УДК 004(075.32)

О МОДИФИКАЦИИ МЕТОДА БЛИЖАЙШИХ СОСЕДЕЙ

Р. К. Стрюков, А. И. Шашкин

Воронежский государственный университет

Поступила в редакцию 12.02.2015 г.

Аннотация. В статье рассматриваются возможные модификации метода ближайших соседей на основе различных функций расстояния и правил принятия решений об отнесении объекта к определенному классу.

Ключевые слова: кластеризация, функция расстояния, ближайший сосед.

Annotation. This article discussesthe possiblemodification of the methodof nearest neighborsbased on differentdistance functionsanddecision rulesto classifyan objectto a class.

Keywords: clustering,distance function, nearest neighbor.

1. ВВЕДЕНИЕ

Задача кластеризации заключается в разбиении заданного множества объектов на группы (классы, кластеры) в определенном смысле однородных объектов, при этом предполагается, что данные внутри групп схожи между собой, а схожесть данных, принадлежащих различным группам, мала.Важное место кластерный анализ занимает в тех областях, которые связаны с изучением массовых явлений и процессов. Кластеризацию применяют для эффективного сжатия и хранения данных, поиска в базах данных, сравнения изображений.

В настоящее время существует значительное количество методов классификации/кластеризации [1, 2, 5, 6, 9, 13, 14, 17], при этом не существует метода, который был бы применим к данным, имеющим произвольную природу. Практически все методы опираются на ряд предположений, среди которых основными являются следующие.

1. В некоторых методах изначально задается количество кластеров, при этом может оказаться, что некоторые кластеры не поддаются содержательной интерпретации. Количество кластеров должно соответствовать количеству естественных подструктур, присутствующих в данных, поэтому адекватность кластера должна оцениваться отдельно уже после кластеризации данных на основе критериев качества классификации. Другой подход определения истинного количества кластеров – это слияние кластеров до тех пока не будет определено «правильное» разбиение [16].

2. Проблема устойчивости кластеризации становится особо острой, когда есть кластеры с изменяющейся плотностью распределения данных и различными объемами. Данная проблема обусловлена чувствительностью к инициализации. Неправильная инициализация данных может привести к неправильному разбиению [7, 8 ,11].

3. Для алгоритмов, использующих прототипы, форма кластеров определяется используемой функцией расстояния. Например, алгоритм С-средних [2, 5, 6] использует евклидово расстояние, и поэтому его следует применять, если можно предположить, что кластеры имеют сферическую форму. В некоторых алгоритмах предусмотрена адаптация функции расстояния к форме кластеров, как это сделано в алгоритме Густавсона-Кесселя [5]. Другим способом повлиять на форму кластеров является выбор прототипов с геометрической структурой [15, 16]. Таким образом, выбор функции расстояния является важнейшим этапом решения задачи кластеризации, при этом важно предусмотреть наличие адаптивных свойств у этой функции.

---

4. В настоящее время существует значительное число критериев качества кластеризации [4, 5, 15]. Выбор критерия, адекватного содержательной постановке задачи, определяет правильность ее решения.

Необходимость развития методов кластеризации и их использования заключается, прежде всего, в том, что такие методы помогают построить научно обоснованные классификации, выявить внутренние связи между единицами наблюдаемой совокупности. Объектом исследования данной статьи является метод ближайших соседей [4], идея которого может быть реализована для различных типов данных. Отличительной особенностью метода является использование весовых коэффициентов при расчете расстояния, что обеспечивает наличие адаптивных свойств функции расстояния. Также предлагается модификация данного метода на основе использования различных правил принятия решений в рамках классифицирующей процедуры.

1. ПОСТАНОВКА ЗАДАЧИ

Пусть X – множество объектов, каждый из которых $x \in X$ характеризуется определенными признаками $\left\{p_i\right\}_{i=\overline{1,n}}$, а каждый признак p_i оценивается в соответствующей шкале S_i. Пространством признаков называется множество $S = S_1 \times \ldots \times S_n$, тогда каждому объекту x соответствует векторная оценка $s^x = \left(s_1^x, \ldots, s_n^x\right) \in S$.

Под *кластеризацией* подразумевается разбиение множества векторных оценок $\left\{s^x\right\}_{x \in X}$ на K групп (кластеров) $G_1, \ldots, G_K$, так что $G_i \bigcap G_j = \varnothing$ для $i \neq j$ и $\bigcup_{j=1}^{K} G_j = X$, при этом минимизируется выбранная функция качества, которая в целом оценивает «правильность» кластеризации.

