Пример использования функции нечеткой кластеризации fanny в программе R

В. К. Солондаев, кандидат психологических наук,
доцент кафедры общей психологии ЯрГУ им. П. Г. Демидова

Ссылка для цитирования

Солондаев В.К. Пример использования функции нечеткой кластеризации fanny в программе R [Электронный ресурс] URL: http://cafedra.narod.ru/solondaev/solond-R-fanny.pdf (дата обращения: чч.мм.гггг).

Во многих исследованиях возникает задача выделения кластеров — групп похожих между собой результатов наблюдений. Современная статистика позволяет решить эту задачу без использования каких-либо содержательных предположений о природе исследуемой реальности и стоящих за результатами закономерностях. Один из инструментов такого рода кластеризации - функция fanny ['fænɪ]свободного статистического пакета `cluster`[1], входящего в состав статистического языка и среды для статистических вычислений `R`[2]. Функция обеспечивает нечеткий кластерный анализ, т. е. позволяет количественно оценить меру принадлежности объекта к каждой из выделяемых групп. Такая возможность позволяет формализовать широко используемое содержательное допущение о возможности существования промежуточных типов. Термин «нечеткий» используется здесь в контексте теории нечетких множеств, общее изложение которой можно найти в работе А. Кофмана[3] и многочисленных публикациях, обзор которых не входит в задачи данной статьи.

При использовании в текстах названия функции следует учитывать, что обозначающее функцию слово fanny ['fænɪ], судя по определениям англо-русского словаря общей лексики и Cambrige Dictionaries Online, многозначно. Так англо-русский словарь предлагает следующие значения:

I сущ.; мор.жестянка; банка; бачок Синоним: tin
fanny full of hot tea — полная кружка горячего чая
II сущ.; сниж.
1) амер. зад, задница; мягкое место
2) брит.; груб. женские половые органы
III
1. сущ.; разг.небылица; болтовня
2. гл.; разг.плести небылицы; заговаривать, вводить в заблуждение, обманывать

Приведем пример использования функции fanny, акцентируя техническую, а не содержательную сторону вопроса. Сделаем нечеткий кластерный анализ и сохраним графические файлы с результатами .

---

1 Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.(2011). cluster: Cluster Analysis Basics and Extensions. R package version 1.14.1.

2 R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

3 Кофман А. Введение в теорию нечетких множеств: Пер. с франц. - М.: Радио и связь, 1982. - 432 с.

Если пакет cluster не установлен, устанавливаем его командой

```
install.packages("cluster", dependencies=TRUE)
```

В данном случае `dependencies=TRUE` означает автоматическую установку связанных пакетов. В многопользовательских операционных системах могут возникать сложности с установкой пакетов для всех пользователей. Тогда установка с правами администратора системы — наиболее очевидный путь решения[4]. Однако в разных операционных системах может использоваться разная политика разделения прав пользователей, но предмет нашего интереса — кластерный анализ.

Описываемые ниже команды в текстовом формате можно сохранить как скрипт в рабочий каталог R и использовать команду для запуска скрипта

```
source("имя скрипта.r", echo=TRUE)
```

Для проведения кластеризации после запуска программы R загружаем пакет cluster

```
library(cluster)
```

В описываемом примере автором использован объект z, содержащий набор оценок различных вариантов решения задач. Данные были получены нами в ходе исследования практического мышления[5].

Поскольку функция fanny требует задать количество кластеров, вначале выделяем 4 кластера.

```
z.f<-fanny(z, 4)
```

Далее создаем графический файл в формате tiff с результатами кластеризации. Выбор формата определяется возможностью изменения размера файла без существенной потери качества. Часто удобным оказывается прозрачный фон. В презентациях он позволяет с минимальными усилиями включить графические объекты в цветовую схему слайда. Прозрачный фон задан командой bg = "transparent" ; в ubuntu 10.04 используем type = "cairo" .

Заметим, что, судя по автоматической подписи (рисунки 1-3), программа R по умолчанию отображает кластеризацию в пространстве двух первых главных компонент.

