

УДК 004.8:004.032.26

Е.В. Бодянский, В.А. Самитова

Харьковский национальный университет радиоэлектроники, Харьков

## НЕЧЕТКАЯ КЛАСТЕРИЗАЦИЯ ДАННЫХ В ПОРЯДКОВОЙ ШКАЛЕ НА ОСНОВЕ СОВМЕСТНОГО ИСПОЛЬЗОВАНИЯ ФУНКЦИЙ ПРИНАДЛЕЖНОСТИ И ПРАВДОПОДОБИЯ

Рассмотрена задача кластеризации данных, заданных в порядковой шкале, в условиях перекрывающихся кластеров. Для классификации предложено использовать подход, основанный на совместном применении функций принадлежности и правдоподобия.

**Ключевые слова:** кластеризация, порядковая шкала, FCM, функции принадлежности, функции правдоподобия, формула Байеса.

### Введение

В настоящее время задачи обработки информации, заданной в нечисловом виде, получили широкое распространение. Подобные задачи часто встречаются в социологии, экономике, медицине, образовании и т.п. В существующих методах кластеризации подобных данных таких, как метод  $k$ -средних [1, 2], “Fuzzy C-means” (FCM) [3, 4], EM-алгоритм [5, 6], чаще всего используется подход, основанный на замене лингвистических переменных их рангами. Однако в большинстве случаев этот прием оказывается некорректным, поскольку предполагает равенство расстояний между соседними числовыми рангами, что не всегда соответствует действительности.

Более естественным представляется подход, развиваемый Р.К. Брауэром [7] и основанный на максимизации функции правдоподобия. Ограничением этого подхода является предположение о гауссовом распределении исходных данных, что во многих приложениях не выполняется, а также способ вычисления правдоподобия для порядковых переменных.

В данной статье предлагается алгоритм нечеткой кластеризации данных в порядковой шкале на основе совместного использования функций принадлежности и правдоподобия. Исходной информацией для решения задачи является упорядоченная последовательность лингвистических переменных  $x^1, x^2, \dots, x^m, 1 < \dots < 1-1 < 1 < 1+1 < \dots < m$ , где  $x^1$  – собственно лингвистическая переменная,  $1$  – соответствующий ранг.

Для обработки информации, заданной в порядковой шкале, в [8 – 10] было предложено осуществлять фазификацию исходных данных на основе анализа распределения частот появления конкретных лингвистических переменных, при этом предполагалось, что эти распределения подчиняются гауссовскому закону. В [11] был использован подход, не связанный с гипотезой нормальности рас-

пределения, который мы и будем использовать в дальнейшем. Таким образом, исходной информацией для решения задачи является выборка наблюдений, сформированная из  $N$   $n$ -мерных векторов признаков  $X = \{x_1, x_2, \dots, x_j, \dots, x_N\}$ , где  $j = 1, \dots, N$ ,  $x_j = \{x_{jk}^l\}$ ,  $k = 1, \dots, n$ ;  $l = 1, \dots, m$  – ранг конкретного значения лингвистической переменной по  $k$ -й координате  $n$ -мерного пространства для  $j$ -го объекта, подлежащего кластеризации.

Результатом работы алгоритма является разбиение исходного массива данных  $X$  на  $s$  классов (кластеров) с вычислением уровня принадлежности  $w_{i,j}$   $j$ -го вектора признаков  $i$ -му кластеру.

### Основная часть

#### Правдоподобие и вероятность

Существует несколько основных подходов к кластеризации данных – иерархический, метрический, итерационный и т.п. [12].

Итерационная кластеризация применяется во многих областях, при этом алгоритм в цикле находит лучшие кластеры, к которым могут принадлежать наблюдения.

Рассмотрим простейший пример, в котором каждое наблюдение имеет четыре атрибута  $x_1, x_2, x_3, x_4$ .

Предполагая, что они являются взаимно независимыми, задача итерационной кластеризации сводится к задаче нахождения кластера  $u$ , путем максимизации правдоподобия  $P(u | x_1 x_2 x_3 x_4)$  для каждого наблюдения с характеристиками  $x_1, x_2, x_3, x_4$ . По формуле Байеса это правдоподобие может быть вычислено следующим образом:

$$P(u | x_1 x_2 x_3 x_4) = \frac{P(x_1 x_2 x_3 x_4 | u) P(u)}{P(x_1 x_2 x_3 x_4)}, \quad (1)$$

то есть, нахождение кластера  $u$  путем максимизации

ции правдоподобия  $P(y|x_1x_2x_3x_4)$  эквивалентно решению этой задачи путем максимизации условной вероятности  $P(x_1x_2x_3x_4|y)$ .