В большинстве методов в качестве такой функции выступает суммарное среднеквадратическое отклонение векторных оценок объектов от центров кластеров, вокруг которых они группируются. По сути, в этом случае функция качества определяет, насколько классы являются компактными в пространстве S, при этом «компактность понимается в визуальном смысле».

Таким образом, один из критериев – это функция вида

$$f\left(c_1, \ldots, c_K\right) = \sum_{j=1}^{K} \sum_{x \in G_j} \left\| s^x - c_j \right\|^2, \qquad (1)$$

где c_j – центр кластера G_j.

Цель кластеризации – найти вектор центров $c = \left(c_1, \ldots, c_K\right)$ и соответствующее разбиение $\left\{G_1, \ldots, G_K\right\}$, минимизирующее (1).

Пусть в результате кластеризации сформированы кластеры $\left\{G_1, \ldots, G_K\right\}$. *Индикаторной функцией* будем называть функцию $g : X \to \left\{1, \ldots, K\right\}$, которая ставит в соответствии каждому объекту $x \in X$ индекс i того класса G_i, которому он принадлежит, т. е. $g(x) = i$, если $x \in G_i$.

Задачу классификации можно рассматривать как задачу нахождения такой индикаторной функции, которая обеспечивает минимум критерия (1).

Заметим, что, как правило, все компоненты векторной оценки (частные оценки) относятся к одному типу, при этом можно выделить несколько основных типов данных [10]. Однако возможна ситуация, когда в векторной оценке присутствует несколько типов данных [18], т. е. информация является смешанной. И в том, и в другом случае важно определить функцию расстояния в S и наделить данное множество свойствами метрического пространства.

2. О СВОЙСТВАХ НЕКОТОРЫХ РАССТОЯНИЙ

Для того, чтобы определить принадлежность объекта $x \in X$ к классу G_i необходимо ввести понятие меры близости.

Меру близости можно ввести несколькими способами: с помощью функции подобия [11], на основе отношений несходства и различия [7, 8], на основе специальных индексов [5], с помощью аксиоматического подхода.

Пусть x и y – элементы заданного множества V. Функцией расстояния называется неотрицательная действительная функция $d : V \times V \to R^+ \cup \{0\}$, которая удовлетворяет следующим условиям[4]:

а) $d(x,y)\geq 0$, причем $d(x,y)=0$ тогда и только тогда, когда $x=y$;

б) $d(x,y)=d(y,x)$;

в) $d(x,y)\leq d(x,z)+d(z,y)$, где $x,y,z\in V$.

Заметим, что x и y могут иметь различную природу. Это могут быть векторы $\left(x_1,\ldots,x_n\right)$ и $\left(y_1,\ldots,y_n\right)$ с числовыми компонентами, и тогда для нахождения расстояния можно воспользоваться формулой Минковского [4]

$$d_p(x,y)=\|x-y\|_p=\left(\sum_{i=1}^{n}(x_i-y_i)^p\right)^{\frac{1}{p}}= \qquad (2)$$

$$=\sqrt[p]{\left|x_1-y_1\right|^p+\ldots+\left|x_n-y_n\right|^p}$$

для различных значений p.

В частности, при $p=1$ получим манхэттэновское расстояние

$$d_1(x,y)=\|x-y\|_1=\sum_{i=1}^{n}(x_i-y_i)= \qquad (3)$$

$$=\left|x_1-y_1\right|+\ldots+\left|x_n-y_n\right|,$$

При $p=2$ имеем расстояние Евклида

$$d_2(x,y)=\sqrt{\sum_{i}^{n}\left(x_i-y_i\right)^2}, \qquad (4)$$

при $p\to\infty$ получается «равномерная» метрика

$$d(x,y)=\|x-y\|_\infty=\max_{i=1,n}\left|x_i-y_i\right|. \qquad (5)$$

К другим известным типам расстояния относятся:

а) расстояние Махаланобиса

$$d_{M^{-1}}(x,y)=\|x-y\|_{M^{-1}}=\sqrt{(x-y)^T M^{-1}(x-y)}= \qquad (6)$$

$$=\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i-y_i)M_{ij}^{-1}(x_j-y_j)},$$

где $M=\left\{m_{ij}\right\}_{n\times n}$ – ковариационная матрица векторов обучающей выборки;

