УДК 519.237.8

МОДИФИКАЦИЯ АЛГОРИТМА КЛАСТЕРИЗАЦИИ С-СРЕДНИХ НА ОСНОВЕ ИСПОЛЬЗОВАНИЯ ОБЪЕМНЫХ ПРОТОТИПОВ И СЛИЯНИЯ СХОЖИХ КЛАСТЕРОВ

А. С. Тарасова

Воронежский государственный университет

Алгоритм кластеризации С–средних (Fuzzy C-Means, FCM) является одним из наиболее часто используемых алгоритмов. Данный алгоритм обладает рядом недостатков, таких как неустойчивость к начальным данным, неспособность распознавать кластеры малого объема. Также существует проблема определения оптимального количества кластеров. В данной работе рассмотрены способы усовершенствования алгоритма FCM с целью устранения этих недостатков.

1. ВВЕДЕНИЕ

Рассмотрим постановку задачи кластерного анализа. Пусть N — количество рассматриваемых объектов, $j = \overline{1,N}$ — номер объекта, $X = \left\{x^j, j = \overline{1,N}\right\}$ — множество рассматриваемых объектов, где каждый объект $x^j = (x^j{}_1, \ldots x^j{}_K)$, характеризуется K признаками. Требуется произвести разбиение данных на кластеры (группы данных, которые близки между собой и далеки друг от друга).

Пусть произведена кластеризация данных и выделено M кластеров, $i = \overline{1,M}$ — номер кластера, $V = \left\{v_i, i = \overline{1,M}\right\}$ — множество центров кластеров, S_i — i-й кластер, u_{ij} — степень принадлежности объекта x_j кластеру S_i.

Кластерный анализ представлен в литературе большим количеством подходов.

Исследуемый нами алгоритм С-средних относится к алгоритмам *минимизации квадрата ошибки разбиения* в группе *разбиенческих методов*, которые продуцируют только одно разбиение в отличие от *иерархических методов*, продуцирующих ряды разбиений, вложенные друг в друга. Алгоритм С-средних не относится ни к *агломеративным подходам*, которые начинают разбиение с каждого паттерна в отдельном кластере (синглтон) и объединяют кластеры, пока не выполнится критерий останова, ни к *разделяющим подходам*, которые начинают с кластера, содержащего все паттерны, и производят разделение, пока не выполнится критерий останова. В алгоритме С-средних степени принадлежности объектов ко всем кластерам находят одновременно. Алгоритм С-средних является *политетичным* алгоритмом, то есть кластеризация основана на сравнении нескольких признаков. Также он относится к *нечетким* алгоритмам, которые присваивают каждому паттерну степень принадлежности в каждом кластере. Алгоритм С-средних относится к *детерминистическим* алгоритмам, то есть оптимизация произведена с использованием традиционных технологий.

Рассмотрим основные положения алгоритма кластеризации С–Средних, который предложил Bezdek [2].

Задача состоит в разбиении N объектов на M кластеров. Пусть u_{ik}, $k = \overline{1,N}$, $i = \overline{1,M}$ – степень принадлежности объекта x_k к кластеру A_i. Нечеткие кластеры опишем следующей матрицей нечеткого разбиения:

$$U = \left\{u_{ik}\right\},\ u_{ik} \in [0,1],\ k = \overline{1,N},\ i = \overline{1,M},$$

k-я строка $(x_{k1}, .., x_{kM})$ содержит степени принадлежности объекта к кластерам $A_1, \ldots\ldots A_M$. Матрица нечеткого разбиения соответственно должна удовлетворять следующим условиям:

$$\sum_{i=1}^{M} u_{ik} = 1,\ u_{ik} \in [0,1], \qquad (1)$$

$$0 < \sum_{k=1}^{N} u_{ik} < N,\ i = \overline{1,M}. \qquad (2)$$

Нечеткое разбиение позволяет просто решить проблему объектов, расположенных на границе двух кластеров – им назначают степени принадлежностей равные 0.5. Недостаток нечеткого разбиения проявляется при работе с объектами, удаленными от центров всех кластеров. Удаленные объекты имеют мало общего с любым из кластеров, поэтому интуитивно хочется назначить для них малые степени принадлежности. Однако, по условию (1) сумма их

