Utilisation et évaluation de la concordance de plusieurs méthodes de classification sur des données cliniques de maladies rares

Emmanuelle Besse & Damien Chimits & Eva-Maria Hüßler & Virginie Stanislas

Ecole Nationale de la Statistique et de l'Analyse de l'Information (Ensai)
emmanube@hotmail.com, damien.chimits@gmail.com, eva.huessler@tu-dortmund.de et virginie.stanislas@orange.fr

Résumé

Ce projet a pour but de comparer différentes méthodes de classification afin de valider ou d'infirmer une classification préalablement réalisée sur des patients atteints de la maladie de Wegener ou de la polyangéite microscopique (PAM). Ces deux maladies rares appartiennent à une même famille, celle des « vascularites associées aux ANCA [1] » et partagent d'autres symptômes communs les rendant difficilement différentiables. D'un premier travail réalisé par Mahr A. et al. [5] est ressorti une classification en cinq classes. Cette étude permet de conclure à l'existence d'une multiplicité de pathologies plutôt qu'à celle de deux maladies distinctes. Nos travaux se basent sur les résultats de cette étude. Ils ont pour but de tester d'autres méthodes de classification et de comparer les rsultats à ceux réalisés dans la publication de Mahr A. et al. [5]. Nous serons ainsi en mesure de valider, ou au contraire d'infirmer la partition initiale. Après avoir mis en œuvre ces différentes méthodes de classification et étudié la pertinence statistique des partitions obtenues, nous avons cherché à identifier des groupes de partitions homogènes, à l'aide de plusieurs outils, afin d'analyser plus simplement leur cohérence (d'un point de vue clinique). La partition la plus pertinente aboutit à 5 classes.
Nous avons analysé la robustesse de cette dernière avant de la comparer à la partition initiale de Mahr A. et al. [5]. Les deux partitions comparées étant similaires, nous pouvons alors conclure à la pertinence de la partition initiale.

Mots-clés. Méthodes de classification, Maladie de Wegener, Polyangéite microscopique

1. Anti-Neutrophilic Cytoplasmic Antibodies

Abstract

The aim of this project was to compare various methods of cluster analysis in order to validate the results of a previous cluster analysis realized on a sample of patients diagnosed with either Wegener's Granulomatosis or Microscopic Polyangiitis. The two diseases are associated with anti-neutrophil cytoplasmic antibodies (ANCA) and have shown similar symptoms. Because of their overlapping features, the differential diagnosis is difficult.

Mahr A. et al. [5] found a five cluster solution suggesting not two but several distinct pathologies. In order to validate these results, we tested several clustering methods and compared them to Mahr A. et al. [5] findings. After implementing these clustering methods and studying the statistical pertinence of the partitions obtained, we tried to find homogeneous groups of partitions in order to analyze their clinical coherence. The most pertinent partition lead to 5 clusters. After analyzing its robustness we compared it to Mahr A. et al [5] findings. Since the two partitions are similar, we concluded that the initial five cluster solution is pertinent.

Keywords. Clustering methods,Wegener's Granulomatosis, Microscopic Polyangiitis

1 Introduction

La maladie de Wegener et la polyangéite microscopique (PAM) sont deux maladies rares qui appartiennent à une même famille, celle des « vascularites associées aux ANCA [2] ». Ce sont toutes deux des maladies auto-immunes [3], cela signifie que les défenses de l'organisme (défenses immunitaires), qui normalement s'attaquent aux éléments « étrangers » (comme les bactéries, les virus...), se retournent contre les cellules propres à celui-ci et les attaquent. Dans ce cas, le système immunitaire produit des anticorps [4] nocifs, appelés auto-anticorps : les ANCA. Ces ANCA entraînent la destruction de certains agents comme les vaisseaux ou les articulations.

