Clustering de Données Complexes Representées par des Formules Booléennes

Abdelhamid Boudane*, Said Jabbour*
Lakhdar Sais* Yakoub salhi*

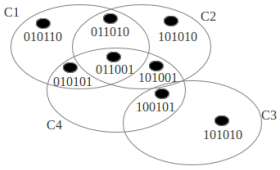
*CRIL - CNRS, Université d'Artois
Rue Jean Souvraz, SP-18
62307, Lens Cedex 3
{boudane,jabbour,sais,salhi}@cril.fr

Résumé. La fouille de données est un domaine de recherche multidisciplinaire. Diverses approches ont été proposées pour résoudre différentes tâches d'extraction de connaissances à partir de données. Cette diversité est souvent due à la multitude des types de données disponibles. Au-delà des données numériques, elles peuvent aussi être de nature symbolique. Dans ce travail, nous introduisons un nouveau problème de clustering de données symboliques représentées par des formules Booléennes. Les algorithmes bien connus de clustering, le k-means, les approches hiérarchiques et divisives, ont été étendus au cadre que nous proposons.

1 Clustering de Formules Booléenes

De nombreuses approches ont été proposées pour le problème de clustering (e.g. Aggarwal et Reddy (2014); Bouguettaya et al. (2015); MacQueen (1967)). Elles dépendent souvent des applications et des types de données considérées qui peuvent être de nature transactionnelles, séquentielles, arborescentes, graphes, textes, etc. Dans de nombreuses situations réelles, les objets à classifier sont souvent exprimés par des attributs numériques. Or, il existe des situations réels où les objets sont complexe et de nature plutôt symbolique. Dans ce travail, nous nous intéressons aux cas des données symboliques représentées par des formules booléennes. Ces données peuvent par exemple provenir des résultats d'un sondage, des préférences des utilisateurs ou des clients.

A titre d'illustration, dans la Figure 1, nous donnons un exemple décrivant les préférences de quatre clients auprès d'un concessionnaire de voitures qui propose certaines marques de voitures avec des options possibles sur la couleur et la motorisation. Chaque option est représentée par une variable Booléenne : r (rouge), b (blanc), d(diesel), e(essence), c(citroen), p(peuguot). Les préférences de chaque client $C_i (1 \leq i \leq 4)$ sont exprimées par les formules Booléennes. Le premier client émet une forte reserve sur la couleur rouge ($\neg r$), le second désire une voiture diesel (d), le troisième préfère une voiture à essence de couleur rouge ($e \wedge r$), alors que le quatrième client préfère une voiture de la marque Peugeot (p). Les modèles sont donnés en suivant l'ordre des variables r, b, d, e, c et p.



Clients	C_1	C_2	C_3	C_4
Formules	$\neg r$	d	$e \wedge r$	p
Modèles	010101 010110 011001 011010	011001 011010 101001 101010	 100101 100110	010101 011001 100101 101001

FIG. 1 – *Concéssionaire de voitures : un exemple*

	C_1	C_2	C_3	C_4
C_1	1	$1/3$	0	$1/3$
C_2	$1/3$	1	0	$1/3$
C_3	0	0	1	$1/5$
C_4	$1/3$	$1/3$	$1/5$	1

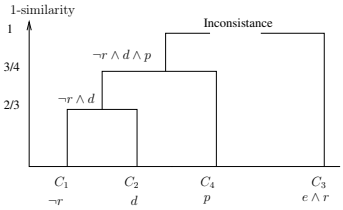


FIG. 2 – *Déroulement de l'algorithme hiérarchique agglomératif*

Cet exemple illustratif donne une idée sur la représentation compacte et la puissance d'expression qu'offre la logique pour représenter les données avec une possibilités d'exprimer des contraintes et de raisonner sur les données.

2 Adaptations du k-means et d'un algorithme hiérarchique agglomératif

Notre adaptation au clustering de formules est obtenue en considérant les choix suivants pour la distance et le représentant d'un cluster :

— **Distance** : on définit la similarité par $S(\varphi, \psi) = \frac{|M(\varphi \wedge \psi)|}{|M(\varphi \vee \psi)|}$, ou φ et ψ sont deux formules Booléennes et M représente l'ensemble des modèles (Tversky et Shafir, 2004). La distance est définit par $d(\varphi, \psi) = 1 - S(\varphi, \psi)$.

— **Représentant d'un cluster** : la conjonction de toutes les formules.

L'objectif est de trouver des clusters en maximisant la taille des intersections entre les modèles des formules. La Figure 2 donne la matrice des distances et le résultat du déroulement de l'algorithme hiérarchique agglomératif sur l'exemple du concessionaire de voiture. Le problème majeur de ces adaptation et l'utilisation d'un grand nombre d'appels à une fonction de mesure de similarité qui n'est pas polynomiale est qui est très coûteuse en temps.

3 Algorithme divisif

En considérant que les formules logiques représentent sémantiquement des ensembles de modèles, l'algorithme divisif consiste à prendre toutes les formules dans un même cluster puis à diviser itérativement le mauvais cluster jusqu'a satisfaction d'un critère d'arrêt. Intuitivement, le mauvais cluster est celui qui possède la plus petite intersection. En représentant un cluster

par un hyper-graphe où les modèles représentent les noeuds et les formules représentent les hyper-arêtes, le problème de division peut être résolu en calculant les transversaux minimum (*minimum hitting set*). À chaque itération de division, on supprime préalablement les éléments de l'intersection I de l'ensemble des noeuds et on calcule ensuite un *hitting set* minimum S. La construction des nouveaux clusters se fait à partir des éléments de S. Nous avons également proposé une approche basée sur la syntaxe sans passer par les modèles de sformules. Cette approche se base sur un codage vers SAT qui permet de résoudre le problème de trouver k cluster consistants dans un ensemble de formules. Elle est utilisés ensuite pour trouver le *hitting set* minimum sans passer par l'enumération des modèles.

Références

Aggarwal, C. C. et C. K. Reddy (Eds.) (2014). *Data Clustering : Algorithms and Applications*. CRC Press.

Bouguettaya, A., Q. Yu, X. Liu, X. Zhou, et A. Song (2015). Efficient agglomerative hierarchical clustering. *Expert Systems with Applications 42*(5), 2785 – 2797.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam et J. Neyman (Eds.), *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability - Vol. 1*, pp. 281–297. University of California Press, Berkeley, CA, USA.

Tversky, A. et E. Shafir (2004). *Preference, Belief, and Similarity : Selected Writings*. Bradford books. MIT Press.

Summary

Data mining is a multidisciplinary research area. Several approaches have been proposed to deal with several data mining tasks. This diversity is often due to the multitude of available types of data. Beyond numeric data, they can also be symbolic in nature. In this paper, we focus on the problem of clustering symbolic data, represented using Boolean formulas.