степеней принадлежностей такая же, как и для объектов, близких к центрам кластеров, т.е. равна единице. Для устранения этого недостатка можно использовать *возможностное разбиение*, которое требует, только чтобы произвольный объект принадлежал хотя бы одному кластеру. Возможностное разбиение получается следующим ослаблением условия (1):

$$\forall k \ \exists i : \left(u_{ik} > 0\right).$$

Для оценки качества нечеткого разбиения используется следующий критерий разброса:

$$F = \sum_{k=1}^{N}\sum_{i=1}^{M}\left(u_{ik}\right)^m \left\|x_k - \nu_i\right\|^2 \rightarrow \min,$$

где величина m носит название экспоненциального веса и задается пользователем. Величина m уменьшает влияние «шума» при вычислении центров кластеров и значения целевой функции. Данная константа выполняет дефазификацию: большие значения степеней принадлежности увеличиваются, а маленькие – уменьшаются. Следует задавать $m > 1$, причем, чем больше m, тем менее нечеткую кластеризацию мы получаем.

Алгоритм продолжает работу до тех пор, пока значение данного критерия на следующей итерации будет отличаться от значения на предыдущей итерации на малую величину ε, заданную пользователем.

Центры кластеров ν_i находят по следующей формуле:

$$\nu_i = \frac{1}{\sum_{k=1}^{N}\left(u_{ik}\right)^m}\sum_{k=1}^{N}\left(u_{ik}\right)^m x_k,$$

Степени принадлежности объектов к кластерам находят по следующей формуле:

$$u_{ik} = \frac{\left(\dfrac{1}{\left\|x_k - \nu_i\right\|^2}\right)^{1/(m-1)}}{\sum_{j=1}^{M}\left(\dfrac{1}{\left\|x_k - \nu_j\right\|^2}\right)^{1/(m-1)}}.$$

В данной статье рассмотрены способы усовершенствования алгоритма С-средних с целью устранения таких недостатков данного алгоритма как неустойчивость к начальным данным, неспособность распознавать кластеры малого объема и определение оптимального количества кластеров.

2. УСОВЕРШЕНСТВОВАНИЕ АЛГОРИТМА С-СРЕДНИХ ПУТЕМ ИСПОЛЬЗОВАНИЯ ОБЪЕМНЫХ ПРОТОТИПОВ

Одна из идей усовершенствования метода FCM — идея использования объемного прототипа. Главное преимущество использования объемных прототипов состоит в снижении чувствительности алгоритма кластеризации к различиям в объемах кластеров и различиям в распределении паттернов данных. Это придает алгоритмам кластеризации большую устойчивость. В дальнейшем прототипы также чрезвычайно полезны при генерации нечетких правил с использованием нечеткой кластеризации. При этом ядра нечетких множеств уже не будут являться точками, а форма нечетких множеств уже будет скорее зависеть от самих данных, чем от свойств алгоритма кластеризации.

Часто элементы данных, близкие к центру кластера могут рассматриваться как полностью принадлежащие кластеру. Особенно это характерно для случая, когда существует несколько кластеров, хорошо отделенных друг от друга. Имеет смысл расширить ядро кластера от одной точки до области в пространстве. Следующее определение объемному прототипу дается в [1]:

Объемный прототип $V \in R^n$ — это n-мерное, выпуклое и компактное подпространство кластерного пространства.

Заметим, что согласно этому определению, прототип может иметь произвольную форму. Естественно выбирать прототипы таким образом, чтобы они «продолжали» расстояние, например, для алгоритма FCM можно выбрать прототипы в форме гиперсфер, а в алгоритме Gustavson—Kessel [6] — гиперэллипсоидов. Элементы, попадающие в прототип, имеют степень принадлежности, равную 1 в соответствующем кластере.

