ClustInvest : outils pour l'analyse et l'interprétation de classifications non supervisées avec variables binaires. Application à des données d'accidents de décompression

Gérard GRÉGOIRE[a] and **Jean-Pierre IMBERT**[b]

[a]Laboratoire LJK
Université Grenoble Alpes
Tour IRMA 51 Rue des mathématiques
Campus de Saint Martin d'Hères
BP 53 38041 Grenoble cedex 09
gerard.gregoire@imag.fr

[b]Divetech
1543 Chemin des vignasses. 06410 BIOT
jpi.divetech@gmail.com

Mots clefs : Classification nonsupervisée. Variables binaires. Accidents de décompression. DCI.

Nous proposons des outils rassemblés dans un package R, ClustInvest, pour faciliter l'analyse et l'interprétation des résultats d'une classification nonsupervisée dans le cas de variables binaires. Après avoir présenté les fonctionnalités du package, nous l'utilisons pour étudier les résultats de classifications effectuées sur des données d'accidents de décompression en plongée professionnelle (DCI).
Dans le cas des données DCI, les variables sont des symptômes manifestés ou pas par l'accidenté (douleurs articulaires, troubles de l'équilibre, nausées, paresthésie-hypoesthésie, fatigue, douleurs dorsales, etc). Pour faciliter la lecture, nous identifions dans ce texte, les variables aux symptômes ($X_j = 1$ si le symptôme est présent, 0 sinon). Il est intéressant, en particulier lorsque le nombre de symptômes manifestés par un individu est plutôt modéré, de disposer de statistiques sur le nombre de symptômes par individu, de statistiques sur le nombre d'occurrences isolées d'un symptôme, le nombre d'occurrences conjointes de 2 symptômes, de 3 symptômes,etc. Il est aussi utile d'identifier les symptômes ayant tendance à apparaître conjointement. Ces informations participent à l'interprétation des groupes. Une aide importante aussi pour l'interprétation d'une classification est le fait de disposer de tableaux permettant de visualiser l'importance de tel ou tel symptôme dans la composition d'un groupe, de visualiser comment les manifestations d'un symptôme se distribuent entre les groupes. Il peut aussi être pertinent de mesurer le caractère plus ou moins discriminant d'un symptôme ou d'un couple de symptômes. Ces fonctionnalités et d'autres sont présentes dans ClustInvest.

ClustInvest(x,C,dist=euclidian)
Utilise les librairies "cluster", "ade4".

- Arguments.
  x: matrice $n \times p$ (ou data.frame) des observations. Les colonnes sont supposées nommées.
  C : vecteur de dimension n contenant la classification à analyser
  dist : distance euclidienne par défaut, sinon matrice de distances ou de dissimilarités.

- Sorties

- stat :
  nombre d'individus, nombre de variables, vecteur du nombre de manifestations de chaque symptôme
- mult :
  tables de multiplicité
  table des couples
  tables des triplets
- influence :
  tables des symptômes observés dans chaque groupe,
  table des pourcentages en ligne, tables des pourcentages en colonnes pour les individus, pour les symptômes.
  Mesure de la qualité discriminante de chaque symptôme par un indice d'entropie et par un indice de variance.
  Valeur de l'indice silhouette par groupe et globale.
- groupes.
  Pour chaque groupe :
  table des occurrences par nom
  table d'occurrences par couples, d'occurrences par triplets,
  tables des pourcentages en lignes et en colonnes.

Le jeu de données DCI contient 785 individus et 17 variables binaires. Tout individu manifeste en réalité au plus 5 symptômes parmi les 17. Le choix de la méthode de classification peut faire débat. En particulier, les méthodes de classification basées sur la distance euclidienne entre individus ne sont pas adaptées à ce contexte. Une des raisons est le fait que,très souvent dans ce type d'observations, 2 individus sont plus proches lorsqu'ils manifestent tous les deux un même symptôme que lorsqu'aucun des 2 ne manifeste ce symptôme. La distance euclidienne ne permet pas de prendre en compte cette dissymétrie. Pour ces raisons nous avons choisi d'utiliser des algorithmes (pam, CAH) avec dissimilarités prenant en compte la dissymétrie (dissimilarités associées aux indices de Jaccard, Sokal et Sneath, Dice et Sorensen, Ochiai. Voir dist.binary du package ade4.). Nous mettons à profit ClustInvest pour interpréter les classes obtenues par ces méthodes. Les informations fournies confirment certains schémas physiologiques généralement acceptés pour le phénomène des bulles de décompression, mais conduisent aussi à certaines interrogations.

Références

[1] J.P. Imbert et al. (2014) Analysis of 605 commercial diving DCS LOGS : trends end underlying mechanisms. 40th Annual Meeting of the Underwater Baromedical Society (EUBS). Wiesbaden, September 24-27th 2014.
[2] Tamer Ozyigit et al. (2010). Decompression Illness Medically Reported by Hyperbaric Treatment Facilities: Cluster Analysis of 1929 Cases. Aviation, Space, and Environmental Medicine x Vol. 81, No. 1.
[3] Holmes Finch (2005). Comparison of Distance Measures in Cluster Analysis with Dichotomous Data. Journal of Data Science 3, 85-100.