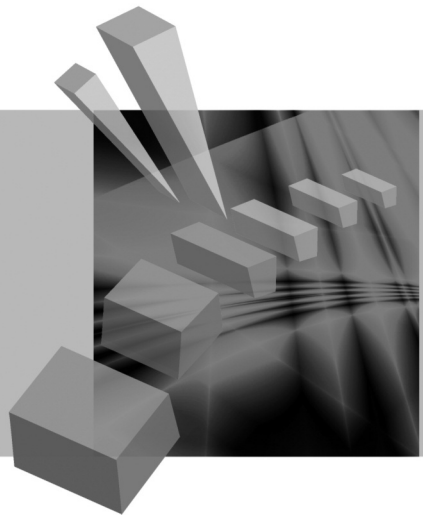
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

242

Taksonomia 19.

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2012

Recenzenci: Eugeniusz Gatnar, Elżbieta Gołata, Tadeusz Kufel, Józef Pociecha,
Mirosław Szreder, Feliks Wysocki

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

Tytuł sfinansowano ze środków Sekcji Klasyfikacji i Analizy Danych PTS
i Uniwersytetu Ekonomicznego we Wrocławiu

Publikacja jest dostępna na stronie www.ibuk.pl

Streszczenia opublikowanych artykułów są dostępne w międzynarodowej bazie danych
The Central European Journal of Social Sciences and Humanities http://cejsh.icm.edu.pl
oraz w The Central and Eastern European Online Library www.ceeol.com,
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon http://kangur.uek.krakow.pl/
bazy_ae/bazekon/nowy/index.php

Informacje o naborze artykułów i zasadach recenzowania znajdują się
na stronie internetowej Wydawnictwa
www.wydawnictwo.ue.wroc.pl

Kopiowanie i powielanie w jakiejkolwiek formie
wymaga pisemnej zgody Wydawcy

© Copyright by Uniwersytet Ekonomiczny we Wrocławiu
Wrocław 2012

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM
Nakład: 320 egz.

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 242
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS 2012

Taksonomia 19. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

Ewa Genge
Uniwersytet Ekonomiczny w Katowicach

ANALIZA SKUPIEŃ OPARTA NA MIESZANKACH UCIĘTYCH ROZKŁADÓW NORMALNYCH

Streszczenie: W artykule przedstawione zostało najnowsze podejście modelowe, polegające na usuwaniu obserwacji oddalonych w trakcie estymacji parametrów mieszanki. W podejściu tym na macierze kowariancji nakładane są pewne ograniczenia, wyznaczane są funkcje wiarygodności uciętych zmiennych losowych, a parametry takiej mieszanki szacowane są za pomocą zmodyfikowanej wersji algorytmu EM, tj. algorytmu TCLUST [García-Escudero i in. 2010, s. 89-109]. Przeprowadzone zostały również badania porównawcze z bardziej popularnym, odpornym modelem mieszanek rozkładów normalnych.

Słowa kluczowe: model mieszanek, algorytm EM, algorytm TCLUST.

1. Wstęp

Analiza skupień oparta na mieszankach rozkładów normalnych nie pozwala na osiągnięcie zadowalających wyników podziału w przypadku, gdy w zbiorze danych występują obserwacje nietypowe. Do tej pory problem ten rozwiązywano dzięki modyfikacji modelu mieszanek, polegającej na dodaniu dodatkowej klasy (rozkładu składowego mieszanki) będącej reprezentantem obserwacji wyraźnie różniących się od obserwacji typowych [Dasgupta, Raftery 1998, s. 294-302].

W artykule przedstawione zostanie najnowsze podejście modelowe, polegające na usuwaniu obserwacji oddalonych w trakcie estymacji parametrów mieszanki. W podejściu tym na macierze kowariancji nakładane są pewne ograniczenia, wyznaczane są funkcje wiarygodności uciętych zmiennych losowych, a parametry takiej mieszanki szacowane są za pomocą zmodyfikowanej wersji algorytmu EM, tj. algorytmu TCLUST [García-Escudero i in. 2008, s. 1324-1345; 2010, s. 89-109]. W artykule dokonane zostanie również porównanie wyników klasyfikacji dla prezentowanego podejścia oraz wspomnianej już modyfikacji modelu mieszanek rozkładów normalnych.