Графический файл создается следующей командой

```
tiff(file = "fanny14.tiff", width = 500, height = 500, units =
"px", pointsize = 12, bg = "transparent", type = "cairo")
```

Как выяснилось, для рисования важно использовать clusplot , чтобы в файле отобразилась именно кластеризация, а не «силуэт» (silhouette plot). Отображаем в файл результаты кластеризации

```
clusplot(z.f)
```

Далее закрываем графическое устройство, после чего программа R сохраняет созданный

---

4 Автор признателен Роману Петрухину за важное уточнение этого момента.

5 Солондаев В. К. Выбор действий субъекта в комплексной ситуации // Рефлексивные процессы и управление. Сборник материалов VIII Международного симпозиума 18-19 октября 2011 г. — М: «Когито-Центр», 2011.С. 240-242

файл в рабочий каталог (рисунок 1).

Для сравнения пробуем варианты с разным числом кластеров (2, 4, 6) . Графический вывод приведен на рисунках 2, 1, 3 соответственно.

Сравнивая рисунки 1-3 видим, что увеличение числа кластеров с 2 не улучшает кластеризацию.

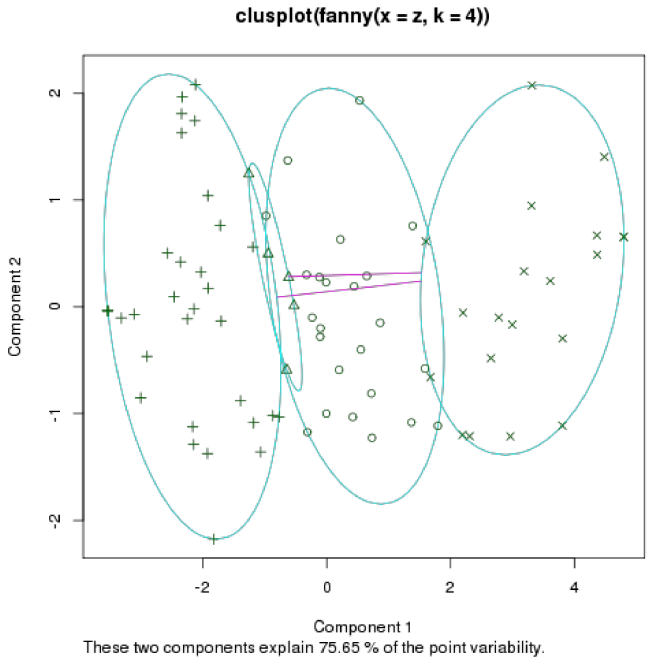

Рисунок 1: Нечеткая кластеризация с выделением четырех кластеров.

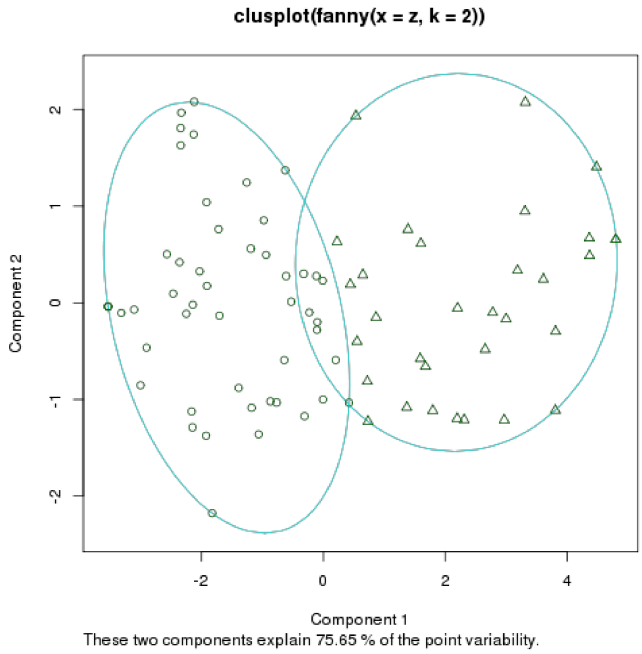

Рисунок 2: Нечеткая кластеризация с выделением двух кластеров.