б) метрика Канберра при $x\neq 0$ или $y\neq 0$

$$d_k(x,y)=\sum_{i=1}^{n}\frac{\left|x_i-y_i\right|}{\left|x_i\right|+\left|y_i\right|}. \qquad (7)$$

Для того, чтобы учесть важность признака в функцию расстояния вводятся весовые коэффициенты $w_i\in[0,1]$ и $\sum_{i=1}^{n}w_i=1$, тогда формула (2) примет следующий вид:

$$d_p(x,y)=\left(\sum_{i=1}^{n}w_i(x_i-y_i)^p\right)^{\frac{1}{p}}= \qquad (8)$$

$$=\sqrt[p]{w_1\left|x_1-y_1\right|^p+\ldots+w_n\left|x_n-y_n\right|^p}.$$

Особого внимания заслуживает случай, когда компоненты векторных оценок не являются количественными. В [10] приводятся основные типы данных и функций расстояния между ними. В [9] рассматривается случай, когда данные являются лингвистическими. Если в векторной оценке представлены данные нескольких типов, то выделяя их и выбирая подходящие функции расстояния можно построить обобщенную функцию расстояния с весовыми коэффициентами, которые отражают значимость того или иного типа информации. Существуют также функции расстояния, которые учитывают влияние окружающих или соседних точек, называемых «контекстом» [5].

Функция расстояния – важнейший инструмент для решения задачи кластеризации. Можно сформулировать следующие особенности, которые должны учитываться при выборе функции расстояния:

a) при использовании (4) на расстояние могут сильно влиять различия между масштабами осей, по координатам которых вычисляется это расстояние;

b) на величину расстояния (4) оказывают влияние отдельные большие выбросы, в то время как это влияние может быть уменьшено, если вместо (4) использовать (3);

c) расстояние (5), которое иначе называется чебышевским, стоит использовать, если работа осуществляется с объектами, особенно различающимися по какой-либо одной координате;

d) евклидово расстояние (4) следует применять, если априори известно, что кластеры имеют сферическую форму, также это расстояние стоит использовать, когда множество данных состоит из «компактных» или «изолированных» кластеров;

e) если имеет смысл учитывать статическую зависимость между признаками, по которым характеризуются объекты, то целесообразно использовать метрику (6), которая приписывает различные веса признакам, основанные на их вариации и попарной корреляции;

f) недостатком расстояния Минковского является то, что признаки с наибольшей шкалой доминируют остальные, поэтому используют нормализацию данных и различные весовые схемы.

g) расстояние Махаланобиса используют для получения кластеров в форме гиперэллипсоидов.

Нормализация данных осуществляется различными способами, наиболее распространенными из которых являются следующие [5]:

$$x_i' = \frac{x_i - \min_i\{x_i\}}{\max_i\{x_i\} - \min_i\{x_i\}}, \tag{9}$$

$$x_i' = \frac{x_i - m}{\delta}, \tag{10}$$

где x_i – «старое» значение, x_i' – новое значение, m – среднее выборочное значение i-й координаты, δ – выборочное среднеквадратичное отклонение для i-й координаты.

Вычисляя расстояния между каждой парой объектов, можно получить матрицу расстояний, которая интерпретируется как матрица отношения несходства. Некоторые алгоритмы работают с матрицей сходства (например, нечеткие алгоритмы кластеризации, основанные на нечетких отношениях [7, 8]).

3. МЕТОД БЛИЖАЙШИХ СОСЕДЕЙ И ЕГО МОДИФИКАЦИЯ

Данный метод подробно описан в [4] и основан на вычислении расстояния до k ближайших соседей. Число k является параметром метода и влияет на результаты классификации. Так если $k=1$, то алгоритм теряет обобщающую способность, и решение об отношение некоторого объекта к классу будет приниматься на основе единственного ближайшего соседа (метод ближайшего соседа), а не нескольких представителей класса. Если же параметр k установить слишком большим, то алгоритм перестанет реагировать на локальные особенности объекта. Таким образом, в зависимости от прикладной задачи необходимо установить оптимальное значение параметра.

Итак, пусть получено разбиение на классы $\{G_1,\dots,G_N\}$ и для данного нового объекта x^* определены k ближайших соседей, которые в наибольше степени с ним схожи. Известно множество расстояний $\{d_1,\dots,d_k\}$ до каждого из соседей $x_{i_1},\dots,x_{i_k}$, , при этом известно, к какому классу принадлежат соседи-объекты. Необходимо определить правило принятия решения об отнесении объекта x^* к некоторому классу.