Obliczenia zostaną przeprowadzone m.in. za pomocą pakietów: `tclust`, `clusterSim`, `mclust` oraz `mlbench` programu **R**.

2. Ucinanie obserwacji w podejściu modelowym

Gallegos [2002] oraz Gallegos i Ritter [2005] zaproponowali probabilistyczny model dla obserwacji oddalonych (*spurious outliers model*). Funkcję wiarygodności takiego modelu dla obserwacji $\mathbf{x}_i$ można zapisać jako:

$$L = \left[\prod_{s=1}^{u} \prod_{i \in P_s} f_s(\mathbf{x}_i \mid \mathbf{\Theta}_s) \right] \left[\prod_{i \in P_0} g_i(\mathbf{x}_i) \right], \tag{1}$$

$\mathbf{x}_i$ – wektor obserwacji, $\mathbf{x}_i = [x_{i1,\ldots,} x_{im}]$,

$f(\mathbf{x}_i \mid \mathbf{\Theta}_s)$ – funkcja gęstości rozkładu składowego P_s, $P_s = P_1, \ldots, P_u$,

$g(\mathbf{x}_i)$ – funkcja gęstości obserwacji nietypowych,

P_0 – rozkład składowy obserwacji nietypowych, $\#P_0 = [n\alpha]$,

α – parametr przycięcia (*trimming parameter*), tj. procent obserwacji nietypowych.

Liczebność zbioru obserwacji właściwych wyznaczana jest ze wzoru:

$$\#P_s = [n(1-\alpha)]. \tag{2}$$

Zakłada się, że obserwacje nietypowe pochodzą z innego rozkładu prawdopodobieństwa o funkcji gęstości oznaczonej symbolem $g(\mathbf{x}_i)$. Obserwacje te nie biorą udziału w maksymalizacji funkcji wiarygodności danej wzorem (1) wtedy, gdy spełniony jest poniższy warunek:

$$\arg\max_{P_\alpha} \max_{\mu_s, \Sigma} \prod_{s=1}^{u} \prod_{i \in P_s} f_s(\mathbf{x}_i \mid \mathbf{\Theta}_s) \subseteq \arg\max_{P_\alpha} \prod_{i \notin \bigcup_{s=1}^{u} P_s} g_i(\mathbf{x}_i), \tag{3}$$

gdzie: $P_\alpha = \left\{ \{P_0, P_1, \ldots, P_u\} : \bigcup_{s=0}^{u} P_s = \{1, \ldots, n\}, P_w \cap P_r = \varnothing \right.$ dla $w \neq r$ i $\left. \#P_0 = [n\alpha] \right\}$ (por. [García-Escudero i in. 2011]).

García-Escudero i in. [2008] dokonali pewnej modyfikacji modelu określonego za pomocą wzoru (1), zakładając różne macierze kowariancji ($\mathbf{\Sigma}_s$) oraz różne wagi dla poszczególnych klas ($\pi_s \geq 0, \sum_{s=1}^{u} \pi_s = 1$). Funkcję wiarygodności takiego modelu można zapisać jako:

$$L = \left[\prod_{s=1}^{u} \prod_{i \in P_s} \pi_s f_s(\mathbf{x}_i \mid \mathbf{\Theta}_s) \right] \left[\prod_{i \in P_0} g_i(\mathbf{x}_i) \right]. \tag{4}$$

W modelu tym zakłada się również, że iloraz największej i najmniejszej wartości własnej jest mniejszy od pewnej stałej c bądź jej równy:

$$M_n / m_n \leq c, \tag{5}$$

$$M_n = \max_{s=1,\ldots,u} \max_{j=1,\ldots,m} \lambda_{j,s}, \tag{6}$$

$$m_n = \min_{s=1,\ldots,u} \min_{j=1,\ldots,m} \lambda_{j,s}, \tag{7}$$

gdzie $\lambda_{j,s}$ to wartości własne macierzy $\boldsymbol{\Sigma}_s$. Podobne ograniczenia na macierz kowariancji nałożone zostały przez Hathawaya [1985] w przypadku mieszanek o jednowymiarowej zmiennej losowej. Zbiór wszystkich parametrów $\boldsymbol{\Theta}$, które spełniają warunek (5), oznaczany jest jako $\boldsymbol{\Theta}_s$. W praktyce maksymalizuje się najczęściej logarytm funkcji wiarygodności określonej jako:

$$L_c(\alpha, s) = \sum_{s=1}^{u} \sum_{i \in P_s} \log(\pi_s f(\mathbf{x}_i; \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)). \tag{8}$$