Если ищется прототип в форме гиперэллипсоида или гиперсферы, то его размер может определяться с помощью радиуса r_i, который может либо задаваться пользователем жестко, либо может быть оценен из имеющихся данных. Естественным образом вычислить радиус r_i можно связав его с размером кластера. Это может быть достигнуто путем рассмотрения нечеткой ковариационной матрицы кластера:

$$P_i = \frac{\sum_{k=1}^{N} u_{ik}^m (x_k - v_i)(x_k - v_i)^T}{\sum_{k=1}^{N} u_{ik}^m}$$

Определитель $|P_i|$ дает размер (объем) кластера. Так как P_i – симметричная и положительно определенная матрица, то она может быть представлена в виде $P_i = Q_i \Lambda_i Q_i^T$, где Q_i — ортонормированная матрица (матрица собственных векторов матрицы), а Λ_i — диагональная матрица собственных значений матрицы P_i. В случае эллипсоида элементы матрицы Λ_i неотрицательны. Пусть $\sqrt{\Lambda_i}$ — матрица, элементы которой равны корням квадратным от элементов матрицы Λ_i. Пусть A_i — матрица, индуцирующая норму $d^2(x,v) = (x_k - v_i)A_i(x_k - v_i)^T$. В данном случае $A_i = P_i$, однако, для формирования расстояния может использоваться и матрица, отличная от ковариационной. Тогда радиус r_i, как показано в [1], задается следующим образом:

$$r_i = \sqrt{\Lambda_i} Q_i^T A_i Q_i \sqrt{\Lambda_i}\ .$$

В алгоритмах, использующих объемные прототипы, используют также модифицированное расстояние, которое измеряют в два этапа:

1. Расстояние $d_{ik}(x_k, v_i)$ от точки до центра кластера,

2. Расстояние от точки до прототипа $\tilde{d}_{ik} = \max(0, d_{ik} - r_i)$.

Одним из недостатков алгоритма FCM является то, что он часто не способен распознать кластеры малого объема. Это происходит потому, что плотные кластеры с большой населенностью «притягивают» точки других паттернов данных (рис. 1)

При использовании объемного прототипа ситуация меняется. Так как точки, принадлежащие прототипу, полностью принадлежат одному кластеру, то они не влияют на остальные кластеры. Это уменьшает способность плотных областей данных «притягивать» данные из других паттернов данных, то есть влияние данных больших кластеров снижается и становится возможным распознать кластеры с маленьким количеством точек.

3. ОПРЕДЕЛЕНИЕ ОПТИМАЛЬНОГО КОЛИЧЕСТВА КЛАСТЕРОВ

Определение естественного количества групп данных необходимо для успешного применения алгоритма кластеризации. Одним из способов является метод, основанный на слиянии схожих кластеров. Метод начинает свою работу с количеством кластеров, равному некоторому верхнему пределу. После оценки подобия кластеров, происходит слияние схожих кластеров. Если подобие кластеров больше, чем некоторая граница $\alpha \in [0,1]$, то наиболее схожие кластеры на каждой итерации объединяют.

