Une Etude Comparative entre Motifs Graduels et Corrélations Statistiques

Maureen Domche*, Jerry Lonlac**, Norbert Tsopze* Engelbert Mephu Nguifo***

*Département d'Informatique, Université de Yaoundé 1, Cameroun
{maureennouno2000,tsopze.norbert}@gmail.com
**IMT Nord Europe, IMT, Université de Lille, CERI SN, Lille, 59653, France
jerry.lonlac@imt-nord-europe.fr
***Université Clermont Auvergne, Clermont Auvergne INP, ENSMSE, CNRS, LIMOS
engelbert.mephu_nguifo@uca.fr

1 Contexte et problématique

Les motifs graduels modélisent les co-variations fréquentes entre attributs de la forme "plus/moins x_1, ..., plus/moins x_n" à partir de données numériques. L'intérêt suscité par ces motifs a conduit à développer des algorithmes efficaces pour leur extraction (Lonlac et Nguifo, 2020). Par ailleurs, la corrélation statistique qui met en évidence des liaisons entre variables permet également d'exprimer des co-variations dans des données numériques. Bien que les motifs graduels et les corrélations statistiques capturent les co-variations dans les données, les motifs graduels fournissent des connaissances plus expressives alors que les corrélations statistiques sont plus faciles à calculer. Ce travail fait une étude comparative entre les motifs graduels extraits en utilisant deux sémantiques de gradualité (Di-Jorio et al., 2009; Laurent et al., 2009) et les corrélations statistiques, en présentant les similarités, différences, avantages de chacune des notions pour le traitement des données numériques. Les expérimentations effectuées confirment l'expressivité des motifs graduels par rapport aux corrélations statistiques.

2 Analyse comparative

Un motif graduel représenté sous forme $(a_1^{*_1}, \ldots, a_k^{*_k})$, est défini comme un ensemble d'items graduels (formé d'un attribut a_i et d'une variation $* \in \{\leq, \geq\}$ croissante/décroissante). Il est évalué à travers un support traduisant sa fréquence. Une corrélation statistique quant à elle évalue la force de la relation entre plusieurs variables quantitatives au moyen d'un coefficient qui informe sur le sens et l'intensité de cette relation. Différentes mesures de corrélation existent (Pearson, Kendall, Spearman, etc) dépendant du type de données et la nature de la relation à évaluer. Les différences entre ces deux concepts sont synthétisées dans le tableau 1.

Le tableau 2 présente les résultats obtenus sur des données portant sur la qualité de l'air. Les motifs de 2, 3 et 4 attributs (avec les sémantiques de gradualité GRAANK et de GRITE) sont relevés et des mesures de corrélation associées sont calculées. Ces motifs informent sur le sens

	Motifs graduels	**Attributs corrélés**
Sens de variation	variation de chaque attribut	variation des attributs non renseignée
Intensité	support : fréquence du motif dans les données	coefficient : force de la relation linéaire (Pearson) ou monotone (Kendall, Spearmann) pour 2 variables
Méthode	algorithme : extrait des informations spécifiques (locale)	formule : résume un ensemble de valeurs (globale)
Complexité	exponentielle	linéaire ou quadratique
Interprétation	facile : résumé synthétique d'une partie des données (extension)	moins facile : nécessite une analyse en plus (graphiques) pour exploiter et comprendre le résultat
Paramétrage de la méthode	support minimum à fournir	aucun
Nombre d'attributs	co-variations entre plusieurs attributs	mesure de corrélation limitée à trois attributs

TAB. 1 – *Récapitulatif de la comparaison entre motifs graduels et attributs corrélés*

de variation de chaque attribut contrairement aux corrélations. Il est plus difficile d'obtenir une mesure générale qui exprime la relation entre plus de deux variables avec les formules existantes pour la corrélation. La sémantique de GRAANK inspirée des corrélations entre rangs, se rapproche plus des mesures de corrélation que celle de GRITE basée sur l'ordre entre objets.

Attributs	Motifs graduels	**GRITE**	**GRAANK**	**r**	ρ	τ	$\mathbf{R}_{x,yz}$	$\mathbf{R}_{y,zx}$	$\mathbf{R}_{z,xy}$	$\bar{R}$
$4, 5$	$(4^{\geq}, 5^{\geq})$	0.49	0.61	0.85	0.9	0.73	-	-	-	-
$6, 8$	$(6^{\geq}, 8^{\geq})$	0.22	0.85	0.82	0.90	0.76	-	-	-	-
$1, 4, 7$	$(1^{\geq}, 4^{\geq}, 5^{\geq})$	< 0.08	0.723	-	-	-	0.08	0.77	0.77	0.54
$1, 6, 8$	$(1^{\geq}, 6^{\geq}, 8^{\geq})$	0.16	0.71	-	-	-	0.68	0.87	0.82	0.79
$2, 4, 5, 9$	$(2^{\geq}, 4^{\geq}, 5^{\geq}, 9^{\geq})$	< 0.08	0.705	-	-	-	-	-	-	-
$2, 4, 5, 10$	$(2^{\geq}, 4^{\geq}, 5^{\geq}, 10^{\geq})$	< 0.08	0.789	-	-	-	-	-	-	-

TAB. 2 – *Quelques résultats sur les données Air Quality*

3 Conclusion

Ce travail [1] présente une étude comparative entre corrélations statistiques et motifs graduels, tous deux révélant des tendances entre variables dans les données numériques, en mettant en évidence leur particularité, leurs avantages. Les motifs graduels permettent la prise en compte d'un grand nombre de variables dans la relation et sont plus expressifs pour expliquer la relation, avec pour défaut majeur la complexité exponentielle de leur énumération. En revanche, les corrélations sont faciles à calculer mais limitées pour expliquer la relation.

Références

Di-Jorio, L., A. Laurent, et M. Teisseire (2009). Mining frequent gradual itemsets from large databases. In *IDA*, pp. 297–308.

Laurent, A., M. Lesot, et M. Rifqi (2009). GRAANK : exploiting rank correlations for extracting gradual itemsets. In *FQAS*, pp. 382–393.

Lonlac, J. et E. M. Nguifo (2020). A novel algorithm for searching frequent gradual patterns from an ordered data set. *Intell. Data Anal. 24*(5), 1029–1042.

1. Est soutenu par le CNRS à travers le projet DSCA AAP-Afrique FDMI-AMG : fdmi.limos.fr.