Classification automatique

Christophe Ambroise

1 Introduction

1.1 Objectif

- ✔ obtenir une représentation simplifiée des données
  - ✓ résumé
  - ✓ compression (avec perte)
- ✔ vérifier une structure existante,
- ✔ ...

FIG. 1 – Compression de Sir R. Fisher

1.2 Terminologie

- ✔ en sciences naturelles : taxinomie (ou taxonomie) désigne l'art ou la science de la classification
- ✔ en médecine : la nosologie est la classification des maladies
- ✔ en marketing : typologie
- ✔ en reconnaissance des formes : classification non supervisée ou classification sans professeur

1.3 Définition

Dans le petit Larousse,

Distribution par classe selon un certain ordre et une certaine méthode

1.4 Questions

Une définition formelle de la classification, qui puisse servir de base à l'automatisation du processus, amène les questions suivantes

- ✔ Comment les objets à classer sont-ils définis ?
- ✔ Comment définir la notion de ressemblance entre objets ?
- ✔ Qu'est-ce qu'une classe ?
- ✔ Comment les classes sont-elles structurées ?
- ✔ Comment comparer deux classifications ?

1.5 Notations et définitions

- ✔ Données : $\Omega = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^t$
- ✔ Ressemblance entres deux objets
  - ✓ Démarche monothétique : deux objets sont semblables s'ils partagent une certaine caractéristique.

    Les mammifères allaitent leurs petits
  - ✓ Démarche polythétique deux objets sont semblables, s'ils sont proches au sens d'une mesure de proximité (i.e. distance ou dissimilarité).

$$d(A, B) = \sqrt{(nicotineA - nicotineB)^2 + (goudronA - goudronB)^2}$$

Cigarettes	Nicotine (mg)	Goudron (mg)
Royal anis	0.45	4.9
Rothmans	1.1	14
Chesterfield Lights	0.6	8
Benson & Hedges	1.1	13
Peter Stuyvesant	1	12.7
Gitanes	1	12
Malboro	1	14
Lucky Strike	0.9	14

1.6 Structure de classification

- ✔ partitions
- ✔ hiérarchies
- ✔ classes empiétantes
- ✔ classe floues
- ✔ ...

2 Structures de classification

2.1 Partition

Ω est un ensemble fini. $P = (P_1, P_2, ..., P_K)$ un ensemble de parties non vides de Ω est une partition si et seulement si :

1. $\forall i \neq j$, $P_i \cap P_j = \emptyset$,
2. $\cup_i P_i = \Omega$.

Exemple

- $P_1 = \{RoyalAnis, Chesterfield\}$,
- $P_2 = \{Benson, Malbor, Gitane, Lucky, Peter, Rothman\}$.

2.2 Notation pratique

Soit $\Omega = (\mathbf{x}_1, ..., \mathbf{x}_n)$ un ensemble partitionné en K classes, Cette partition peut être décrite par une matrice de classification :

$$\mathbf{c}(P) = \mathbf{c} = \begin{pmatrix} c_{11} & \cdots & c_{1K} \\ \vdots & \ddots & \vdots \\ c_{n1} & \cdots & c_{nK} \end{pmatrix}$$

où $c_{ik} = 1$ ssi $\mathbf{x}_i \in P_k$, et $c_{ik} = 0$ sinon.

- $\sum_{k=1}^{K} c_{ik} = 1$
- $\sum_{i=I}^{n} c_{ik} = n_k > 0$

2.3 Partition floue

Extension des travaux de Zadeh (1965) sur les ensembles flous à la classification

- ✔ Un individu appartient à toutes les classes avec différents degrés d'appartenance

une partition floue est définie par une matrice de classification floue vérifiant

1. $\forall k \in \{1, \ldots, K\},\ \forall \mathbf{x}_i \in \Omega,\ c_{ik} \in [0, 1]$
2. $\forall k \in \{1, \ldots, K\},\ 0 < \sum_{i=1}^{n} c_{ik} < n$
3. $\forall \mathbf{x}_i \in \Omega,\ \sum_{k=1}^{K} c_{ik} = 1$

3 Hiérarchie indicée

Ω est un ensemble fini. Un ensemble H de parties non-vides de Ω est une hiérarche sur Ω si

- $\Omega \in H$
- $\forall \mathbf{x} \in \Omega, \{\mathbf{x}\} \in H$
- $\forall h,\ h' \in H,\ h \cap h' = \emptyset\ ou\ h' \subset h\ ou\ h \subset h'$

3.1 Indice

On appelle indice sur une hièrarchie H une fonction i de H dans $\mathbb{R}^+$ vérifiant les propriétés

– $h \subset h' \Rightarrow i(h) < i(h')$

– $\forall \mathbf{x} \in \Omega,\ i(\{\mathbf{x}\}) = 0$

Le couple (h, i) est alors appelé hiérarchie indicée.

3.2 Partition et hiérarchie

- ✔ à chaque niveau d'une hiérarchie indicée correspond une partition
- ✔ $\{\Omega, P1, P2, \ldots, P_K, x_1, .., x_n\}$ forme une hiérarchie

3.3 Aspects combinatoires

Le nombre de hiérarchies et partitions possibles devient vite explosif lorsque le cardinal de Ω augmente :

Exemple

- ✔ n éléments
- ✔ K classes
- ✔ $S(n,K) = \frac{1}{K!}\sum_{k=0}^{K}(-1)^{k-1}C_k^K k^n$ partition possible

4 Objectif de la classification

- ✔ Idéalement, 2 objets d'une classe donnée devrait être plus proches que deux objets originaires de classes différentes.

  $\Rightarrow$ objectif impossible à atteindre

- ✔ démarche numérique
  - ✓ définition d'un critère
  - ✓ impossible de trouver l'optimum global
  - ✓ optimisation locale