Étant caractérisées par la présence d'ANCA, ces maladies possèdent d'autres symptômes communs les rendant difficilement différentiables. De plus, selon le Dr. Alfred Mahr, dans 30 à 50% des cas, les médecins ont du mal à déterminer de quelle maladie le patient est atteint. Afin de répondre à ce problème, des étudiants de l'Ensai ont travaillé sur un projet ayant pour objectif d'établir des groupes homogènes de patients au sens des caractéristiques cliniques et biologiques. Ces groupes visaient à retrouver les deux entités (la maladie de Wegener et la polyangéite microscopique) ou au contraire d'affirmer la

2. Anti-Neutrophilic Cytoplasmic Antibodies

3. Bien que la cause de la PAM ne soit pas connue, il s'agit très probablement d'une maladie auto-immune

4. Substance défensive engendrée par l'organisme

multiplicité des maladies. Le dernier objectif de la classification était de déterminer si les sous-groupes obtenus avaient des profils évolutifs distincts (en termes de rechute et de mortalité).

De ce travail est ressorti une étude complémentaire prenant en compte des troubles digestifs qui n'étaient pas présents au préalable. Une nouvelle classification a été réalisée, composée de cinq classes ayant fait l'objet d'une publication scientifique dans la revue : *Annals of the Rheumatic Diseases*, « Revisiting the classification of clinical phenotypes of anti-neutrophil cytoplasmic antibody-associated vasculitis : a cluster analysis »[6].

À partir de ces résultats, le but de ce projet est d'infirmer ou de valider les résultats de la classification publiée, obtenue à partir d'une des méthodes les plus souvent utilisées, à savoir : une ACM suivie d'une CAH avec critère de Ward puis consolidée par la méthode des Kmeans.

2 Méthodologie

Afin de vérifier la pertinence de la classification publiée, diverses méthodes de classification ont été implémentées via le logiciel R. Ces dernières diffèrent sur plusieurs points. En effet certaines ne sont utilisables qu'après une analyse factorielle, alors que d'autres peuvent aussi être mises en place à partir d'un tableau de données binaires. Ces méthodes demandent d'effectuer plusieurs choix, comme la distance qui permettra de calculer la proximité entre deux individus (ou deux classes), ou encore le critère d'agrégation. Ainsi, ont été utilisées les méthodes hiérarchiques (agrégation ascendante et descendante), par partitionnement (algorithmes Kmeans, PAM et CLARA), mais également des méthodes se basant sur la densité (DBSCAN) ou sur une approche probabiliste (algorithme EM).

À partir de chacune de ces méthodes un grand nombre de partitions a pu être réalisé. Le choix du nombre de classes à retenir pour chacune de ces partitions s'est fait à partir de plusieurs critères (graphiques, selon le R-Square ou encore le BIC). De cette façon, nous avons sélectionné 55 partitions comparables les unes aux autres par le critère de la silhouette. Par la suite, une classification de ces partitions a été réalisée. Cette dernière visait à regrouper celles qui se ressemblent le plus afin de pouvoir se concentrer sur des groupes de partitions et non plus sur les méthodes unes à unes. L'analyse du dendrogramme de cette classification nous a poussés à conserver un découpage en sept groupes. Un autre outil, la heatmap, a mis en évidence deux groupes homogènes. Pour chaque groupe de partitions, une d'entre elles a été choisie (selon le critère de la silhouette) afin de les représenter.

3 Résultats

Nous nous intéressons désormais aux caractéristiques des partitions de chacun des sept groupes.

Le premier et le deuxième groupes sont constitués principalement de partitions issues des méthodes Kmeans et CAH. On y retrouve des partitions similaires à celles de la publication.
Le troisième est constitué de partitions issues de la méthode PAM qui permettent de bien distinguer le rôle de l'atteinte digestive dans les risques de décès.
Les quatrième, cinquième et sixième groupes regroupent des partitions issues de méthodes variées qui mettent à jour le rôle de la positivité MPO dans les risques de décès.
Le dernier groupe est composé des partitions issues de méthodes variées qui mettent à jour le rôle des signes digestifs et cardiovasculaires dans les risques de décès.

Parmi tous nos résultats, nous nous sommes intéressés à sélectionner les groupes de partitions ayant le plus grand sens clinique et statistique. Les deux premiers groupes sont ceux les plus cliniquement interprétables. Au sein de ces deux groupes, la partition Kmeans de l'algorithme McQueen est celle qui a le plus grand sens statistique (au regard du critère de la silhouette). Le consensus clustering montre, de plus, la robustesse de cette partition.

