УДК 681.3.06 (075.32)
DOI: 10.15827/0236-235X.132.599-604

Дата подачи статьи: 24.09.20
2020. Т. 33. № 4. С. 599–604

Разработка теоретических основ классификации и кластеризации нечетких признаков на основе теории категорий

К.Д. Русаков [1], *научный сотрудник, rusakov.msk@yandex.ru*
Д.Е. Селиверстов [2], *к.т.н., ведущий научный сотрудник, seliverstov_dmitriyy@rambler.ru*
С.Ш. Хиль [3], *к.т.н., доцент, skhill@mail.ru*
С.Б. Савилкин [4], *к.ф.-м.н., доцент, старший научный сотрудник, Savilkin@mail.ru*

[1] *Институт проблем управления им. В.А. Трапезникова РАН, г. Москва, 117997, Россия*
[2] *Российский экономический университет им. Г.В. Плеханова, г. Москва, 117997, Россия*
[3] *Московский авиационный институт (национальный исследовательский университет), г. Москва, 125993, Россия*
[4] *Центр визуализации и спутниковых информационных технологий ФНЦ НИИСИ РАН, г. Москва, 117218, Россия*

В статье дается обоснование выбора меры неопределенности сведений. Описывается современный подход, основанный на применении фундаментальных алгебраических конструкций теории категорий. Особенностью множества отношений эквивалентности является непосредственное (прямое) установление отношения эквивалентности между объектом и классом.

Показано, что в настоящее время существует ряд актуальных прикладных задач в области классификации, требующих иного подхода к установлению отношения эквивалентности – использования модели каскадного фильтра с промежуточными состояниями. Для обоснования меры неопределенности об объекте предлагается использовать теоретические положения на основе математического аппарата теории ультраоператоров. Данный аппарат также оперирует сведениями в терминах определений неэлементарных сведений.

К особенностям рассматриваемого аппарата можно отнести следующие: предложение оперировать не сведениями, а их неопределенностями, не рассматриваемыми в аппарате ультраоператоров; в некоторых задачах рассматриваются элементарные сведения, что является частным случаем в аппарате ультраоператоров и облегчает вычисления; область применения сужается до чисел (то есть сведения-множества могут быть только числовой природы, компактами, в том числе многомерными); оперирование числовыми множествами-сведениями в некоторых случаях исключает необходимость применения в явном виде решетки (и соответствующих шкал) понятий и позволяет оперировать в неявном виде с бесконечными решетками.

Предлагаемый авторами подход и представленные математическая модель и мера информационной неопределенности являются составной частью разрабатываемого метода классификации и кластеризации состояний сложных систем на основе теоретико-множественного подхода и позволяют рассматривать процесс получения четких классов с точки зрения снижения информационной энтропии с использованием каскадного фильтра.

Ключевые слова: *классификация, кластеризация, категория, функтор, информационная энтропия.*

Необходимость совершенствования математического аппарата теории классификации уже рассматривалась ранее [1]. Авторами предложен современный подход, основанный на применении фундаментальных алгебраических конструкций теории категорий [2]. Адекватность выбора данной теории обусловлена ее фрагментарным использованием в ряде работ, посвященных решению задачи классификации [3]. При этом введен ряд определений, таких как однозначно не идентифицируемый объект A, категория однозначно не идентифицируемых объектов $\tilde{A}$, ковариантный функтор однозначно не идентифицируемых объектов F, дуальная категория однозначно не идентифицируемых объектов $\tilde{A}^d$. В основу разрабатываемых теоретических положений включена система аксиом Бернайса–Геделя.