Stała c pozwala w pewien sposób kontrolować ograniczenia nakładane na macierz kowariancji. W przypadku gdy $c = 1$, macierz kowariancji jest najbardziej ograniczona, co odpowiada ważonej metodzie uciętych *k*-średnich (*weighted version of trimmed k-means*) [Cuesta-Albertos i in. 1997]. García-Escudero i in. [2010] pokazali, że dla c = 50 możliwe jest znalezienie klas o największym rozproszeniu (obserwacji właściwych).

Oceny parametrów funkcji wiarygodności danej wzorem (8) wyznaczane są za pomocą algorytmu TCLUST [García-Escudero i in. 2008]. Algorytm ten jest pewną modyfikacją algorytmu EM [Dempster i in. 1977] składającego się z kroku estymacji parametrów oraz kroku maksymalizacji funkcji największej wiarygodności, rozszerzoną o tzw. krok koncentracji (*concentration step*).

Algorytm TCLUST

1. Wyznacz wartości początkowe dla wartości przeciętnych $\hat{\mu}_1^0, \ldots, \hat{\mu}_s^0$, macierzy kowariancji $\hat{\boldsymbol{\Sigma}}_1^0, \ldots, \hat{\boldsymbol{\Sigma}}_s^0$ oraz początkowe wagi $\hat{\pi}_1^0, \ldots, \hat{\pi}_s^0$ dla każdego z rozkładów składowych.

2. W *l*-tej iteracji:

- Dla każdej obserwacji wyznacz prawdopodobieństwa *a posteriori*:

$$d_i = \pi_s^l f(\mathbf{x}_i; \mu_s^l, \boldsymbol{\Sigma}_s^l). \tag{9}$$

- Usuń [$n\alpha$] obserwacji o najniższych wartościach d_i (przypisz je do klasy P_0). Pozostaw („skoncentruj") pozostałe obserwacje i przydziel je do jednej z klas $P_s = P_1, \ldots, P_u$, aby spełniony był warunek (10):

$$d_i^* = \max_{s=1,\ldots,u} \pi_s^l f(\mathbf{x}_i; \mu_s^l, \mathbf{\Sigma}_s^l). \tag{10}$$

- Dla każdej klasy P_s wyznacz liczebność n_s, wartość przeciętną $\boldsymbol{\mu}_s$ oraz macierz kowariancji $\mathbf{\Sigma}_s$. Jeżeli wartości własne wyznaczonej w ten sposób macierzy kowariancji nie spełniają warunków określonych równaniem (5), problem musi zostać rozwiązany ponownie.
- Wyznacz nowe wartości parametrów $\mathbf{\Theta}^{l+1} = (\pi^{l+1}, \mu^{l+1}, \sum^{l+1})$, tak by funkcja wiarygodności dana wzorem (8) osiągnęła maksimum.

3. Procedurę iteracyjną powtarzaj kilkukrotnie.

Szczegółowe informacje na temat algorytmu TCLUST można znaleźć w pracy [García-Escudero i in. 2008; 2010; 2011].