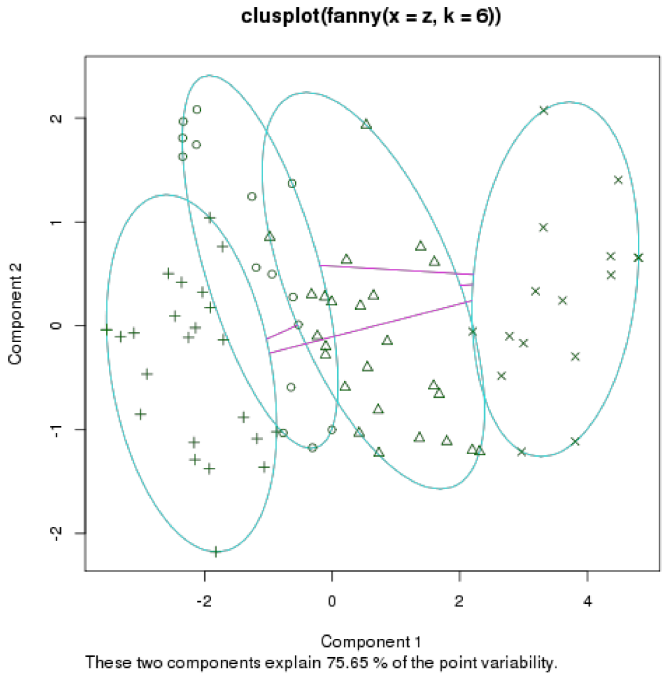

Рисунок 3: Нечеткая кластеризация с выделением шести кластеров.

Вывод о бесперспективности увеличения числа кластеров в описанном случае имеет содержательную интерпретацию. Поскольку оцениваемые решения бывают правильными и неправильными, использование k>2 вполне закономерно не улучшает кластеризацию . Сделать такой вывод мы можем и по числовым результатам кластеризации. Для этого изучим объекты, создаваемые функцией fanny при разном значении параметра k.

Создадим объект fanny при k=2 и проанализируем его содержание

```
> z2.f<-fanny(z, 2)
> z2.f
Fuzzy Clustering object of class 'fanny' :
m.ship.expon.        2
objective     120.4843
tolerance        1e-15
iterations          31
converged            1
maxit              500
n                   81
Membership coefficients (in %, rounded):
      [,1] [,2]
 [1,]   58   42
 [2,]   66   34
 [3,]   53   47
 [4,]   63   37
 [5,]   63   37
 [6,]   52   48
 [7,]   62   38
 [8,]   50   50
...
```

Затем повторим действия при k=4.

```
> z4.f<-fanny(z, 4)
> z4.f
Fuzzy Clustering object of class 'fanny' :
m.ship.expon.        2
objective     59.24158
tolerance        1e-15
iterations         150
converged            1
maxit              500
n                   81
Membership coefficients (in %, rounded):
      [,1] [,2] [,3] [,4]
 [1,]   31   31   26   11
 [2,]   32   32   29    8
 [3,]   31   31   25   13
 [4,]   29   29   30   11
 [5,]   31   31   29   10
 [6,]   32   32   24   11
 [7,]   29   29   30   12
 [8,]   32   32   24   12
```

Сравнивая два объекта, мы видим, что коэффициенты, характеризующие отношения принадлежности v (определение v приведено ниже) объекта 1 к первому кластеру v(1, 1) и ко второму кластеру v(1, 2) отличаются при k=2 и совпадают при k=4. Подобные результаты показывают пересечение кластеров. В реальности редко какое решение бывает полностью верным или полностью ошибочным, но различение «полностью верного», «менее верного», «частично ошибочного», «полностью ошибочного» вариантов решения (возможная интерпретация 4 кластеров) не улучшает ситуацию, поскольку заметное количество объектов в равной степени принадлежат одновременно к нескольким нечетким подмножествам решений разной правильности. Аналогичные изменения отношений принадлежности мы видим и по другим объектам. Это позволяет даже по кластеризации первых восьми объектов предпочесть минимально возможное k=2.

