Eq. 3
G3	$1 - p_k^2(q)$	0	Eq. 1	Eq. 3
G4	$p_k^2(q)$	0	Eq. 1	Eq. 3
IOF	1	$\frac{1}{1+\log f_k(x_{ik}) \log f_k(x_{jk})}$	Eq. 1	Eq. 4
LIN	$2 \log \hat{p}_k(x_{ik})$	$2 \log(\hat{p}_k(x_{ik}) + \hat{p}_k(x_{jk}))$	Eq. 2	Eq. 4
LIN1	$\sum_{q \in Q} \log \hat{p}_k(q)$	$2 \log \sum_{q \in Q} \log \hat{p}_k(q)$	Eq. 2	Eq. 4
OF	1	$\frac{1}{1+\frac{n}{\log f_k(x_{ik})} \cdot \frac{n}{\log f_k(x_{jk})}}$	Eq. 1	Eq. 4

TAB. 1 – *Tableau de mesures de similarité (Boriah et al., 2008; Šulc et Řezanková, 2019).*

Parmi les nombreuses mesures de dissimilarité existantes (Boriah et al., 2008; Šulc et Řezanková, 2019), nous considérons uniquement les mesures listées dans la table 1 et les équations qui leur sont associées. Ainsi, les mesures G1, G2, G3 et G4, sont des extensions de la mesure de Goodall (Goodall, 1966). Les mesures G3, G4 et LIN (Lin, 1998) s'appuient sur la fréquence relative des données catégorielles et se calculent avec $\hat{p}_k(x)$. Il en va de même pour G1, G2 et LIN1, une extension de LIN.

3 L'approche proposée pour la sélection de mesures de similarité

Nous allons détailler notre approche sur la conception d'un système de méta-apprentissage pour la sélection automatique de mesures de similarité dans le cadre du partitionnement de données catégorielles. L'idée est de s'appuyer sur des partitionnements déjà effectués pour construire un modèle prédictif. Il nous faut construire un méta-tableau qui va servir d'entrée au système de méta-apprentissage qui va produire le modèle prédictif. Les tâches principales sont donc de construire le méta-tableau et d'en définir les caractéristiques.

Les caractéristiques des tableaux de données.

Les caractéristiques décrivant les tableaux de données en ligne portent sur les attributs et la nature des attributs, et doivent pouvoir être calculées de façon efficace. Elles sont communément utilisées dans la sélection d'algorithmes (Kalousis, 2002; Brazdil et al., 2008; Pimentel et de Carvalho, 2019). De plus, il existe une dépendance entre ces caractéristiques et la tâche où les mesures sont utilisées, ici le partitionnement (Šulc et Řezanková, 2019).

— N_A : nombre d'attributs (colonnes),
— N_I : nombre d'objets (lignes),
— le rapport N_A/N_I qui indique la densité du tableau de données,
— l'histogramme de l'entropie des attributs,
— l'histogramme de l'asymétrie ("skewness") des attributs,
— l'histogramme de l'aplatissement ("kurtosis") des attributs.

Histogramme	Intervalles
Entropie des attributs	[0,0.2],(0.2,0.4],(0.4,0.6],(0.6,0.8],(0.8,1]
Asymétrie	$(-\infty,-1],(-1,-0.5],(-0.5,0.5],(0.5,1],(1,+\infty)$
Aplatissement	$(-\infty,-1],(-1,-0.5],(-0.5,0.5],(0.5,1],(1,+\infty)$

TAB. 2 – *Les caractéristiques et les histogrammes associés.*

Certaines caractéristiques sont décrites sous la forme d'histogrammes, qui résument les informations nécessaires relatives à un attribut (Ferrari et De Castro, 2015). Tous les histogrammes sont calibrés de la même façon, autour de 5 intervalles, comme le montre la table 2.

L'objectif.

Le méta-tableau de données d'entrée est complété avec l'introduction d'un attribut "objectif" pour chaque ensemble de données en ligne, qui fait référence à la "meilleure" mesure de similarité. De cette façon, il est possible d'associer un ensemble de données en ligne, ses caractéristiques et le potentiel d'une mesure de similarité lors d'un partitionnement. Pour ce faire, nous utilisons l'index de Rand (ARI pour "Adjusted Rand Index") (Hubert et Arabie, 1985) comme critère pour comparer des partitions et pour indiquer qu'une mesure de similarité fournit la meilleure partition d'un ensemble de données. Pour chaque tableau de données en ligne, nous appliquons l'algorithme de partitionnement en utilisant toutes les mesures de similarités proposées en table 1. Ensuite, nous évaluons les partitions obtenues avec ARI, en comparant avec une partition de référence (voir la section suivante) et nous choisissons la mesure donnant l'ARI le plus élevé. La figure 1 montre le "workflow" associé au processus qui vient d'être décrit.