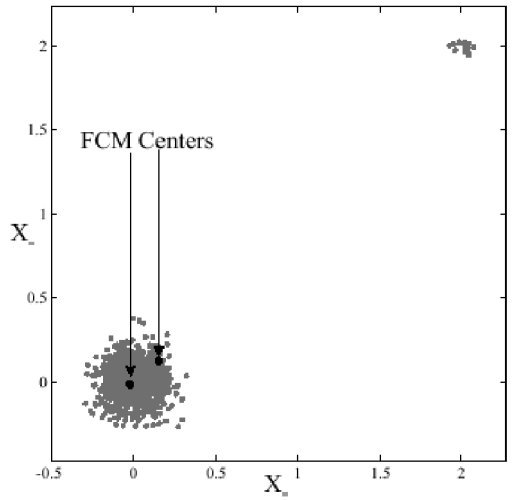

Рис. 1. Кластеризация данных с большими колебаниями плотности распределения

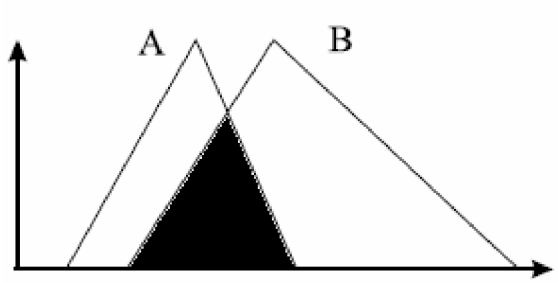

Рис. 2. Степень подобия множества A и B низка, однако степень включения A в B высока

Индекс Жаккарда [7] $S_{AB} = \frac{|A \cap B|}{|A \cup B|}$ является хорошим инструментом для измерения подобия нечетких множеств. Однако в кластеризации необходимо получить хорошо разделенные кластеры данных. Для этих целей предлагается использовать *степень включения*. На рис. 2. множество A с высокой степенью включено во множество B. Соответственно *индексу Жаккарда* эти два множества имеют низкую степень подобия. Для целей кластеризации множество A может рассматриваться как по-

добное с высокой степенью B, так как оно описывается в значительной степени также множеством B. Данное качество измеряется *степенью включения*. Пусть даны два нечетких кластера, степени принадлежности $u_i(x_k)$, $u_j(x_k)$ определены поточечно на X, тогда *нечеткая степень включения* определена следующим образом:

$$I_{ij} = \frac{\sum_{k=1}^{N} \min(u_{ik}, u_{jk})}{\sum_{k=1}^{N} u_{ik}}.$$

Заметим, что *степень включения* представляет отношение мощности пересечения двух нечетких множеств (в данном случае A и B), деленное на мощность одного из них (A). *Степень включения* является ассиметричной и может быть использована для конструирования симметричной *меры подобия*, которая назначает степень подобия $S_{ij} \in [0,1]$ по формуле

$$S_{ij} = \max(I_{ij}, I_{ji}),$$

где $S_{ij} = 1$ соответствует ситуации, когда $u_i(x_k)$ полностью включено в $u_j(x_k)$ или наоборот.

Граница α, которая определяет, будет ли происходить слияние, зависит от характеристик рассматриваемого множества данных, таких как, отделимость групп друг от друга, плотность (мера внутренней связи) кластера, размеры кластеров, и т.д, а также от параметров кластеризации, например меры нечеткости m. Мера подобия также зависит от других кластеров в разбиении ввиду того факта, что сумма степеней принадлежности каждого элемента данных к кластерам равна единице. Для случаев, когда определение границы α затруднительно, предлагается оценивать α как адаптивную границу, зависящую от количества кластеров в данных в любой момент времени. Эмпирическим путем установлено, что адаптивная граница работает наилучшим образом, когда ожидаемое количество кластеров в данных относительно мало (меньше 10).

Предлагается использовать $\alpha^{(l)} = \frac{1}{M^{(l)} - 1}$ как адаптивную границу, где $M^{(l)}$ — количество кластеров на итерации l. Смысл использования такой величины в качестве адаптивной границы заключается в следующем. Когда в начале алгоритма количество кластеров велико, величина $\alpha^{(l)}$ мала и слиянию подвергаются многие кластеры. С уменьшением количества кластеров растет адаптивная граница $\alpha^{(l)}$ и, следовательно, слиянию подвергается все меньше кластеров. Кластеры сливают только тогда, когда изменение максимального подобия кластеров от итерации к итерации превышает некоторую величину ε и подобие превышает границу α. Сливают только наиболее подобные кластеры и количество кластеров уменьшают на 1 при каждом слиянии. Алгоритм заканчивает свою работу, когда изменение элементов матрицы разбиения ниже, чем предопределенная граница ε.