Мы намеренно рассмотрели пример с данными, плохо поддающимися кластеризации. Этот результат интересен для нас в содержательном плане. Оцениваемые объекты — три варианта решения трех разных задач. Разные варианты решения значительно различаются по объективным параметрам. Результаты кластеризации показывают, что испытуемые в своих оценках учитывают не только эти параметры.

Кластеризованы оценки трех вариантов решения трех разных задач (по 9 оценок для каждого испытуемого) по общему набору шкал. Общий набор шкал позволяет объединить результаты разных испытуемых и оценки разных решений, разных задач. Если испытуемые в своих оценках полностью учитывают предметную специфику задач, мы должны получить три

четко различающихся кластера. Если испытуемые полностью учитывают различия решений задач, мы также должны получить три четко различающихся кластера. Если испытуемые полностью учитывают и предметную специфику задач, и различия решений, мы должны получить кратное трем число кластеров.

Но, как показано выше, четыре и шесть кластеров не дают более точного разделения, чем два кластера. А при попытке выделения трех кластеров мы получаем сообщение об ошибке

```
> zf3<-fanny(z, 3)
Предупреждение
In fanny(z, 3) :
  FANNY algorithm has not converged in 'maxit' = 500 iterations
```

Поэтому мы можем интерпретировать результаты кластеризации как проявление влияния некоторого фактора (в нематематическом смысле), независимого от содержания задач и различий вариантов решения.

Аналогичная интерпретация была получена нами ранее с помощью других статистических методов. Но функция fanny дает один интересный результат — значение коэффициента принадлежности, который позволяет не только оценить «силу тяготения» объекта к той или иной группе (кластеру). Внимательное изучение связей коэффициентов принадлежности разных объектов к одному кластеру со значениями переменных, по которым происходила кластеризация, может дать содержательную информацию об основаниях, лежащих в основе выделения данных кластеров.

Для понимания предметного основания, отражаемого выделенными по статистическим основаниям кластерами, можно экспортировать эти коэффициенты (в выводе функции они названы `Membership coefficients`). К сожалению, нам удалось экспортировать коэффициенты лишь через буфер обмена. Команда `write.table` выдала ошибку:

```
Ошибка в as.data.frame.default(x[[i]], optional = TRUE,
stringsAsFactors = stringsAsFactors) :
  cannot coerce class 'c("fanny", "partition")' into a data.frame
```

Экспортировав коэффициенты принадлежности, в рассматриваемом случае мы объединили информацию с данными о выборе для реализации решений, оцениваемых в проведенном исследовании. Поскольку выше уже было показано, что выделение более двух кластеров не увеличивает точность кластеризации, коэффициенты были взяты из объекта z2.f для двух кластеров.

Далее, по аналогии с созданием графических объектов R, приведенных на рисунках 1-3, мы построили боксплоты коэффициентов принадлежности для каждого кластера в зависимости от выбора решения. Результаты приведены на рисунках 4-5.

```
boxplot(z2$V2 ~ z2$V3, xlab="решения", ylab="коэфф. принадлежности
к кластеру 1 (в %, округленно)", names=c("не выбранные",
"выбранные"))
```

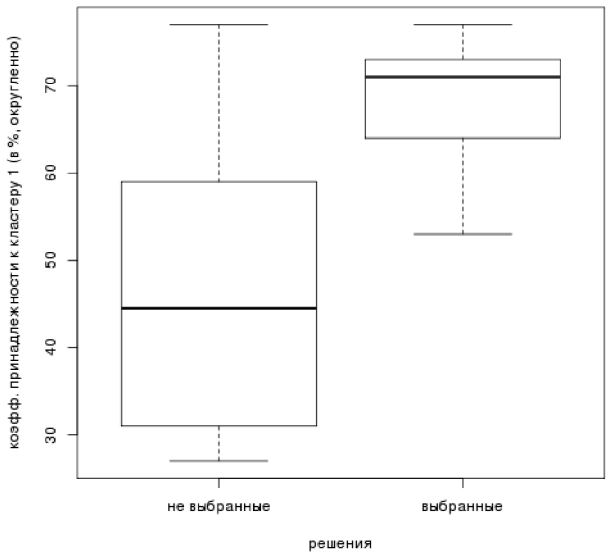

Рисунок 4: Боксплоты коэффициентов принадлежности к кластеру 1 в зависимости от выбора решения.