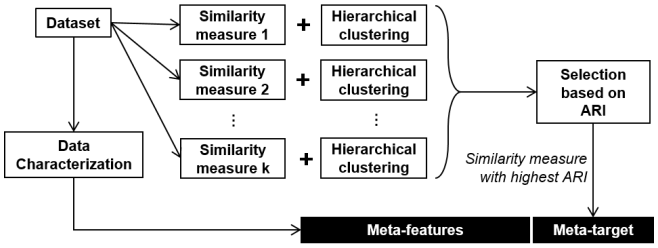

FIG. 1 – *Le "workflow" de traitement pour un tableau de données en ligne dans le méta-tableau.*

4 Expérimentations et discussion

Les conditions expérimentales.

Pour évaluer notre approche, nous avons mis en place une expérimentation qui porte sur 60 tableaux de données synthétiques [3]. Ces tableaux ont été engendrés avec la fonction *genRandomClust* du package *clusterGeneration R*. Nous avons fait varier les paramètres pour engendrer des tableaux de données avec différentes caractéristiques, par exemple nombre d'attributs,

3. Les tableaux peuvent être téléchargés à `http://guilhermealves.eti.br/datasets`

Paramètres	Valeurs, ensemble ou intervalle
Nombre d'attributs	{4,6,8,10}
Taille des partitions	[25,100]
Taille des domaines d'attributs	{2,3,4}, {2,3,5,6}, {6,8,10}
Nombre d'attributs bruités	0
Nombre de valeurs aberrantes	0

TAB. 3 – *Paramètres utilisés pour engendrer les données expérimentales.*

domaine des attributs et partitionnement. Effectivement, une partition est engendrée et associée au tableau de données, qui va servir de référence. Après la génération, les attributs numériques sont transformés en attributs catégoriels. Tous les détails des différents tableaux sont indiqués dans la table 3. La taille des tableaux de données varie entre 5 et 14 attributs (colonnes) et 139 et 354 instances (lignes). Comme les tableaux de données sont synthétiques et possèdent une partition de référence, il devient possible d'utiliser l'index de Rand ARI pour comparer une partition calculée avec une certaine mesure de similarité et la partition de référence.

Ensuite, nous utilisons un algorithme de classification ascendante hiérarchique (CAH) avec le lien moyen. Chaque tableau de données est traité avec toutes les mesures de similarités listées dans la table 1. Nous avons repris les paramètres par défaut de la CAH disponibles dans la librairie Scikit-learn (Pedregosa et al., 2011). Toutes les expérimentations ont été effectuées sur la même plate-forme, "Intel Core i7 8th generation (2.11 GHz)" avec 16 GB de RAM et Windows 7 x64, en utilisant le langage Python [4].

Lors des expérimentations, nous avons fixé le nombre de classes dans la partition et nous avons suivi le protocole "leave-one-out" pour faire les tests. les classifieurs Random Forest (RF (Breiman, 2001)), SVM, K-NN (K=5), AdaBoost, Naive Bayes et Decision Tree (DT) ont été utilisés comme systèmes de méta-apprentissage, en accord avec les recherches sur la sélection d'algorithmes (Brazdil et al., 2008; Pimentel et de Carvalho, 2019).

Résultats expérimentaux.

L'objectif principal des expérimentations est d'évaluer la potentialité de l'approche en comparant les résultats avec (1) le choix aléatoire d'une mesure de similarité, (2) une classification de référence ("ground truth"), (3) et la mesure " Overlap" qui est la mesure la plus simple entre deux attributs (voir la table 1). Pour (2), nous voulons avoir une borne supérieure de comparaison et pour cela nous simulons un oracle qui choisirait à chaque fois la mesure ayant le score ARI le plus élevé, La table 4 montre les résultats en termes d'exactitude ("accuracy") du modèle construit en utilisant les différents classifieurs par rapport à la stratégie aléatoire et à la mesure "Overlap". Les différents classifieurs entraînés sur le méta-tableau obtiennent de meilleurs résultats que la stratégie aléatoire et "Overlap", et en particulier Random Forest a le score le plus élevé.

Nous avons aussi étudié les qualités des partitions obtenues, ce qui est indiqué dans la seconde ligne de la table 4. Nous avons comparé le score obtenu par la mesure de similarité courante, avec celui de l'oracle qui est le plus élevé, le score lié au choix aléatoire et celui de la mesure "Overlap". On peut remarquer que les classifieurs montrent de meilleurs résultats que

4. Le code source est disponible à `http://github.com/asilvaguilherme/cat-sim-measure-selection`.

	Random	**Overlap**	**Naive**	**DT**	**SVM**	**5NN**	**AdaBoost**	**RF**	**Oracle**
Accuracy	0.100	0.083	0.300	0.450	0.333	0.333	0.333	0.467	-
ARI	0.444	0.484	0.539	0.558	0.600	0.585	0.563	0.597	0.653

TAB. 4 – *Les valeurs de l'exactitude des classifieurs ("accuracy") et mesure de la qualité des partitions associées (ARI).*

l'aléatoire et "Overlap", un peu moins bon que l'oracle, ce qui montre que notre approche est prometteuse et doit être davantage étudiée.