Данную задачу можно представить как задачу группового выбора на основе голосования [17]: r-ым голосом считается номер класса, к которому относится r-ый ближайший сосед.

Таким образом, k ближайшим соседям соответствует выборка длины k, каждый элемент которой – номер класса. Введем различные характеристики классов и ближайших соседей объекта x^*.

Пусть из k ближайших соседей n_1 относятся к классу G_1, n_2 – к классу $G_2,\dots$, n_N – к классу G_N, так что $n_1+\dots+n_N=k$. *Весом класса* G_r назовем величину $w_r=\frac{n_r}{k}$ – относительное количество объектов в классе.

Рангом $rang(G_r)$ класса G_r назовем порядковый номер класса в ранжировании классов по невозрастанию весов w_r (или количеству ближайших соседей объекта x^*).

Средним расстоянием до класса G_R назовем величину

$$\bar{d}\left(x^*,G_r\right)=\frac{1}{n_r}\sum_{\{y:y\in G_r\}} d\left(x^*,y\right). \tag{11}$$

Кратчайшим расстоянием $\rho_{\min}^{G_r}$ до класса G_r назовем минимальное расстояние от x^* до ближайших соседей из класса G_r, т. е.

$$\rho_{\min}^{G_r}=\min_{\{y:y\in G_r\}} d\left(x^*,y\right). \tag{12}$$

Вкладом объекта x^* в класс G_r называется величина

$$contrib\left(G_r\right)=\sum_{\{y:y\in G_r\}}\frac{1}{d^2(x^*,y)}. \tag{13}$$

Сформулируем несколько типов правил об отнесении объекта x^* к некоторому классу.

1) *правило простого большинства* [17]: новый объект следует отнести к тому классу G_{r^*}, объекты которого занимают в выборке более половины мест, т. е. $w_{r^*}=\frac{n_{r^*}}{k}\ge 0.5$;

2) *правило относительного большинства* [17]: новый классифицируемый объект будет отнесен к тому классу, элементов которого окажется больше в выборке из k ближайших соседей, т. е. к тому классу, который наберет наибольшее количество голосов

$$r^* = \arg\max_r \left\{ n_r \right\}; \qquad (14)$$

3) *правило взвешенного большинства* заключается в выборе такого класса r^*, номер которого определяется формулой

$$r^* = \arg\max_r \left\{ \frac{w_r}{\rho_{\min}^{G_r}} \right\}; \qquad (15)$$

4) *правило среднего*: новый объект относится к тому классу G_{r^*}, до которого среднее расстояние минимально, т. е.

$$r^* = \arg\max_r \overline{d}\left(x^*, G_r \right); \qquad (16)$$

5) *правило, основанное на критериях качества классификации*, заключается в том, что объект относится к тому классу, чтобы вновь полученное разбиение оптимизировало выбранный критерий качества. Например, одним из критериев является компактность классов, поэтому считая, что чем ближе объект класса к объекту x^*, тем больше он вносит вклад в классификацию, можно предложить следующее правило классификации: объект x^* относится к тому классу, для которого его вклад $contrib\left(G_r\right)$ максимальный.

Заметим, что если расстояния от x^* до каждого из ближайших соседей приблизительно одинаковы, то можно воспользоваться правилами простого или относительного большинства. Если ближайшие соседи разбиваются на классы приблизительно одинаковой мощности, то, наоборот, имеет смысл использовать те правила, которые в большей степени учитывают расстояния (например, правило среднего). Правило взвешенного большинства учитывает и количество объектов в классе (через весовой коэффициент w_r), и расстояние до этих объектов. Если классов достаточно много и веса w_r приблизительно одинаковы, то вначале осуществляется выборка таких классов, чтобы их суммарный вес был больше, чем 0.5, а затем применяется одно из правил.

Учитывая различные критерии качества классификации, можно получить и другие правила.

4. ВЫЧИСЛИТЕЛЬНЫЙ ЭКСПЕРИМЕНТ

Проведем сравнительный анализ различных правил на известном примере классификации ирисов Фишера [3]. Выборка цветов состоит 150 штук и содержит 3 вида цветов: IrisSetosa, IrisVersicolour, IrisVirginica (по 50 штук каждого вида). Считается, что каждый цветок описывается двумя характеристиками – это длина чашелистика и длина лепестка. Считается, что длина лепестка вдвое важнее длины чашелистика. Необходимо классифицировать новый цветок со значениями длины чашелистика 6,1 см и лепестка 4,8. Расположение нового цветка, среди известных, представлено на рис. 1.