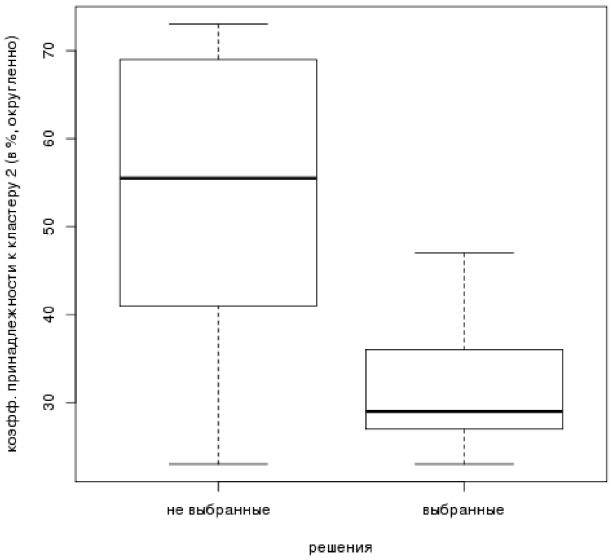

Рисунок 5: Боксплоты коэффициентов принадлежности к кластеру 2 в зависимости от выбора решения.

Нас не интересует статистическая значимость приведенных результатов, поскольку проверка статистических гипотез логически противоречит основной идее нечеткой кластеризации. Если по формальным причинам при выполнении научно-квалификационных работ такая проверка потребуется, ее можно провести.

Вернемся к содержательной интерпретации результатов нечеткой кластеризации.

На рисунке 2 отчетливо заметна область пересечения двух кластеров. Содержательно это могло бы интерпретироваться так: для испытуемых все оцениваемые варианты решения делятся на правильные, неправильные и промежуточную группу.

Очевидно, что врач не выбирает тот вариант решения (вариант лечения), который считает неправильным. Не менее очевидно, что однозначно оценить правильность варианта решения может быть, мягко говоря, затруднительно. Медицинская задача часто является логически нестрогой: эффективное при конкретном *заболевании* лекарство может иметь ряд противопоказаний, существенных для конкретного *больного*. Поэтому закономерно, что мы редко получаем однозначные оценки.

Рисунки 1-3 явно могут быть «прочитаны» как диаграммы диаграммами Эйлера-Венна, но аналогично могут интерпретироваться и рисунки 4-5. Нас интересует предметная интерпретация основания кластеризации, произведенной «механически» - без учета предметного значения результатов, количественно отраженных в массиве данных.

Предположим, что все оцениваемые решения делятся для наших испытуемых на два нечетких множества правильных и неправильных, причем испытуемые выбирают те варианты, которые считают правильными. В каждом отдельном случае мы не увидим никакой нечеткости. Один испытуемый выберет единственный вариант решения и отвергнет остальные. Чтобы «увидеть» нечеткость, нам необходимо много раз повторить исследование с одним испытуемым, или по одному разу провести со многими испытуемыми. Тогда мы обнаружим, что выбор испытуемых не определяется полностью ни особенностями самих испытуемых, ни их оценками вариантов решения. Наш пример иллюстрирует последний вариант. Разделение решений на два кластера (по существу — два множества), интерпретируемых нами как правильные и неправильные, происходило лишь по шкальным оценкам вариантов, без учета их выбора. В интерпретации мы исходим из допущения незлонамеренности врача, т. е. из того, что врач всегда выбирает только те решения, которые считает правильными. Незлонамеренность вплне допускает, что по разным причинам врач не всегда выбирает «наилучшие» или «самые правильные» решения. На рисунках 4 и 5 мы для выбранных решений видим противоположную картину. Рисунок 4 показывает, что испытуемые не выбирают решения, коэффициент принадлежности которых к кластеру правильных меньше 50% (медиана выше 70%), а рисунок 5 показывает, что испытуемые не выбирают решения, коэффициент принадлежности которых к кластеру неправильных больше 50% (медиана ниже 30%). Заметим, что распределение коэффициентов принадлежности не выбранных решений на рисунках 4 и 5 явно перекрывает диапазон коэффициентов принадлежности выбранных решений, т. е. в психологическом плане выбранные решения не являются полной противоположностью не выбранных, что дополнительно подтверждает