В теории нечетких множеств используют индексы включения для сравнения нечетких множеств [8]. Индекс включения $I(A,B)$ определяется аксиоматически и удовлетворяет следующим свойствам:

1. $I(A,B) = 1 \Leftrightarrow \overline{A} \cup B = U$
2. $I(A,B) = 0 \Leftrightarrow A \cap B = \varnothing$
3. $I(A,B)$ зависит от $g(\overline{A} \cup B)$, где g - универсальная оценочная функция

В качестве универсальных оценочных функций могут выступать:

a) $g_{\min}(A) = \min_{x \in U} \{\mu_A(x)\}$

b) $g_{\Sigma}(A) = \frac{1}{|U|} \sum_{x \in U} \mu_A(x)$

c) $g_{1/2}(A) = \frac{1}{2}\left(\max_{x \in U} \{\mu_A(x)\} + \min_{x \in U} \{\mu_A(x)\}\right)$

Если $S(A) \cap S(B) = \varnothing$, то независимо от операции объединения $\overline{A} \cup B = \overline{A}$, следовательно $g(\overline{A} \cup B) \in [g(\overline{A}), 1]$, и $I(A,B)$ можно получить в нормированном виде:

$$I(A,B) = \frac{g(\overline{A} \cup B) - g(\overline{A})}{1 - g(\overline{A})}$$

Для исследования возможностей использования приведенных выше индексов было проведено исследование для треугольных функций принадлежности.

Мы рассмотрели два треугольных нечетких множества A, B (рис. 3) и выявили некоторые

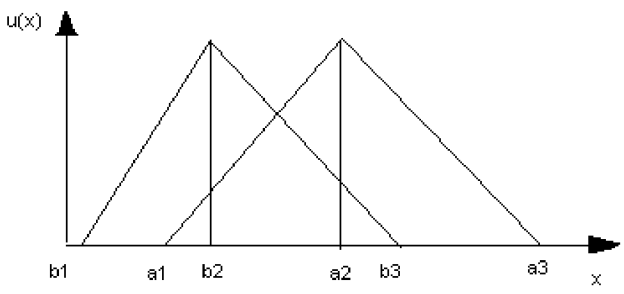

Рис. 3. Треугольные нечеткие множества A, B. A включено в B

свойства предложенной степени включения и индексов включения.

Обозначим через Sqw площадь фигуры. Тогда *степень включения* можно выразить как

$$I_{AB} = \frac{(b_3 - a_1)Sqw(A \cap B)}{(a_3 - a_1)Sqw(A)}.$$

Это следует из того, что

$$\min(u_A, u_B) = A \cap B,$$

$$\frac{1}{b_3 - a_1}\sum_{k=1}^{N}\min(u_{Ak}, u_{Bk}) \underset{N\to\infty}{\to} Sqw(A \cap B),$$

$$\frac{1}{a_3 - a_1}\sum_{k=1}^{N} u_{Ak} \underset{N\to\infty}{\to} Sqw(A)$$

То есть *степень включения* характеризует отношения площади пересечения двух нечетких множеств к площади одного из них в отличие от *индекса Жаккарда,* который представляет собой отношение площади пересечения к площади объединения двух нечетких множеств.

Исходя из геометрической интерпретации треугольных нечетких множеств A, B можно также предложить индекс включения следующего вида:

$$I_h = \begin{cases} \dfrac{(b_3 - a_1)h}{(a_3 - a_1)}, & a_3 > b_3 > a_1, \\ \dfrac{(a_3 - b_1)h}{(a_3 - a_1)}, & a_3 > b_1 > a_1, \\ 1 & (b_1 < a_1 < a_3 < b_3), \\ \dfrac{b_3 - b_1}{a_3 - a_1}, & (a_1 < b_1 < b_3 < a_3), \end{cases}$$

где h — высота треугольника, соответствующего пересечению множеств (рис. 4) и выражается следующим образом:

$$h = \frac{b_3 - x_h}{b_3 - b_2},\ x_h = \frac{b_3 a_2 - b_2 a_1}{b_3 - b_2 + a_2 - a_1},$$