Находим трех ближайших соседей: *A*(6,1; 4,7) – IrisVersicolour, *B*(6; 4,8), *C*(6,2 4,8) – IrisVirginica. Заметим, что по длине лепестка новый цветок относится к виду IrisVirginica, а по длине чашелистика он ближе к виду IrisVersicolour. Расстояния от нового цветка до каждого из классов представлены в табл. 1.

Анализ табл. 1 показывает, что согласно правилам простого, относительного и взвешенного большинства, цветок должен быть отнесен к виду IrisVirginica.

Вычислим величины

$$contrib(IrisVersicolour) = 50,$$

$$contrib(IrisVirginica) = 200.$$

Так как

$$contrib(IrisVirginica) > contrib(IrisVersicolour),$$

то, согласно правилу, основанному на критерии компактности, получим, что цветок классифицируется как IrisVirginica.

ЗАКЛЮЧЕНИЕ

Несмотря на простоту реализации, данный метод показывает хорошие результаты при классификации. Среди его преимуществ перечислим следующие:

– гибкость в выборе правила отнесения нового объекта к одному из классов, что позволяет оптимизировать алгоритм;

– объясняющие способности и возможность интерпретации результатов классификации;

– простота программной реализации.

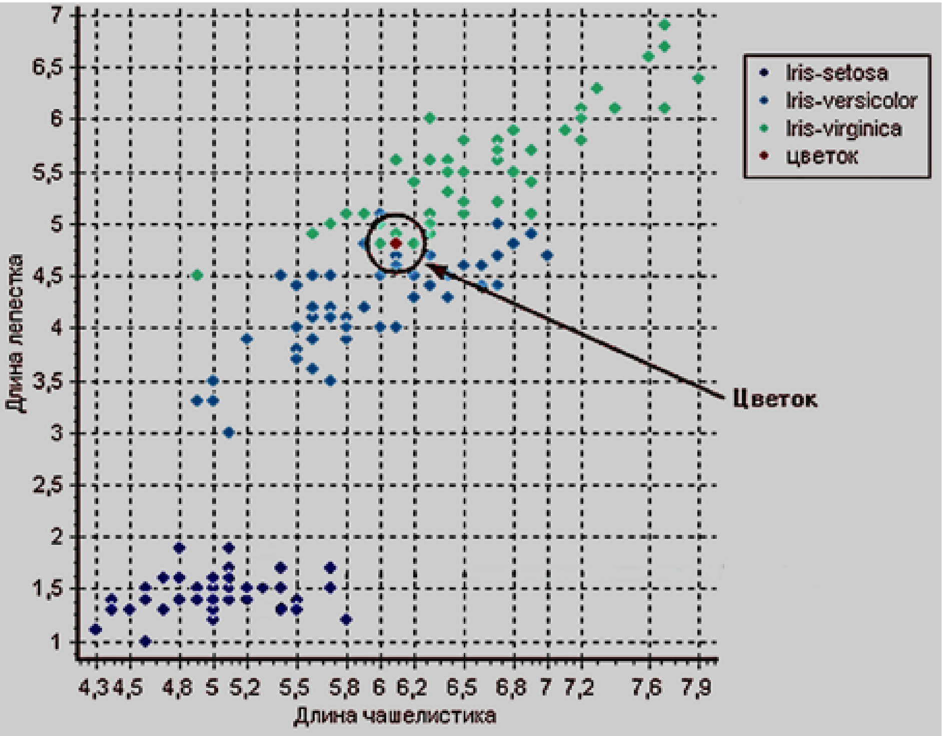

Рис. 1. Размещение цветка среди уже известных цветов

Таблица 1

Три ближайших соседа для цветка

Запись	Длина чашелистика	Длина лепестка	Расстояние	Класс
Цветок	6,1	4,8	–	
A	6,1	4,7	0,14	IrisVersicolour
B	6	4,8	0,1	IrisVirginica
C	6,2	4,8	0,1	IrisVirginica

Вместе с тем качество работы метода ближайших соседей зависит от репрезентативности данных и их качества. Перспективным направлением развития метода ближайших соседей является использование критериев качества классификации.