3. Analiza empiryczna

Celem analizy empirycznej jest zbadanie jakości podziału uzyskanego za pomocą prezentowanego w artykule podejścia modelowego, polegającego na usuwaniu obserwacji oddalonych. Podejście to zostało dodatkowo porównane z bardziej popularną, odporną analizą skupień opartą na mieszankach rozkładów normalnych [Dasgupta, Raftery 1998; Fraley, Raftery 2002; Witek 2008][1]. W badaniu wykorzystano 4 zbiory generowane w przestrzeni dwuwymiarowej oraz 4 zbiory sztuczne, wykorzystywane w badaniach porównawczych metod taksonomicznych, dostępne w bibliotekach programu **R**. Zbiory te cechują się różnym stopniem separowalności (wśród zbiorów można znaleźć zbiory trudno i łatwo separowalne) o różnej liczbie klas. Poniżej przedstawiono krótką charakterystykę analizowanych zbiorów:

1. `6dnormals` – 6 niełatwo separowalnych klas, zawierających łącznie 500 obserwacji, wygenerowanych za pomocą funkcji `mlbench.2dnormals` z pakietu `mlbench`.

2. Zb_gen_1 – 5 łatwo separowalnych klas o kształcie sferycznym, zawierającym łącznie 2500 obserwacji, generowanych za pomocą modelu 8 funkcji `cluster.Gen` z pakietu `clusterSim`.

3. Zb_gen_2 – 5 niełatwo separowalnych klas, zawierających łącznie 2500 obserwacji, generowanych za pomocą modelu 6 funkcji `cluster.Gen` z pakietu `clusterSim`.

[1] W podejściu tym dokonywana jest modyfikacja modelu mieszanek, polegająca na dodaniu dodatkowej klasy (rozkładu składowego mieszanki) będącej reprezentantem obserwacji wyraźnie różniących się od obserwacji typowych. Obserwacje dodatkowej klasy przyjmują zwykle rozkład jednostajny lub rozkład Poissona.

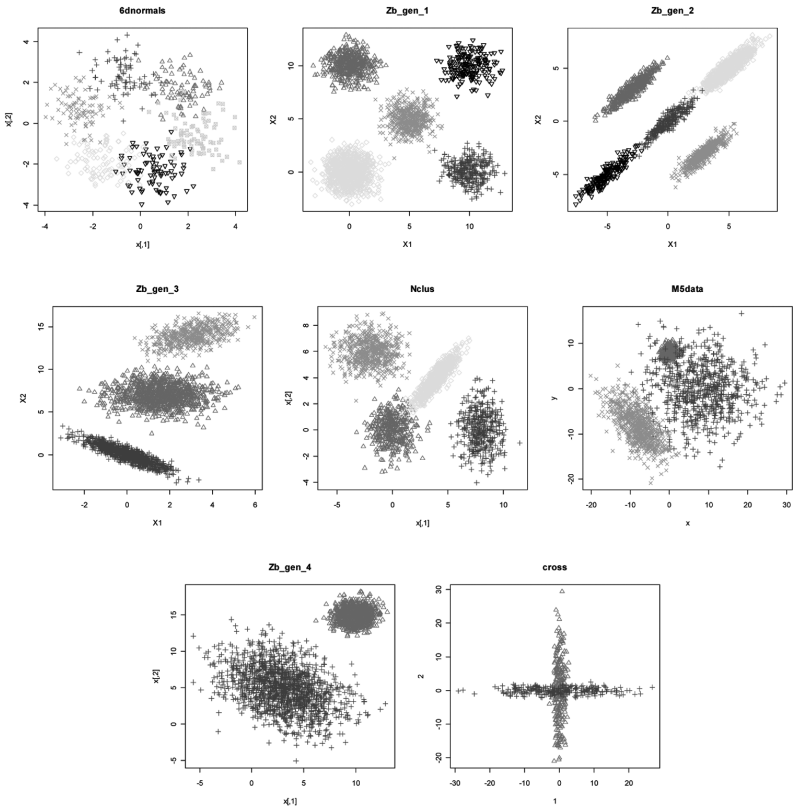
4. `Nclus` – 4 niełatwo separowalne klasy, zawierające łącznie 2200 obserwacji, zbiór dostępny w pakiecie `flexmix`[2].

5. Zb_gen_3 – 3 łatwo separowalne klasy o kształcie eliptycznym, zawierające łącznie 2500 obserwacji, generowanych za pomocą modelu 13 funkcji `cluster.Gen` z pakietu `clusterSim`.

6. `M5data` – 3 trudno separowalne klasy, zawierające łącznie 1800 obserwacji, zbiór dostępny w pakiecie `tclust`.