5 Conclusion

Dans cet article, nous avons proposé une approche pour la sélection automatique d'une mesure de similarité en partitionnement de données catégorielles. L'approche s'appuie sur un système de méta-apprentissage qui traite un méta-tableau composé de tableaux de données en ligne et de caractéristiques globales de ces tableaux en colonne. Les partitionnements construits sont comparés grâce à l'index de Rand. Pour les expérimentations, des tableaux synthétiques sont engendrés avec une classification associée. Cela permet de tester différentes mesures de similarité et différents classifieurs dont Decision Tree, Random Forest et SVM entre autres, de mesurer leur exactitude et leur pouvoir prédictif, c'est à dire leur potentiel à découvrir la meilleure mesure de similarité pour un nouvel ensemble de données à tester.

Plusieurs extensions sont envisageables. Ainsi, il faut d'abord évaluer l'approche sur des données réelles. Ensuite il nous faut étudier la combinaison de classifieurs où le choix des algorithmes et des mesures associées pourraient être automatisés.

Références

Abdulrahman, S. M., P. Brazdil, J. N. van Rijn, et J. Vanschoren (2018). Speeding up algorithm selection using average ranking and active testing by introducing runtime. *Machine learning 107*(1), 79–108.

Ahmad, A. et L. Dey (2007). A k-mean clustering algorithm for mixed numeric and categorical data. *Data & Knowledge Engineering 63*(2), 503–527.

Ahmad, A. et S. S. Khan (2019). Survey of state-of-the-art mixed data clustering algorithms. *IEEE Access 7*, 31883–31902.

Alamuri, M., B. R. Surampudi, et A. Negi (2014). A survey of distance/similarity measures for categorical data. In *2014 International joint conference on neural networks (IJCNN)*, pp. 1907–1914. IEEE.

Andritsos, P. et P. Tsaparas (2017). Categorical data clustering. In *Encyclopedia of Machine Learning and Data Mining*, pp. 188–193. Springer.

Barioni, M. C. N., H. Razente, A. M. Marcelino, A. J. Traina, et C. Traina Jr (2014). Open issues for partitioning clustering methods : an overview. *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery 4*(3), 161–177.

Boriah, S., V. Chandola, et V. Kumar (2008). Similarity measures for categorical data : A comparative evaluation. In *Proceedings of the 2008 SIAM international conference on data mining*, pp. 243–254.

Brazdil, P., C. G. Carrier, C. Soares, et R. Vilalta (2008). *Metalearning : Applications to Data Mining*. Springer.

Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

dos Santos, T. R. et L. E. Zárate (2015). Categorical data clustering : What similarity measure to recommend ? *Expert Systems with Applications 42*(3), 1247–1260.

Ferrari, D. G. et L. N. De Castro (2015). Clustering algorithm selection by meta-learning systems : A new distance-based problem characterization and ranking combination methods. *Information Sciences 301*, 181–194.

Goodall, D. W. (1966). A new similarity index based on probability. *Biometrics*, 882–907.

Guha, S., R. Rastogi, et K. Shim (2000). Rock : A robust clustering algorithm for categorical attributes. *Information systems 25*(5), 345–366.

Hubert, L. et P. Arabie (1985). Comparing partitions. *Journal of classification 2*(1), 193–218.

Kalousis, A. (2002). *Algorithm selection via meta-learning*. Ph. D. thesis, University of Geneva.

Lin, D. (1998). An Information-Theoretic Definition of Similarity. In *Proceedings of ICML*, pp. 296–304.

Nguyen, T.-H. T., D.-T. Dinh, S. Sriboonchitta, et V.-N. Huynh (2019). A method for k-means-like clustering of categorical data. *Journal of Ambient Intelligence and Humanized Computing*, 1–11.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, et E. Duchesnay (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research 12*, 2825–2830.

Pimentel, B. A. et A. C. de Carvalho (2019). A new data characterization for selecting clustering algorithms using meta-learning. *Information Sciences 477*, 203–219.

Šulc, Z. et H. Řezanková (2019). Comparison of similarity measures for categorical data in hierarchical clustering. *Journal of Classification 36*(1), 58–72.

Summary

Data clustering is a well-known task in data mining and it often relies on distances or, in some cases, similarity measures. The latter is indeed the case for real world datasets that comprise categorical attributes. Several similarity measures have been proposed in the literature, however, their choice depends on the context and the dataset at hand. In this paper, we address the following question: *given a set of measures, which one is best suited for clustering a particular dataset?* We propose an approach to automate this choice, and we present an empirical study based on categorical datasets, on which we evaluate our proposed approach.