Более того, предположение о том, что характеристики взаимно независимы, позволяет записать очевидное соотношение:

$$P(x_1x_2x_3x_4|y) = P(x_1|y)P(x_2|y) \times P(x_3|y)P(x_4|y). \quad (2)$$

Следовательно, проблема нахождения кластера  $y$  представляет собой проблему максимизации правой части уравнения (2).

Таким образом, можно говорить о том, что проблема нахождения кластеров решается путем максимизации произведения индивидуальных условных вероятностей характеристик наблюдения.

Заметим, что вероятность  $P(x_j|y)$  выражает, как часто наблюдение  $x_j$  появляется в выборке со всеми одинаковыми значениями характеристик в кластере  $y$ , т.е. вероятность  $P(x_j|y)$  выражает определенный вид частоты встречаемости наблюдения  $x_j$  с одинаковыми значениями параметров в кластере  $y$ .

**Алгоритм нечеткой кластеризации  
порядковых данных на основе  
совместного использования  
функций принадлежности  
и функции правдоподобия**

Предлагаемый алгоритм имеет достаточно близкую алгоритмическую структуру к алгоритму "Fuzzy C-means" (FCM).

Задача кластеризации с помощью алгоритма "Fuzzy C-means" (FCM) для количественных характеристик решается путем минимизации целевой функции:

$$Q = \sum_{i=1}^c \sum_{j=1}^N w_{i,j}^\beta \|x_j - v_i\|^2 \quad (3)$$

при ограничениях:

$$\begin{aligned} w_{i,j} &\geq 0, \forall i = 1, \dots, c; \forall j = 1, \dots, N, \\ \sum_{i=1}^c w_{i,j} &= 1, \forall j = 1, \dots, N, \\ \sum_{j=1}^N w_{i,j} &> 0, \forall i = 1, \dots, c, \end{aligned} \quad (4)$$

где  $w_{i,j}$  – уровень принадлежности  $j$ -го наблюдения к  $i$ -му кластеру;

$\beta$  – неотрицательный параметр фаззификации.

При этом уровень принадлежности и прототипы кластеров вычисляются по формулам

$$w_{t,j} = \frac{1}{\sum_{i=1}^c \left( \frac{\|x_j - v_t\|}{\|x_j - v_i\|} \right)^{\frac{2}{\beta-1}}}, \quad (5)$$

$$\forall t = 1, \dots, c; \forall j = 1, \dots, N;$$

$$v_t = \frac{\sum_{j=1}^N w_{t,j}^\beta x_j}{\sum_{j=1}^N w_{t,j}^\beta}, \quad (6)$$

$$\forall t = 1, \dots, c.$$

Из выражений (5) и (6) видно, что при вычислении уровня принадлежности конкретного наблюдения к кластеру  $w_{i,j}$  используется расстояние между наблюдением и соответствующими центроидами кластера  $v_i$ . Далее пересчитывается  $v_i$  на основе уровней принадлежности к кластерам  $w_{i,j}$ . Вычисления производятся итерационно, пока не будет выполнено условие останова алгоритма.

Идея предлагаемого алгоритма состоит в том, чтобы использовать правдоподобия наблюдений для определения кластеров вместо расстояний в алгоритме "Fuzzy C-means" (FCM). Таким образом, задача решается путем максимизации целевой функции:

$$Q = \sum_{i=1}^c \sum_{j=1}^N w_{i,j}^\beta L_{i,j}, \quad (7)$$

или соответственно минимизации:

$$Q = \sum_{i=1}^c \sum_{j=1}^N w_{i,j}^\beta U_{i,j} \quad (8)$$

при ограничениях:

$$w_{i,j} \geq 0, \forall i = 1, \dots, c; \forall j = 1, \dots, N,$$

$$\sum_{i=1}^c w_{i,j} = 1, \forall j = 1, \dots, N, \quad (9)$$

$$\sum_{j=1}^N w_{i,j} > 0, \forall i = 1, \dots, c,$$

где  $L_{i,j}$  – правдоподобие принадлежности  $j$ -го наблюдения к  $i$ -му кластеру;  $U_{i,j}$  – логарифм несходства  $j$ -го наблюдения с  $i$ -м кластером.