7. Zb_gen_4 – 2 niełatwo separowalne klasy, zawierające łącznie 2500 obserwacji, generowanych za pomocą funkcji `rmvnorm` z pakietu `mvtnorm`.

8. `cross` – 2 nakładające się na siebie klasy (nierozłączne), zawierające łącznie 500 obserwacji, zbiór dostępny w pakiecie `mclust`.



Rys. 1. Zbiory danych wykorzystane w analizie empirycznej

Źródło: obliczenia własne.

[2] W badaniu wygenerowano zbiór za pomocą funkcji `ExNclus()` o tej samej strukturze, lecz czterokrotnie większej liczbie obserwacji niż zbiór `Nclus`.

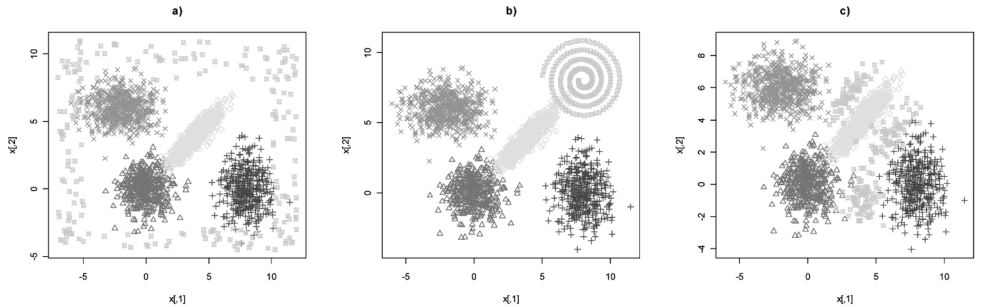
Wykorzystane w analizie zbiory zostały przedstawione na rys. 1. Do każdego z tych zbiorów dodano 10% obserwacji nietypowych, które znajdowały się:

a) poza obszarem wszystkich klas,

b) tworzyły dodatkową klasę o kształcie nieregularnym (spirali),

c) pomiędzy klasami (na granicy ich seperowalności).

Rozważane warianty dodawania obserwacji oddalonych dla jednego z analizowanych zbiorów zilustrowano graficznie na rys. 2.



Rys. 2. Trzy sposoby dodawania (generowania) obserwacji oddalonych dla zbioru `Nclus`

Źródło: obliczenia własne.

Tabela 1. Miary Randa (R_{HA}) dla zbiorów danych

Zbiór	Obserwacje nietypowe	`mclust`	`tclust`
6dnormals	a)	0,74	0,71
	b)	0,59	0,70
	c)	0,66	0,66
Zb_gen_1	a)	0,98	0,98
	b)	0,96	0,98
	c)	0,90	0,96
Zb_gen_2	a)	0,98	0,98
	b)	0,94	0,98
	c)	0,92	0,95
Nclus	a)	0,95	0,95
	b)	0,89	0,95
	c)	0,87	0,92
Zb_gen_3	a)	0,98	0,98
	b)	0,88	0,98
	c)	0,88	0,97
M5data	a)	0,83	0,89
	b)	0,68	0,94
	c)	0,78	0,82
Zb_gen_4	a)	0,96	0,96
	b)	0,80	0,94
	c)	0,81	0,85
cross	a)	0,76	0,73
	b)	0,76	0,74
	c)	0,71	0,65

Źródło: obliczenia własne.

W analizowanych zbiorach znana jest przynależność obiektów do poszczególnych klas. Informacja ta jest jednak traktowana jako informacja *a priori*, która wykorzystana została do badania zgodności podziałów uzyskanych za pomocą obu odpornych podejść w modelowej analizie skupień.

Porównania jakości klasyfikacji dokonano za pomocą skorygowanej miary Randa (R_{HA}) [Hubert, Arabie 1985, s. 198]. Obliczenia wykonano w programie **R** z zastosowaniem m.in. funkcji `mclust` (podejście modelowe z klasą dodatkową) oraz `tclust` (podejście modelowe polegające na ucinaniu obserwacji nietypowych). Uzyskane wyniki przedstawiono w tab. 1.