содержательную адекватность нечеткой кластеризации.

В заключение статьи приведем более полное описание функции fanny[6]

Функция fanny производит нечеткую кластеризацию данных на k кластеров.

Использование

```
fanny(x, k, diss = inherits(x, "dist"), memb.exp = 2,
      metric = c("euclidean", "manhattan", "SqEuclidean"),
      stand = FALSE, iniMem.p = NULL, cluster.only = FALSE,
      keep.diss = !diss && !cluster.only && n < 100,
      keep.data = !diss && !cluster.only,
      maxit = 500, tol = 1e-15, trace.lev = 0)
```

Аргументы функции fanny

`x` - объект обработки - матрица или фрейм (таблица -?) данных, или матрица различий, в зависимости от значения аргумента `diss`

В случае матрицы или таблицы данных, каждая строка соответствует наблюдению, и каждый столбец соответствует переменной. Все переменные должны быть числовыми.

Отсутствующие значения (`NAs`) не допускаются.

В случае матрицы различий, `x`, как правило, выход функции `daisy` или `dist`. Также допускается вектор длины n*(n-1)/2 (где n-число наблюдений), который может интерпретироваться как выход вышеперечисленных функций. Отсутствующие значения (NAs) не допускается.

`k` - целое число, задающее требуемое количество кластеров. Должно находиться в диапазоне $0 < k < n/2$ где n число наблюдений.

`diss` - логическое значение (т. н. флаг) если `diss=TRUE` (по умолчанию для объектов - расстояний или различий), то x принимается как матрица различий. Если `diss=FALSE` то `x` рассматривается как матрица наблюдений переменных.

`memb.exp` - число r строго больше, чем 1, задающее основание экспоненты, используемой в критерии соответствия (membership exponent used in the fit criterion); см. «подробности» ниже. Значение по умолчанию равно 2 - фиксированное значение внутри fanny.

`metric` - символьный аргумент, задающий метрики, которые будут использоваться для расчета расстояния между наблюдениями. Варианты значений: `"euclidean"` (по умолчанию), `"manhattan"`, `"SqEuclidean"`. Евклидовы расстояния равны квадратному корню суммы квадратов расстояний, и манхэттенские расстояния раны сумме абсолютных расстояний, квадратичные евклидовы расстояния равны сумме квадратов расстояний. Используя этот последний вариант, получаем эквивалент (немного более медленный) для вычисления так называемых "fuzzy C-means" (нечеткой кластеризации методом К-средних).

---

6 перевод описания из R Documentation выполнен автором

Если x - уже матрица различий, этот аргумент игнорируется.

stand - логическое значение; если stand=TRUE то измерения в x стандартизуются до вычисления различий. Измерения стандартизуются для каждой переменной (столбец), путем вычитания из переменной среднего значения и деления на среднее абсолютное отклонение переменной. Если x - уже матрица различий, этот аргумент игнорируется.

iniMem.p - числовая n * k матрица или NULL (по умолчанию); может использоваться для указания начальной матрицы принадлежности, т.е. матрица неотрицательных чисел, с одинаковой суммой каждой строки.

cluster.only - логическое значение, если cluster.only=TRUE, не вычисляется и не выдается информация о силуэте, см. Подробности.