если $a_3 > b_3 > a_1$ и

$$h = \frac{x_h - b_1}{b_2 - b_1},\ x_h = \frac{a_3 b_2 - b_1 a_2}{a_3 - a_2 + b_2 - b_1},$$

если $a_3 > b_1 > a_1$. Данный индекс включения, по сути, представляет собой то же самое, что и степень включения, в случае, если рассматривается большое число элементов данных. С другой стороны, данный индекс представляет собой конкретизацию степени включения для треугольных функций принадлежности и тем самым является более удобным и вычислительно менее затратным способом измерения степени включения.

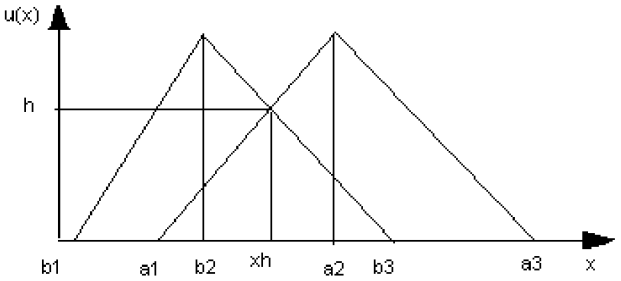

Рис. 4. Основная идея индекса включения I_h

Мы также исследовали поведение индекса включения, основанного на использовании универсальной оценочной функции $g_{\min}(A) = \min_{x\in U}\{\mu_A(x)\}$. То есть мы рассмотрели, функцию $g_{\min}(\overline{A} \cup B)$ и с помощью геометрической интерпретации получили следующий результат:

$$g_{\min}(\overline{A} \cup B) = \begin{cases} 0, & b_3 \le a_2 \\ \overline{h}, & b_3 > a_2 \end{cases},$$

где $\overline{h} = u_{\overline{A}}(x) = u_B(x)$, то есть $\overline{h}$ — это значение функций принадлежности в точке их пересечения (рис. 5).

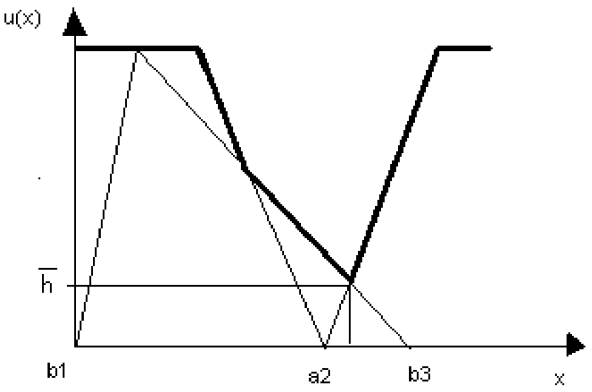

Рис. 5. Геометрический расчет индекса включения, основанного на оценочной функции $g_{\min}$

Таким образом, индекс включения, основанный на использовании универсальной оценочной функции $g_{\min}$ распознает только достаточно большую степень включения одного нечеткого множества в другое. Заметим, что для оценочной функции $g_{\min}$, как мы показали, не выполняется аксиома 2 для индекса включения.

Для индекса включения I_h в сравнении с индексом, основанным на оценочной функции $g_{\min}$, характерно следующее поведение:

$I_h > (a_2 - a_1)h$, если $b_3 > a_1$ и $I_h \le (a_2 - a_1)h$, если $b_3 \le a_1$. Значение $I_h = 0$ достигается только в случае отсутствия пересекающихся областей.

Оценочная функция $g_\Sigma(A)$ принимает значение площади фигуры под функцией принадлежности нечеткого множества A.