Jeżeli chodzi o jakość podziałów dla zbiorów, w których obserwacje nietypowe generowane były z rozkładu jednostajnego poza obszarem klas, obie metody dały porównywalne wyniki. Jakość podziału wyrażona za pomocą miary Randa dla analizy skupień opartej na mieszankach uciętych rozkładów normalnych jest wyraźnie lepsza w przypadku, gdy obserwacje nietypowe tworzą dodatkową klasę o nieregularnym kształcie lub znajdują się pomiędzy klasami. Metoda ta jednak daje nieco gorsze wyniki w przypadku zbiorów o klasach nierozłącznych (zbiór „cross").

4. Podsumowanie

Podejście modelowe, polegające na usuwaniu obserwacji oddalonych, pozwala na osiągnięcie zadowalających wyników podziału w przypadku, gdy w zbiorze danych występują obserwacje nietypowe. W przypadku obserwacji nietypowych tworzących dodatkową klasę o nieregularnym kształcie lub znajdujących się pomiędzy klasami metoda ta dała znacznie lepsze wyniki w przypadku zbiorów niejednorodnych (z wyjątkiem zbioru „cross"). W dalszych badaniach analizie poddane zostaną również zbiory o różnych wymiarach i różnych udziałach obserwacji nietypowych.

Literatura

Cuesta-Albertos J.A., García-Escudero L.A., Gordaliza A., *Trimmed k-means: an attempt to robusify quantizers*, „The Annals of Statistics" 1997, no 25.

Dasgupta A., Raftery A.E., *Detecting features in spatial point processes with clutter via model-based clustering*, „Journal of the American Statistical Association" 1998, no 93.

Dempster A.P., Laird N.M., Rubin D.B., *Maximum likelihood for incomplete data via the EM algorithm (with discussion)*, "Journal of the Royal Statistical Society" 1977, no B (39).

Fraley C., Raftery A.E., *Model-based clustering, discriminant analysis, and density estimation*, "Journal of the American Statistical Association" 2002, no 97.

Gallegos M.T., *Maximum Likelihood Clustering with Outliers*, [w:] K. Jajuga, A. Sokołowski, H. Bock, *Classification, Clustering and Data Analysis: Recent Advances and Applications*, Springer-Verlag, 2002.

Gallegos M.T., Ritter G., *A robust method for cluster analysis*, "Annals of Statistics" 2005, no 33(1).

García-Escudero L.A., Gordaliza A., Matrán C., Mayo-Iscar A., *A general trimming approach to robust cluster analysis*, „The Annals of Statistics" 2008, no 36(4).

García-Escudero L.A., Gordaliza A., Matrán C., Mayo-Iscar A., *A review of clustering methods*, „Advances of data analysis and classification", Springer 2010, no 4.
García-Escudero L.A., Gordaliza A., Matrán C., Mayo-Iscar A., *Exploring the number of groups in robust model-based clustering*, "Statistics and Computing" 2011, no 21(4).
Hathaway R.J., *A constrained formulation of maximum likelihood estimation for normal mixture distributions*, "The Annals of Statistics" 1985, no 13.
Hubert L.J., Arabie P., *Comparing partitions*, „Journal of Classifiacation" 1985, no 1.
Witek E., *Obserwacje nietypowe w analizie skupień – podejście modelowe*, [w:] J. Harasim, Warsztaty Doktoranckie'07, *Zarządzanie–Finanse–Ekonomia*, Wydawnictwo AE, Katowice 2008.

TRIMMING APPROACH TO THE MIXTURES OF NORMAL DISTRIBUTIONS

Summary: The paper presents trimming approach to the mixtures of normal distributions. In this approach the proportion of the most outlying observations is trimmed (during parameter estimation), different constraints on the cluster scatter matrices are assumed and trimming likelihood is defined. To estimate the parameter of this kind of mixture mostly the modified version of EM, i.e. TCLUST algorithm is applied [García-Escudero et al. 2010, p. 89-109]. The results of clustering given by trimming approach and well known robust mixtures of normal distributions are compared.

Keywords: mixture model, EM algorithm, TCLUST algorithm.