Правдоподобие  $L_{i,j}$  в (7) вычисляется согласно формуле

$$L_{i,j} = \prod_{k=1}^n p_{i,j,k}, \quad (10)$$

где  $p_{i,j,k}$  – условная вероятность появления определенного значения  $k$ -й характеристики  $j$ -го наблюдения в  $i$ -м кластере, и вычисляется следующим образом:

$$P_{i,j,k} = P(x_{j,k} | y_i). \quad (11)$$

Логарифм несходства в (8) определяется следующим образом:

$$U_{i,j} = -\ln L_{i,j}, \quad (12)$$

при этом целевую функцию (8) можно переписать в виде:

$$\begin{aligned} Q &= \sum_{i=1}^c \sum_{j=1}^N w_{i,j}^\beta U_{i,j} = \\ &= \sum_{i=1}^c \sum_{j=1}^N w_{i,j}^\beta \left(-\ln \prod_{k=1}^n P_{i,j,k}\right) = \\ &= -\sum_{i=1}^c \sum_{j=1}^N w_{i,j}^\beta \sum_{k=1}^n \ln P_{i,j,k}. \end{aligned} \quad (13)$$

Для вычисления  $w_{i,j}$  используется выражение [7]:

$$w_{t,j} = \frac{1}{\sum_{i=1}^c \left(\frac{U_{t,j}}{U_{i,j}}\right)^{m-1}}, \quad \forall t=1, \dots, c; \forall j=1, \dots, N, \quad (14)$$

а для подсчета условных вероятностей  $P_{i,j,k}$  используются функции принадлежности, описанные ниже.

Недостатком данного подхода является то, что рассматриваемый объект «размазывается» по всем существующим кластерам, что в ранговой шкале ведет к потере физического смысла.

В связи с этим представляется целесообразным после вычисления центроидов, пересчитать все расстояния  $d(x_j, v_i)$ , провести их ранжирование по возрастанию и выбрать наименьшее  $d_{\min \min}(x_j, v_i)$  и следующее за ним  $d_{\min}(x_j, v_1)$ . Принимая в расчет два наименьших расстояния, можно воспользоваться формулами [11]:

$$w_{ji} = \frac{d_{\min \min}^{-2}(x_j, v_i)}{d_{\min \min}^{-2}(x_j, v_i) + d_{\min}^{-2}(x_j, v_1)}, \quad (15)$$

$$w_{j1} = \frac{d_{\min}^{-2}(x_j, v_1)}{d_{\min \min}^{-2}(x_j, v_i) + d_{\min}^{-2}(x_j, v_1)}. \quad (16)$$

Таким образом, алгоритм имеет следующий вид:

1. Инициализация

$P_{i,j,k}, \forall i=1, \dots, c; \forall j=1, \dots, N; \forall k=1, \dots, n$  случайными значениями.

2. Подсчет  $w_{i,j}, \forall i=1, \dots, c; \forall j=1, \dots, N$  с помощью формулы (14).

3. Подсчет

$P_{i,j,k}, \forall i=1, \dots, c; \forall j=1, \dots, N; \forall k=1, \dots, n$  с помощью формулы (23).

4. Шаг 2 и 3 повторяется итерационно до выполнения условия:

$$\varepsilon \leq \max_{i,j} \left\{ \left| \text{old}_{\mu_{i,j}} - \text{new}_{\mu_{i,j}} \right| \right\}.$$

5. Расчет всех расстояний  $d(x_j, v_i) = \|x_j - v_i\|$  и выделение двух наименьших расстояний  $d_{\min \min}(x_j, v_i)$  и  $d_{\min}(x_j, v_1)$ , где 1 может принимать значение или  $i-1$ , или  $i+1$ ;

6. Расчет уровней принадлежности  $x_j$  к двум соседним кластерам по формулам (15) и (16).

**Вычисление условной вероятности  $P_{i,j,k}$  и фаззификация исходных данных**

Процесс фаззификации последовательности ранговых лингвистических переменных рассмотрим на примере одномерной выборки  $x_1, \dots, x_N$ , где каждому из наблюдений  $x_j$  может быть приписан один из рангов  $l, l=1, \dots, m$ .