Эксперименты показали следующие результаты применения индексов включения, а также рассмотренной выше степени включения к оценке включения двух нечетких множеств A, B, которые относительно друг друга расположены как и множества на рис.2.

$$u_A(x) = \begin{cases} 0 & (x \le 0) \vee (x \ge 6) \\ 1/5\, x & x \in [0,5] \\ 6 - x & x \in [5,6] \end{cases}$$

$$u_B(x) = \begin{cases} 0 & (x \le 1) \vee (x \ge 10) \\ \dfrac{x-1}{5} & x \in [1,6] \\ \dfrac{10-x}{4} & x \in [6,10] \end{cases}$$

В таблице представлены результаты вычислений степени включения нечеткого множества A в B, степени включения нечеткого множества B в A и соответственно их меры подобия $S_{AB} = \max(I_{AB}, I_{BA})$.

ВЫВОД

Все использованные нами способы измерения степени включения показали, что в данном случае степень включения нечеткого множества A в нечеткое множество B не менее 0.5, а также, что степень включения нечеткого множества B в нечеткое множество A меньше, чем степень включения A в B. Данный результат адекватен по отношению к тестируемым данным.

Таким образом, индексы включения с использованием оценочных функций $g_\Sigma(A)$, $g_{1/2}(A)$, $g_{\min}(A)$, а также индекс включения I_h также являются хорошим инструментом для измерения степени включения нечетких множеств, и их можно рекомендовать для использования в описанном выше алгоритме.

Описанные здесь модификации алгоритма FCM применяют для алгоритмов, основанных на использовании целевой функции. Введение объемных прототипов ослабляет влияние различий в распределении и плотности данных на результаты кластеризации. Слияние кластеров на основе подобия помогает определить подходящее количество кластеров. Усовершенствованный алгоритм С-средних более устойчив по отношению к начальным данным и, тем самым, уменьшается зависимость кластеризации от инициализации данных.

ЛИТЕРАТУРА

1. *Kaymak U.* Extended fuzzy clustering algorithms / U. Kaymak, M. Setnes // Erim Report Series Research in Management, November 2000, 24 p.

2. *Bezdek J.* Pattern Recognition with Fuzzy Objective Function. Plenum Press, New York, 1981.

3. *Gath I.* Unsupervised optimal fuzzy clustering / I. Gath, A. B. Geva //IEEE Transactions on Pattern Analysis and Machine Intelligence, 11(7):773—781, July 1989. 22

4. *Kaymak U.* Compatible cluster merging for fuzzy modeling/ U. Kaymak. and R. Babuška // Proceedings of the Fourth IEEE International Conference on Fuzzy Systems, vol. 2, P. 897—904, Yokohama, Japan, Mar. 1995.

5. *Kaymak U.* Methods for simplification of fuzzy models. // U. Kaymak., R. Babuška, M. Setnes, H. B. Verbruggen, H. R. van Nauta Lemke // Intelligent Hybrid Systems, P/ 91–108. Kluwer Academic Publishers, Boston, 1997.

6. G*ustafson E.* Fuzzy clustering with a fuzzy covariance matrix/ E. Gustafson, W. Kessel // Proc. IEEE Conf. Decision Control, P. 761—766, 1979.

Таблица

Степень включения множества А в В, В в А и степень сходства данных множеств с использованием различных индексов включения

	На основе индекса включения	На основе $g_{\min}$	На основе $g_{1/2}$	На основе g_Σ	На основе I_h
Степень включения А в В	0.6967	0.5000	0.5000	0.5082	0.6900
Степень включения В в А	0.5667	0.0250	0.2250	0.4133	0.4600
Сходство В и А	0.6967	0.5000	0.5000	0.5082	0.6900

7. *Kaymak U.* Compatible cluster merging for fuzzy modeling/ U. Kaymak, R. Babuška//Proceedings of the Fourth IEEE International Conference on Fuzzy Systems, vol. 2, P. 897—904, Yokohama, Japan, Mar. 1995.

8. *Леденева Т.М.* Обработка нечеткой информации / Воронеж: Воронежский государственный университет, 2006. — 233 с.

Статья принята к опубликованию
25 декабря 2006 г.