В известной литературе предложен ряд классических схем отнесения исследуемых

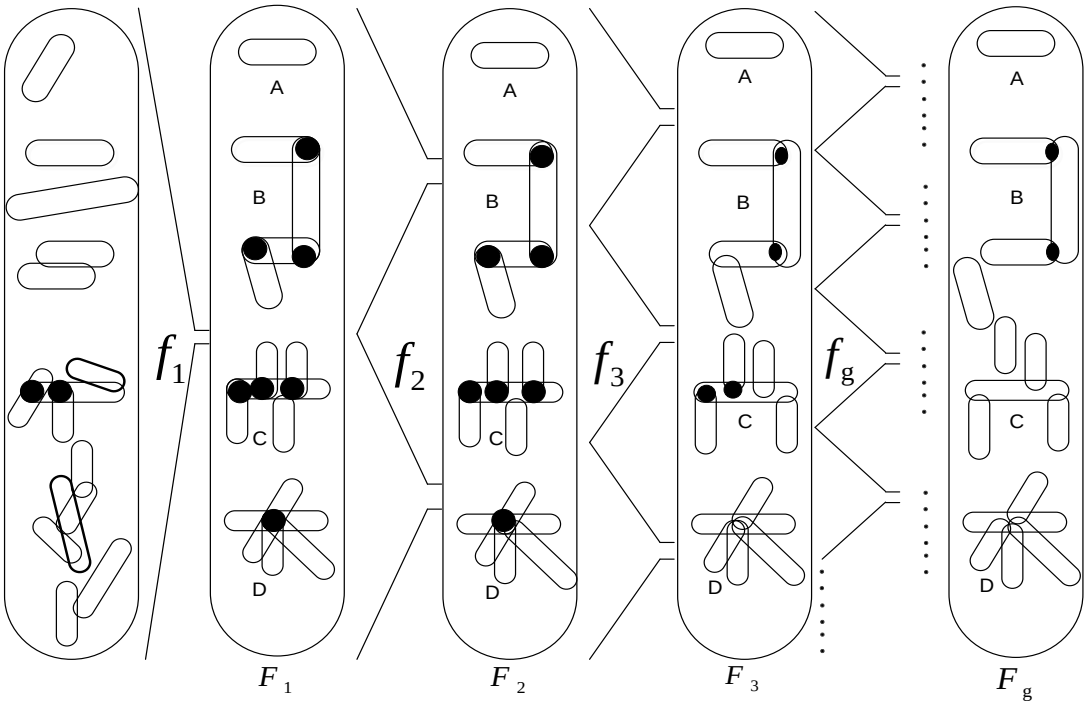
объектов к тому или иному классу [3]. Пусть на категории однозначно не идентифицируемых объектов $\tilde{A}$ определены некоторое множество X_n, $n=\overline{1,m}$, и множество их классов толерантности K_P^T, $p=\overline{1,q}$. При этом $K_P^T \subset F_\upsilon$, $\upsilon=\overline{1,g}$ – множество промежуточных состояний; f = {f_1, f_2, f_3, …, f_{g-1}} – отношения эквивалентности (морфизмы) F_υ. Их особенностью является непосредственное (прямое) установление отношения эквивалентности между объектом X_n, $n=\overline{1,m}$, и классом K_p, $p=\overline{1,q}$, где $K_p \subseteq F$. Однако в настоящее время существует ряд актуальных прикладных задач в области классификации, требующих иного подхода к установлению отношения эквивалентности, а именно: вместо прямого отношения предлагается использовать модель каскадного фильтра с промежуточными состояниями, представленную на рисунке.

При установлении f_1 определяется соответствие исследуемого элемента некоторому классу толерантности K_P^T. При этом K_P^T может четко идентифицировать объект, тогда $K_P^T = K_P^\ni$ есть класс эквивалентности (класс *A* на рисунке). В противном случае K_P^T есть пересечение классов идентифицируемых объектов (классы *B*, *C*, *D* на рисунке).

После наделения исследуемых классов структурой идентифицируемые объекты образуют пересечения в масштабах классов K_P^T. Объем пересечений K_P^T представляет собой не что иное, как информационную неопределенность об объекте. Это дает основание задать на ней меру – меру информационной неопределенности (энтропии).

В данном случае энтропия интерпретируется как недостаток сведений о состоянии исследуемого объекта. Информация в той или иной мере устраняет эту неопределенность. Однако сведения, несущие информацию, могут содержать некую неопределенность, причем неопределенность двух типов: неточность сведения о состоянии объекта и определенную степень истинности данного сведения, то есть сведение может быть неабсолютно точным и неабсолютно истинным. Таким образом, в широком смысле неопределенность – векторный показатель, характеризующий неточность *H* и неистинность *P* сведений.

В части, касающейся рассмотрения аспекта неточности сведений, в нормативных документах допустимые значения неопределенностей



Модель каскадного фильтра классификации нечетких объектов с промежуточными состояниями и элементами неопределенности

Stage filter model for classification of fuzzy objects with intermediate states and elements of uncertainty

результатов переработки информации задаются предельными погрешностями в единицах измерения соответствующих физических величин. Следовательно, применение мер неопределенности необходимо рассмотреть с точки зрения представления результатов в соответствующих единицах измерения физических величин, а не в условных (бит, дит и т.д.). Для обоснования приемлемой меры неопределенности предлагается использовать теоретические положения математического аппарата теории ультраоператоров [4–6].

