

Vers une modularité pour données vectorielles

David Combe*, Christine Largeron*
Előd Egyed-Zsigmond**, Mathias Géry*

*Université de Lyon, F-42023, Saint-Étienne, France,
CNRS, UMR 5516, Laboratoire Hubert Curien, F-42000, Saint-Étienne, France
Université de Saint-Étienne, Jean-Monnet, F-42000, Saint-Étienne, France
{david.combe, christine.largeron, mathias.gery}@univ-st-etienne.fr

**Université de Lyon
UMR 5205 CNRS, LIRIS
7 av J. Capelle, F 69100 Villeurbanne, France
elod.egyed-zsigmond@insa-lyon.fr

Résumé. La modularité, introduite par Newman pour mesurer la qualité d'une partition des sommets d'un graphe, ne prend pas en compte d'éventuelles valeurs associées à ces sommets. Dans cet article, nous introduisons une mesure de modularité complémentaire, basée sur l'inertie, et adaptée pour évaluer la qualité d'une partition d'éléments représentés dans un espace vectoriel réel. Cette mesure se veut un pendant pour la classification non supervisée de la modularité de Newman. Nous présentons également 2Mod-Louvain, une méthode utilisant ce critère de modularité basée sur l'inertie conjointement à la modularité de Newman pour détecter des communautés dans des réseaux d'information. Les expérimentations que nous avons menées ont montré qu'en exploitant à la fois les données relationnelles et vectorielles, 2Mod-Louvain détectait plus efficacement les communautés que des méthodes utilisant un seul type de données et qu'elle était robuste face à des dégradations des données.

1 Introduction

Dans le cas de la classification automatique comme pour la détection de communautés, après avoir construit une partition à l'aide d'une méthode, il convient d'évaluer sa qualité. Pour ce faire, on peut faire appel à des critères externes ou internes. Les premiers permettent de comparer le résultat obtenu avec un résultat attendu, par exemple une partition faisant office de "vérité terrain" alors que les seconds quantifient la qualité de la partition proposée. Parmi les critères externes, utilisables aussi bien en classification automatique que pour la détection de communautés, on peut citer le taux de bien classés, la pureté, l'indice de Rand ou sa version ajustée, l'entropie ou encore l'information mutuelle, éventuellement normalisée ou ajustée, mais aussi la V-mesure, l'homogénéité ou la complétude (Hubert et Arabie (1985); Vinh et al. (2010); Rosenberg et Hirschberg (2007)).

En classification automatique, les critères internes peuvent eux-mêmes être subdivisés en critères dont l'usage est spécifique à une distance ou à une méthode, comme par exemple l'iner-