Classification d'accidents de décompression en plongée professionnelle

Gérard GRÉGOIRE [1] & Jean-Pierre IMBERT [2]

[1] *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
[2] *Divetech 1543 Chemin des vignasses. 06410 BIOT jpi.divetech@gmail.com*

Résumé. Nous nous intéressons à des données d'accidents de décompression en plongée professionnelle. Le jeu de données comprend 785 individus et 17 variables binaires. Les variables sont des indicatrices de symptômes manifestés ou pas par l'accidenté. Le but de l'étude est d'identifier, à l'aide des symptômes manifestés, des groupes suggérant une interprétation physiologique de l'accident observé. Nous utilisons des méthodes de classification basées sur des dissimilarités prenant en compte la dissymétrie. Utilisant un package R que nous avons développé, nous analysons finement la structure du fichier ainsi que les classifications obtenues de manière à en faciliter l'interprétation. Nous calculons aussi un indice de qualité de la classification basé sur l'entropie. Les informations fournies ont tendance à confirmer certains schémas physiologiques généralement acceptés dans l'étude des phénomènes de bulles de décompression.

Mots-clés. Classification nonsupervisée. Variables binaires. Accidents de décompression. DCI.

Abstract. We are interested in decompression illness concerning professional diving activity. Our data set contains 785 observations and 17 binary variables. Each variable X_i is related to one symptom and $X_i = 1$ or 0 as the symptom is present or not. We are to identify groups which could help to undersatnd the illness. We use classification methods based on dissimilarities. Using an R package we developped we perform a thorough analysis of the data set as well as of the classifications we got. To compare classifications we compute a quality index based on entropy. Globally the results we get are not very far from what physicians specialists of DCI think.

Keywords. Clustering. Unsupervised clustering. Decompression illness. DCI.

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 mÃªme 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.). Considérons la table:

	Indiv2	
Indiv1	1	0
1	a	b
0	c	d

a est le nombre de variables égales à 1 pour les 2 individus, b le nombre de variables égales à 1 pour le 1èr individu et 0 pour l'autre, etc. On a $a+b+c+d=p$ où p est le nombre total de variables.
Les dissimilarités $d(i,j)$ sont souvent généralement définies à partir de similarités $s(i.j)$.

- Dissimilarité de Jaccard:

$$s(i,j)=\frac{a}{a+b+c},\quad d(i,j)=\frac{b+c}{a+b+c}$$

- Dissimilarité de Sokal-Sneath:

$$s(i,j)=\frac{a}{a+b+c+d},\quad d(i,j)=\frac{b+c+d}{a+b+c+d}$$

- Dissimilarité de Dice or Sorensen:

$$s(i,j)=\frac{2a}{2a+b+c},\quad d(i,j)=\frac{b+c}{2a+b+c}$$

- Dissimilarité de Ochiai:

$$s(i,j)=\frac{a}{\sqrt{(a+b)(a+c)}},\quad d(i,j)=1-\frac{a}{\sqrt{(a+b)(a+c)}}$$

Ces dissimilarités donnent plus d'importance aux co-occurrences de "1" et minimisent l'impact des co-occurrences de 0. Beaucoup d'autres dissimilarités sont définies dans la litérature. Nous utilisons ces dissimilarités en particulier avec les méthodes de classification PAM et CAH.

Nous mettons à profit le package ClustInvest pour interpréter les classes obtenues par ces méthodes.

Les fonctions de ClustInvest

- **Analyse du fichier de données**
  a) *Informations de base*
  Nombre d'individus, nombre de variables
  Nombre total de symptômes exprimés
  Nombre maximal de symptômes pour un individu
  Nombre total de patterns différents
  Vecteur du nombre de manifestations de chaque symptôme
  Entropie de l'ensemble des variables
  b) *Multiplicités*
  Tables de multiplicités
  Tables des couples
  Table des triples

- **Analyse d'une classification**
  a) Table des effectifs, tables des pourcentages
  b) Exploration des groupes
  c) Calcul de l'entropie conditionnelle. Calcul de l'indice entropique de la classification. Calcul d'indices d'entropie pour les variables.

- **Application**
  Comparaison de résultats de méthodes différentes.

Nous donnons ci-dessous un tableau d'analyse de la qualité de différentes méthodes à l'aide de l'indice d'entropie que nous définissons:

Méthode	**Nbclasses** $= 4$	**Nbclasses** $= 5$
kmeans	0.5430	0.5882
CAH euclidienne	0.5876	0.6424
CAH diss. meth. 1	**0.5989**	0.6185
CAH diss. meth. 3	0.5831	0.6139
CAH diss. meth. 5	0.5935	0.6459
CAH diss. meth. 7	0.5942	**0.6465**
pam euclidien	0.5366	0.5713
pam diss. meth. 1	0.5366	0.5760
pam diss. meth. 3	0.5366	0.5766
pam diss. meth. 5	0.5366	0.5710
pam diss. meth. 7	0.5366	0.5728
mélange binaire	0.5043(*)	0.5556 (*)
mélange gaussien	0.5413	0.5752

(*) en moyenne sur 40 essais.

Le résultat de notre analyse conduit dans la plupart des classifications à observer:

- Un groupe "problèmes dans les articulations" contenant tous les individus pour lesquels "Joint" est le seul symptôme, plus éventuellement des individus qui ont des symptômes additionnels. Gravité faible de l'événement. (Groupe 1)
- Une classe "vestibulaire (troubles de l'oreille interne)-cérébral". Les manifestations sont principalement Troubles de l'équilibre/Nausée-vomissements/Mouvements des yeux (nystagmus)/troubles de l'audition/fatigue auxquels s'ajoutent éventuellement état de choc/dyspnée/douleurs dans les articulations. Gravité élevée. (groupe 2)
- Une classe "médullaire-neurologique" caractérisée par Fatigue/troubles de la vision/problèmes de dos/Difficulté à se tenir debout(paralysie)/Skin marmorata/état de choc/parèsthésie. Gravité élevée. (Groupe 3)
- Une classe constituée essentiellement d'individus présentant les 2 symptômes "douleurs dans les articulations" et "parésthésie". Ces symptômes semblent atribuables à des atteintes nerveuses locales ou médullaires légères. Gravité modérée. (Groupe 4)

1 Bibliographie

[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*, Vol. 81, No. 1, 1-7.
[3] Holmes Finch (2005). Comparison of Distance Measures in Cluster Analysis with Dichotomous Data. *Journal of Data Science* 3, 85-100.
[4]G. Grégoire and J.P. Imbert. (2015). ClustInvest. Classification nonsupervisée avec données binaires. Applications à des données d'accidents de décompression. Rencontres R. Grenoble 24-25 Juin.