Nous présentons ici les cinq groupes obtenus à l'aide de cette partition dans l'ordre croissant par rapport au niveaux de mortalité observé (le premier présentant le niveau de mortalité le plus faible) :

- « Non renal » : Les patients qui ne présentent aucun signe rénal, mais des manifestations cliniques telles que des signes ORL et occulaires. Ces patients présentent majoritairement une positivité PR3-ANCA.
- « Atteinte Rénale et PR3-ANCA » : Les patients ont des caractéristiques similaires à la classe précédente mais sont tous atteints de troubles rénaux.
- « Atteinte Rénale sans PR3-ANCA » : Les individus de ce groupe sont tous atteints de troubles rénaux et se caractérisent par une positivité de type MPO-ANCA et non PR3-ANCA.
- « Atteinte cardiovasculaire » : Tous les individus de ce groupe présentent des troubles cardiovasculaires (contrairement aux groupes précédents), ils se caractérisent également par la présence de symptômes rénaux et par une positivité ANCA aussi bien de type MPO que PR3.
- « Atteinte digestive » : Ce groupe comprend tous les individus atteints de troubles digestifs, certains d'entre eux présentent également des troubles cardiovasculaires. Ils se caractérisent aussi par la présence de symptômes rénaux et par une positivité ANCA aussi bien de type MPO que PR3.

Étant donné que cette partition est identique à celle de la publication, nous pouvons finalement accepter la partition initiale comme étant la plus pertinente sur nos données.

Effectuer une avant les Kmeans permet une consolidation plus rapide de celle-ci. Il serait donc intéressant de savoir s'il est plus « rentable » d'utiliser les Kmeans seuls ou non.

Un grand nombre de méthodes a été ici écarté car elles ne donnaient pas de résultats probants sur nos données. On pourrait alors chercher dans quel cadre, et sur quel(s) type(s) de données elles s'appliqueraient de manière optimale. Dans le cadre de l'étude initiale, il serait intéressant d'étudier l'impact des traitements fournis aux patients suite à la connaissance de cette partition auprès des médecins qui l'utilisent.

Bibliographie

[1] BOUBOU M. (2007), *Contribution aux méthodes de classification non supervisée via des approches prétopologiques et d'agrégation dopinions*, Thèse de doctorat d'université, Lyon : Université Claude Bernard Lyon I.
[2] ESTER M., KRIEGEL H.-P., SANDER J., XU X. (1996), « A Density-Based Algorithm for Discovering Clusters in Large Spatial Databses with Noise », *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining*, KDD, p. 226-231.
[3] FRALEY C., RAFTERY A. E. (2002), « Model-Based Clustering, Discriminant Analysis, and Density Estimation », *Journal of the American Statistical Association* 97, p. 611-632.
[4] KAUFMAN L., ROUSSEEUW P. J.(1990), *Finding groups in data : an introduction to cluster analysis*, Wiley.
[5] MAHR A., KATSAHIAN S., VARET H., GUILLEVIN L., HAGEN C., HÖGLUND P., MERKEL P., RASMUSSEN N., WESTMAN K., JAYNE D.(2012), « Revisiting the Classification of Clinical Phenotypes of Anti-Neutrophil Cytoplasmic Antibody-Associated Vasculitis : a Cluster Analysis », *Annals of the Rheumatic Diseases*.
[6] MONTI S., TAMAYO P., MESIROV J., GOLUB T. (2003), « Consensus Clustering : A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data », *Machine Learning*, 52, p. 91-118.
[7] MORISSETTE L., CHARTIER S. (2013), « The k-means clustering technique : General considerations and implementation in Mathematica », *Tutorials in Quantitative Methods for Psychology*, 9(1), p.15-24.
[8] ROUSSEEUW P. J. (1987), « Silhouettes : a graphical aid to the interpretation and validation of cluster analysis », *Journal of Computational and Applied Mathematics*, 20, p. 53-65.
[9] TUFFERY S. (2011), *Data Mining and Statistics for Decision Making*, Wiley.