AriCode un package R pour calculer efficacement l'ARI et d'autres mesures pour comparer des classifications.

J. Chiquet[a], **V. Dervieux**[a,b] and **G. Rigaill**[a,b]

[a]MIA Paris
UMR 518 AgroParisTech/INRA
julien.chiquet@agroparistech.fr

[b] LaMME & IPS2
UMR CNRS/UEVE/ENSIIE/INRA
UMR CNRS/INRA/UEVE/Uni. Paris-Diderot/Uni. Paris-Sud
guillem.rigaill@inra.fr

Mots clefs : Classification, ARI, NID,

Les mesures permettant de comparer des classifications - comme l'ARI (Adjusted Rand Index) ou le NID (Normalized Information Distance) [1] - sont essentielles pour évaluer la qualité d'une classification ou plus généralement pour sélectionner une clasification stable et robuste [2].

Considérons deux classifications $\mathcal{C}_1$ et $\mathcal{C}_2$ de n observations en respectivement c_1 et c_2 classes. Pour la majorité des mesures (en particulier l'ARI et le NID), il faut calculer le nombre d'observations qui sont simultanément dans la classe j de la classification $\mathcal{C}_1$ et la classe k de la classification $\mathcal{C}_2$ pour toutes paires (j, k) possibles. Il y en a $c_1 c_2$. Un calcul naïf de ces quantités a une complexité de $\mathcal{O}(n + c_1 c_2)$ en temps et en espace. Cela est prohibitif quand n et c_1 et c_2 sont grands.

Nous proposons un algorithme simple de complexité $\mathcal{O}(n)$ utilisant le "bucket sorting" [3]. Nous avons implementé cet algorithme dans le package R ariCode (https://github.com/jchiquet/aricode) Ce package est typiquement un ordre de magnitude plus rapide que des implémentations classiques de l'ARI en R.
Nous illustrons l'utilisation d'aricode sur des données de métagénomique ou n, c_1 et c_2 sont grands. Nous illustrons également comment l'ARI ou le NID peuvent être utilisées pour évaluer la robustesse d'une classification grace à du rééchantillongage.

Références

[1] Nguyen Xuan Vinh, Julien Epps and James Bailey. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance, Journal of Machine Learning Research, 2009

[2] Ulrike von Luxburg, Clustering stability: an overview. Foundations and Trends in Machine Learning, 2 (3), 235-274, 2010.

[3] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. Introduction to Algorithms, MIT press, 2009.