Пусть значение  $x_j$ , соответствующее  $l$ -му рангу, встречается в выборке  $N_l$  раз. Тогда в рассмотрение вводятся относительные частоты появления  $l$ -го ранга

$$f_l = \frac{N_l}{N}, \quad (17)$$

при этом естественно выполняется условие

$$\sum_{i=1}^m f_i = 1. \quad (18)$$

На основе относительных частот формируются усредненные частоты встречаемости наблюдений, при этом для их вычисления удобно воспользоваться рекуррентным соотношением

$$\begin{aligned} c_1 &= 0.5f_1, \\ c_l &= c_{l-1} + 0.5(f_{l-1} + f_l), \quad \forall l=2, \dots, m. \end{aligned} \quad (19)$$

Далее все порядковые данные заменяются соответствующими усредненными частотами встречаемости наблюдений. Этап фаззификации представлен в виде гистограммы на рис. 1.

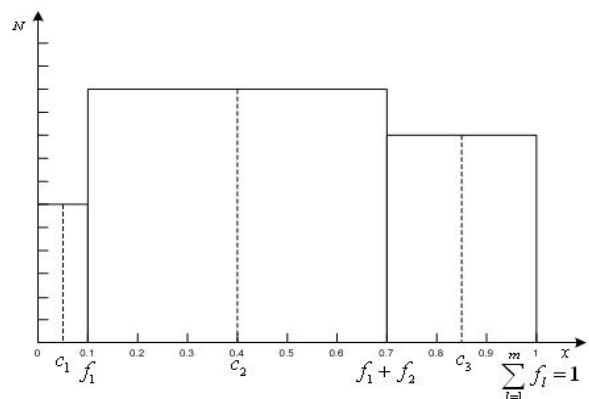


Рис. 1. Гистограмма распределения порядковых переменных по частоте встречаемости в выборке

Предполагая, что уровень принадлежности наблюдений к кластерам

$$\mu_{i,j}, \forall i = 1, \dots, c; \forall j = 1, \dots, N$$

известен, вычисляется мода для каждой характеристики по каждому из кластеров

$$x_{i,k}^*, \forall i = 1, \dots, c; \forall k = 1, \dots, n.$$

Далее, учитывая полученные моды, строится ассиметричная функция принадлежности следующим образом:

Если  $x_{i,k}^* > 0.5$ , то функция принадлежности имеет вид, представленный на рис. 2.

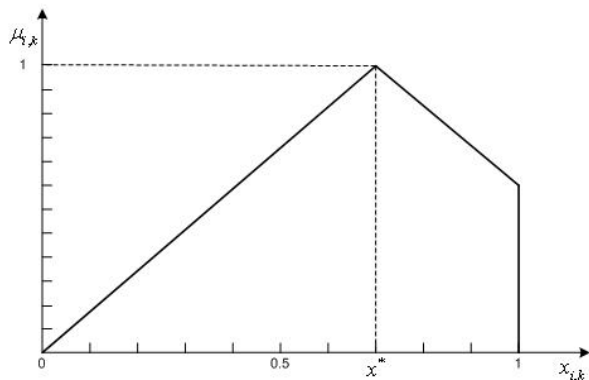


Рис. 2. Функция принадлежности для выборки ранговых переменных, когда  $x_{i,k}^* > 0.5$

и описываемый формулой

$$\mu_{i,j,k} = \begin{cases} \frac{x_j}{x_{i,k}^*}, & x \in [0, x_{i,k}^*], \\ \frac{2x_{i,k}^* - x_j}{x_{i,k}^*}, & x \notin [0, x_{i,k}^*]. \end{cases} \quad (20)$$

Если  $x_{i,k}^* < 0.5$ , то функция принадлежности имеет вид, представленный на рис. 3.