СПИСОК ЛИТЕРАТУРЫ

1. *Айвазян С. А.* Классификация многомерных наблюдений / С. А. Айвазян, З. И. Бежаева, О. В. Староверов. – М. : Статистика, 1974. – 238 с.

2. *Бююль А.* SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей. – СПб. : ДиаСофтЮП, 2002. – 608 с.

3. Википедия, свободная энциклопедия [электронный ресурс]. URL: https:// ru.wikipedia.org/wiki/ Ирисы Фишера.

4. *Воронцов К. В.* Лекции по метрическим алгоритмам классификации [электронный ресурс]. URL:http:// www.ccas.ru/frc/papers/ voron04mpc.pdf.

5. *Гайдышев И. П.* Анализ и обработка данных: специальный справочник / И. П. Гайдышев. – СПб. : Питер, 2001. – 762 с.

6. *Дюк В. А.* DataMining: учебный курс / В. А. Дюк, А. С. Самойленко. – СПб. : Питер, 2001. – 368 с.

7. *Каплиева Н. А.* Исследование различных типов транзитивности в приложении к нечеткой классификации / Н. А. Каплиева, Т. М. Леденева// Вестник Воронежского государственного университета. Серия: Физика. Математика, 2006. – № 2. – С. 206–216.

8. *Леденева Т. М.* Влияние различных типов транзитивности на декомпозиционное дерево в задаче нечеткой классификации / Т. М. Леденева, Н. А. Каплиева // Нечеткие

системы и мягкие вычисления, 2013. – Т. 8. – № 1. – С. 5–25.

9. *Леденева Т. М.* Алгоритм нечеткой классификации для объектов с оценками в лингвистической шкале / Т. М. Леденева, К. С. Погосян, НгуенНгок Хуи // Системы управления и информационные технологии, 2012. – Т. 49. – № 3. – С. 20–23.

10. *Леденева Т. М.* О представлении информации в задачах классификации / Т. М. Леденева, НгуенНгок Хуи // Вестник Воронежского государственного технического университета, 2012. – Т. 8. – № 7–1. – С. 33–37.

11. *Леденева Т. М.* Влияние функции подобия на результаты нечеткой кластеризации / Т. М. Леденева, НгуенНгок Хуи // Информационные технологии, 2011. – №11. – С. 14–20.

12. *Леденева Т. М.* О свойствах нечеткого отношения сходства / Леденева Т. М., Р. К. Стрюков // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии, 2014. – № 4. – С. 75–79.

13. *Лепский А. Е.* Математические методы распознавания образов / А. Е Лепский, А. Г. Броневич. – Таганрог, 2009.

14. *Моисеев С. А.* Эволюционный алгоритм для решения задачи нечеткой кластеризации / С. А. Моисеев, Т. М. Леденева // Вестник Воронежского государственного технического университета, 2012. – Т. 8. – № 2. – С. 4–8.

15. *Тарасова А. С.* Методы определения оптимальной геометрической формы в задачах кластерного анализа / А. С. Тарасова // Информационные технологии, 2007. – № 11. – С. 14–21.

16. *Тарасова А. С.* Модификация алгоритма С-средних на основе использования объемных прототипов и слияния схожих кластеров / А. С. Тарасова // Вестник Воронежского ун-та. Серия: Системный анализ и информационные технологии. – 2007. – № 1. – С. 68–74.

17. *Робертс Ф. С.* Дискретные математические модели с приложениями к социальным, биологическим и экономическим задачам / Ф. С. Робертс. – М. : Наука, 1986. – 496 с.

18. *Yang M. S.* Fuzzy clustering algorithms for mixed feature variables / M. S. Yang, P. Y. Hwang, D. H. Chen // Fuzzy Set and Systems, 2004. – № 141. – Pp. 301–317.

Шашкин Александр Иванович – доктор физико-математических наук, профессор, заведующий кафедрой математического и прикладного анализа, Воронежский государственный университет.
E-mail: shashkin@amm.vsu.ru

Стрюков Руслан Константинович – аспирант кафедры математического и прикладного анализа, Воронежский государственный университет.
E-mail: 79204605031@ya.ru

Shashkin Aleksandr Ivanovich – Doctor of Physical and Mathematical Sciences, Professor, Head of Department ofMathematicsandApplied Analysis, Voronezh State University.
E-mail shashkin@amm.vsu.ru

Ctryukov Ruslan K. – Post-graduate student of the Department of Mathematics and Applied Analysis, Voronezh State University.
E-mail 79204605031@ya.ru