keep.diss, keep.data - логические значения, указывающие необходимость сохранения в результате функции различий и/или входных данных x. Установка значения FALSE может дать меньше результатов, что экономит выделяемый объем памяти и время.

maxit, tol - максимальное количество итераций, и по умолчанию устойчивость (tolerance) к схождению (convergence) (относительное схождение к критерию соответствия) для алгоритма fanny. По умолчанию фиксированные значения в алгоритме: maxit = 500 и tol = 1e-15 .

trace.lev - целое число, указывающее уровень трассировки для печати диагностики в ходе C-внутреннего (C-internal) алгоритма. По умолчанию равно 0 т. е. не печатать ничего; более высокие значения - печатать все больше и больше.

Подробности

В нечеткой кластеризации каждое наблюдение "рассыпается" по различным кластерам. Обозначим u(i,v) отношение принадлежности наблюдения i к кластеру v.

Эти отношения являются неотрицательными, для каждого наблюдения i их сумма равна 1. В частности, метод fanny вытекает из главы 4 Kaufman and Rousseeuw (1990) (см. руководство к daisy) Martin Maechler расширил метод, для любого, указанного пользователем memb.exp, iniMem.p, maxit.tol, и др.

fanny направлена на минимизацию целевой функции

SUM_[v=1..k] (SUM_(i,j) u(i,v)^r u(j,v)^r d(i,j)) / (2 SUM_j u(j,v)^r)

где n - число наблюдений, k - количество кластеров, r - основание экспоненты (?) memb.exp и d(i,j) - различие между наблюдениями i и j.

Обратите внимание, что r -> 1 дает более четкую кластеризацию, r -> Inf приводит к полной нечеткости. K\&R(1990), на стр.191 указывают, что значения слишком близкие к 1, могут привести к медленной сходимости. Далее заметим, что даже по умолчанию, r = 2 может

привести к полной нечеткости, т.е.u(i,v) == 1/k. В этом случае будет выведено предупреждение, и рекомендация пользователю выбрать меньшее `memb.exp` (=r).

По сравнению с другими нечеткими методами кластеризации, функция fanny имеет следующие особенности:

(a) принимает матрицы различий;

(b) более робастна (устойчива) к допущению сферических кластеров;

(c) предоставляет "силуэт" (silhouette plot) (см. plot.partition).

`Value` - объект класса "fanny", представляющий собой кластеризации. Подробности см. fanny.object.

См. также

Для общего понимания и ссылок рекомендуется изучить функции `agnes;`

`fanny.object, partition.object, plot.partition, daisy, dist.`

Примеры

создадим 10+15 объектов в двух кластерах, плюс 3 "обманчивых" объекта между кластерами.

```
x <- rbind(cbind(rnorm(10, 0, 0.5), rnorm(10, 0, 0.5)),
cbind(rnorm(15, 5, 0.5), rnorm(15, 5, 0.5)),
cbind(rnorm( 3,3.2,0.5), rnorm( 3,3.2,0.5)))
fannyx <- fanny(x, 2)
```

Рассматривая объект `fannyx`, обратите внимание, что наблюдения с 26 по 28 - "нечеткие" (ближе к # 2)

```
fannyx
summary(fannyx)
plot(fannyx)
```

есть вариант увеличения / уменьшения «степени четкости»

```
(fan.x.15 <- fanny(x, 2, memb.exp = 1.5)) # увеличиваем четкость для
```

объектов 26:28

```
(fanny(x, 2, memb.exp = 3)) # большая нечеткость в целом
```

используем набор данных ruspini, состоящий из 75 точек в четырех группах, широко используемый для иллюстрации техники кластеризации.

```
data(ruspini)
f4 <- fanny(ruspini, 4)
stopifnot(rle(f4$clustering)$lengths == c(20,23,17,15))
plot(f4, which = 1)
```

Изображение, похожее на рисунок №6 в книге Stryuf et al (1996)

```
plot(fanny(ruspini, 5))
```