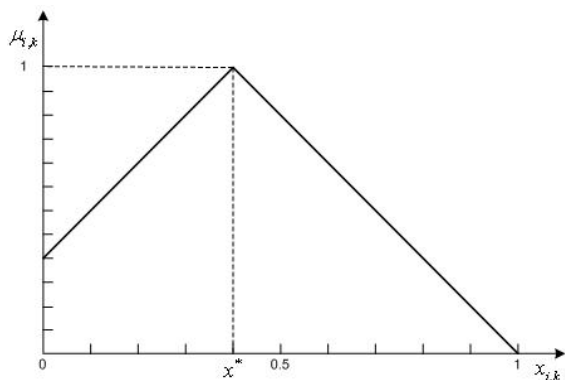


Рис. 3. Функция принадлежности для выборки ранговых переменных, когда  $x_{i,k}^* < 0.5$

и описываемый формулой

$$\mu_{i,j,k} = \begin{cases} \frac{1-x_j}{1-x_{i,k}^*}, & x \in [x_{i,k}^*, 1], \\ \frac{x_j - 2x_{i,k}^* + 1}{1-x_{i,k}^*}, & x \notin [x_{i,k}^*, 1]. \end{cases} \quad (21)$$

Если  $x_{i,k}^* = 0.5$ , то функция принадлежности имеет вид, представленный на рис. 4.

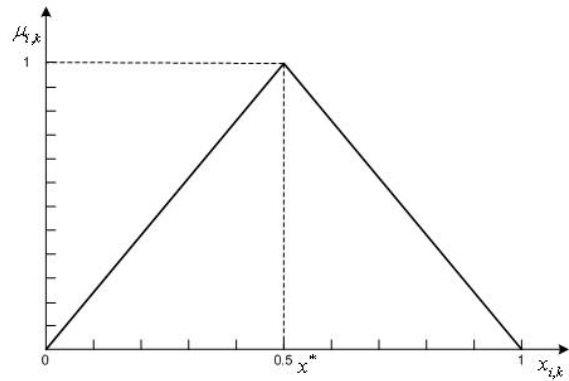


Рис. 4. Функция принадлежности для выборки ранговых переменных, когда  $x_{i,k}^* = 0.5$

и описываемый формулой

$$\mu_{i,j,k} = \begin{cases} \frac{x_j}{x_{i,k}^*}, & x \in [0, x_{i,k}^*], \\ \frac{1-x_j}{1-x_{i,k}^*}, & x \in [x_{i,k}^*, 1]. \end{cases} \quad (22)$$

Поскольку условная вероятность  $p_{i,j,k}$  напрямую зависит от частоты встречаемости конкретного значения характеристики в выборке, а порядковые данные идут в четко заданном порядке от самого малого к самому большому, то можно сказать, что

$$p_{i,j,k} = \mu_{i,j,k}. \quad (23)$$

### Численное моделирование

Для проверки работоспособности предложенного алгоритма были использованы данные об успеваемости студентов первого курса Харьковского национального университета радиоэлектроники. Набор данных содержит оценки по трем предметам для 1108 человек.

В результате работы алгоритма были определены центроиды для каждого из рангов (оценок) по каждой из переменных. Результаты представлены в табл. 1.

Таблица 1

Центроиды рангов наблюдений

№/Оценка	«2»	«3»	«4»	«5»
1	0,011	0,18	0,66	0,99
2	0,01	0,28	0,69	0,92
3	0,006	0,29	0,75	0,96

Далее данные были разбиты на 4 кластера: "отличник", "хорошист", "учащийся удовлетворительно" и "неудовлетворительно". По результатам работы алгоритма были получены центроиды кластеров, представленные в табл. 2.

Таблица 2

## Центроиды кластеров

Кластеры / Параметры	Предмет №1	Предмет №2	Предмет №3
1	0,66	0,92	0,75
2	0,66	0,69	0,75
3	0,66	0,28	0,29
4	0,18	0,28	0,29

15% учащихся было отнесено к кластеру "отличник", 26% учащихся было отнесено к кластеру "хорошист", 30% – к кластеру "учащийся удовлетворительно" и 29% – к кластеру "неудовлетворительно".

Результаты эксперимента на 98% совпадают с классификацией, сделанной деканатом.

**Выводы**

Предложен алгоритм нечеткой классификации данных в порядковой шкале на основе совместного использования функций принадлежности и правдоподобия. Данный подход позволяет эффективно обрабатывать информацию благодаря учету характера распределения обрабатываемых данных. Метод фаззификации данных и способ определения условной вероятности появления конкретных наблюдений в каждом кластере  $p_{i,j,k}$  позволяют быстро и точно классифицировать выборку.

Одним из преимуществ данного подхода, является его устойчивость к выбросам благодаря использованию порядка следования переменных при построении функций принадлежности.

