MixAll: Un logiciel de classification non-supervisée pour données mixtes avec des valeurs manquantes [1]

S. Iovleff [a]

[a]Université Lille 1 (UMR 8524), Inria Lille Nord Europe (Modal Team)
59650 Villeneuve d'Ascq - France
serge.iovleff@inria.fr

[1]Travail en partie financé par le défi Mastodons

Mots clefs : Classification non supervisée, Modèles de mélange, Données mixtes, C++.

Introduction

Le package MixAll ([3]) permet de faire de la classification des données mixtes en proposant des modèles de mélanges Gaussiens, gamma, poisson et catégoriel. Il permet aussi de prendre en compte les données manquantes. Les traitements sont réalisés en utilisant le code C++ du projet "Clustering" de la librairie STK++ (The Statistical ToolKit, `http://www.stkpp.org`). STK++ est une librairie écrite en C++ dédiée aux statistiques. La librairie est divisée en différent projets. Le noyau est distribué à travers le package rtkore ([1],[2]) dont dépend donc le package MixAll.

Présentation rapide des modèles de mélange

Soit $\mathcal{X}$ un espace mesurable arbitraire et soit $\mathbf{x} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ n vecteurs indépendants de $\mathcal{X}$ tels que $\mathbf{x}_i$ est généré par un *modèle de mélange* de densité

$$f(\mathbf{x}_i|\theta) = \sum_{k=1}^{K} p_k h(\mathbf{x}_i|\boldsymbol{\lambda}_k, \boldsymbol{\alpha}) \tag{1}$$

où les p_k's représentent les proportions de mélange et $h(\cdot|\boldsymbol{\lambda}_k, \boldsymbol{\alpha})$ représente une densité de probabilité $\mathcal{X}$ paramétrée par $\boldsymbol{\lambda}_k$ et $\boldsymbol{\alpha}$. Le vecteur des paramètres à estimer est noté $\theta = (p_1, \ldots, p_K, \boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_K, \boldsymbol{\alpha})$ et son estimateur est obtenu en maximisant la log-vraisemblance observée

$$L(\theta|\mathbf{x}_1, \ldots, \mathbf{x}_n) = \sum_{i=1}^{n} \ln \left(\sum_{k=1}^{K} p_k h(\mathbf{x}_i|\boldsymbol{\lambda}_k, \boldsymbol{\alpha}) \right). \tag{2}$$

Le cas des données mixtes est traité en faisant une hypothèse d'indépendance conditionnelle, c'est à dire en supposant que $\mathcal{X} = \mathcal{X}^1 \times \ldots \times \mathcal{X}^L$ et que conditionnellement à la classe

$$h(\mathbf{x}_i|\boldsymbol{\lambda}_k, \boldsymbol{\alpha}) = h^1(\mathbf{x}_i^1|\boldsymbol{\lambda}_k^1, \boldsymbol{\alpha}^1) \ldots h^1(\mathbf{x}_i^L|\boldsymbol{\lambda}_k^L, \boldsymbol{\alpha}^L).$$

Lorsque certains échantillons présentent des valeurs manquantes, c'est à dire lorsque certains $\mathbf{x}_i$ se décompose en un vecteur $\mathbf{x}_i^o$ de valeurs observées et un vecteur $\mathbf{x}_i^m$ de valeurs manquantes, il faut en théorie maximiser la log-vraisemblance intégrée

$$L(\theta|\mathbf{x}_1^o, \ldots, \mathbf{x}_n^o) = \sum_{i=1}^{n} \int \ln \left(\sum_{k=1}^{K} p_k h(\mathbf{x}_i^o, \mathbf{x}_i^m|\boldsymbol{\lambda}_k, \boldsymbol{\alpha}) \right) d\mathbf{x}_i^m. \tag{3}$$

Estimation des modèles de mélange à l'aide de MixAll

MixAll permet d'estimer les modèles de mélange suivants

- les mélanges gaussiens diagonaux (8 modèles) en utilisant la fonction `clusterDiagGaussian`,
- les mélanges gamma (24 modèles) en utilisant la fonction `clusterGamma`,
- les mélanges catégoriels (4 modèles) en utilisant la fonction `clusterCategorical`,
- les mélanges de Poisson (6 modèles) en utilisant la fonction `clusterPoisson`,
- un modèle spécial appelé "mixed data" en utilisant la fonction `clusterMixed`.

L'estimation peut se faire en utilisant l'algorithme EM et ses variantes (CEM, SEM, SemiSEM) et le comportement de ces algorithmes peut être ajusté par l'utilisateur en utilisant la fonction `clusterAlgo`. Afin d'éviter de tomber dans un minimum local il est possible de définir une stratégie d'estimation en utilisant la fonction `clusterStrategy`. Une stratégie consiste à lancer un petit nombre d'itérations d'un algorithme depuis différentes configurations choisies au hasard, puis à choisir parmi les résultats obtenus la classification la plus prometteuse avant de lancer un algorithme d'estimation à partir de cette configuration.
Lorsque l'échantillon présente des valeurs manquantes MixALL propose deux algorithmes d'estimation adaptés : le `SEM` et le `SemiSEM`. La quantité (3) est alors estimée en utilisant une méthode de Monte-Carlo au cours des itérations.

Perspectives
Le package MixAll a vocation a être étendu à tous les types de données. Dans le cadre du projet CloHé financé par le défi Mastodons, il pourra traiter les données fonctionnelles à l'aide de modèles génératifs et par des méthodes à noyaux.

Références
[1] Serge Iovleff (2015). Rtkpp: Un package pour faire l'interface entre R et la bibliothèque STK++. *Quatrièmes Rencontres R.*
[2] Serge Iovleff (2015). rtkore: STK++ Core Library Integration to R using Rcpp. *http://cran.r-project.org/web/packages/rtkore/*
[3] Serge Iovleff (2015). MixAll: Clustering using Mixture Models. *http://cran.r-project.org/web/packages/*