**Список литературы**

1. MacQueen J. *Some methods of classification and analysis of multivariate observations* / J. MacQueen // *Berkely*

*Symposium on Mathematical Statistics and Probability. – Berkeley, 1967. – Vol. 1. – P. 281-297.*

2. Lloyd S.P. *Least squares quantization in PCM* / S.P. Lloyd // *IEEE Transactions on Information Theory. – 1982. – Vol. IT-28. – P. 129-137.*

3. Bezdek J.C. *Pattern Recognition with Fuzzy Objective Function Algorithms* / J.C. Bezdek. – N.Y.: Plenum Press, 1981. – 272 p.

4. Jang J.-Sh.R. *Neuro-Fuzzy and Soft Computing* / J.-Sh.R. Jang, Ch.-T. Sun, E. Mizutani. – Upper Saddle River, NJ: Prentice Hall, 1997. – 614 p.

5. Dempster A.P. *Maximum-likelihood from incomplete data via the EM algorithm* / A.P. Dempster, N.M. Laird, D.B. Rubin // *Journal of the Royal Statistical Society. – 1977. – Vol. B. – P. 1-38.*

6. Zhong S. *A unified framework for model-based clustering* / S. Zhong, J. Ghosh // *Journal of Machine Learning Research. – 2003. – Vol. 4. – P. 1001-1037.*

7. Mahnhoon L. *Likelihood based fuzzy clustering for data sets of mixed features* / L. Mahnhoon, R.K. Brouwer // *IEEE Symp. on Foundations of Comput. Intell. FOCI 2007. – 2007. – P. 544-549.*

8. Brouwer R.K. *A feedforward neural network for mapping vectors to fuzzy sets of vectors* / R.K. Brouwer, W. Pedrycz // *Proc.Int.Conf. on Artificial Neural Networks and Neural Information Processing ICANN/ICOMIP 2003. – Istanbul, Turkey, 2003. – P.45-48.*

9. Butkiewicz B.S. *Robust fuzzy clustering with fuzzy data* / B.S. Butkiewicz // *Lecture Notes in Computer Science. – Berlin – Heidelberg: Springer-Verlag, 2005. – V. 3528. – P. 76-82.*

10. Brouwer R.K. *Fuzzy set covering of a set of ordinal attributes without parameter sharing* / R.K. Brouwer // *Fuzzy Sets and Systems. – 2006. – 157. – №13. – P. 1775-1786.*

11. Бодянский Е.В. *Нечеткая кластеризация данных, заданных в порядковой шкале* / Е.В. Бодянский, В.А. Опанасенко, А.Н. Слипченко // *Системы обработки информации. – X.: XV ПС, 2007. – Вып. 4(62). – С. 5-9.*

12. Hoepfner F. *Fuzzy-Clusteranalysis* / F. Hoepfner, F. Klawonn, R. Kruse. – Braunschweig: Vieweg, 1997. – 280 p.

Поступила в редколлегию 22.06.2010

**Рецензент:** д-р техн. наук, проф. В.А. Филатов, Харьковский национальный университет радиоэлектроники, Харьков.

**НЕЧІТКА КЛАСТЕРИЗАЦІЯ ДАНИХ У ПОРЯДКОВІЙ ШКАЛІ НА ОСНОВІ СУМІСНОГО ЗАСТОСУВАННЯ ФУНКЦІЙ НАЛЕЖНОСТІ ТА ПРАВДОПОДІБНОСТІ**

Є.В. Бодяньський, В.О. Самітова

Розглядається задача кластеризації даних, що задані в порядковій шкалі, в умовах кластерів, що перекриваються. Для класифікації запропоновано підхід, на основі сумісного використання функцій належності та правдоподібності.

**Ключові слова:** кластеризація, порядкова шкала, FCM, функції належності, функції правдоподібності, формула Байєса.

**FUZZY CLUSTERIZATION OF DATA IN ORDINAL SCALE BASED ON MEMBERSHIP AND LIKELIHOOD FUNCTIONS**

Ye. V. Bodyanskiy, V.A. Samitova

Fuzzy clusterization of data in ordinal scale based on membership and likelihood functions taking into account the overlapped clusters is considered.

**Keywords:** clusterization, index scale, FCM, functions of belonging, functions of verisimilitude, formula of Bayes.