Пусть дано множество W, содержащее точку W_0, $W_0 \in \delta \subset W$. Если δ содержит те и только те точки W_i, которые обладают некоторым свойством, то возможно отождествление подмножества δ с данным свойством [4–6]. При этом истинное высказывание представляет собой точку W_0 из множества W, обладающую свойством δ и являющуюся элементарным сведением о точке W_0, представимой в виде одноместного предиката $\delta(W_0)$.

Таким образом, любое подмножество есть компакт $\delta(W_0)$, такой что $W_0 \in \delta(W_0)$, при этом $\delta(W_0) \subset W$ – элементарное сведение о точке W_0.

Пусть даны сведение $\delta(W_0) \subset W$ и семейство $\tilde{I}$ сведений о точке W_0, являющееся фильтром над подмножеством $\delta(W_0)$. Семейство $\tilde{I}$ подмножеств некоторого множества называется фильтром, если выполняются условия: $\phi \notin \tilde{I}$, $B \subset A$, $B \in \tilde{I} \Rightarrow A \in \tilde{I}$, $A, B \in \tilde{I} \Rightarrow A \cap B \in \tilde{I}$ [7].

Математическое определение информации введено в работах [6, 8, 9]. Семейство $\tilde{I}$ подмножеств множества W-сведений о точке W_0, являющееся фильтром над подмножеством $\delta(W_0)$, есть элементарная информация о точке W_0. Иными словами, элементарная информация – это семейство всех истинных следствий и умозаключений, полученных средствами математической логики из истинного высказывания – элементарного сведения $\delta(W_0)$.

Если множества W, δ измеримы (по Лебегу), то введенная на них мера может служить мерой неопределенности H сведения $\delta(W_0)$ о точке W_0 [8]. Неопределенность сведения $\delta(W_0)$ измеримого множества есть мера (Лебега) данного множества: $H(\delta(W_0)) = mes(\delta(W_0))$.

Мерой *mes* (необязательно Лебега) на полукольце множеств называется неотрицательная функция, принимающая конечные значения и являющаяся аддитивной [7, 8]:

$A \cap B = \phi \Rightarrow mes(A \cup B) = mes(A) + mes(B)$, $mes(\phi) = 0$.

Исходя из определения меры введенная мера неопределенности неотрицательна и конечна, то есть ограничена, и удовлетворяет общим требованиям [8] к мерам неопределенности (неотрицательности, равенства нулю при отсутствии неопределенности, аддитивности):

1. $H(\delta(W_0)) \geq 0$;

2. $\delta(W_0) = W_0 \Rightarrow H(\delta(W_0)) = 0$, $\delta(W_0) \neq W_0 \Rightarrow \Rightarrow H(\delta(W_0)) > 0$;

3. $\delta_1(W_0) \cap \delta_2(W_0) = W_0 \Rightarrow H(\delta_1(W_0) \cup \cup \delta_2(W_0)) = H(\delta_1(W_0)) + H(\delta_2(W_0))$.

Допустим, множество W есть множество действительных чисел R. Пусть некоторое действительное число $r \in R'$, $R' \subset R$, где R' – множество возможных значений r, являющееся отрезком на числовой оси. Тогда элементарным сведением о значении числа r является некоторый отрезок ΔR, $r \in \Delta R$, $\Delta R \subset R'$.

Неопределенность H_r сведения о значении r есть длина $\Delta R \subset R'$, $r \in \Delta R$, то есть $H_r = |\Delta R|$.

Очевидно, что $\max H_r = H_r(\Delta R \equiv R')$ единиц измерения, $\min H_r = H_r(\Delta R \equiv r) = 0$.

Значения неопределенностей H_{xj} обусловливают значение неопределенности H_y, однако при некоторых, в частности, нелинейных преобразованиях $y = F(X)$, неопределенность H_y является функцией не только значений неопределенностей H_x аргументов, но и расположения областей неопределенностей H_x на множестве возможных значений аргументов.

Поскольку априори неизвестно, каковы будут значения аргументов и, следовательно, где будут расположены области их неопределенностей, для однозначности определения меры неопределенности результата преобразования в различных задачах следует вводить дополнительные условия. Исходя из принципа гарантированного результата следует ввести требование экстремальности (max, min, sup, inf) значения неопределенности H_y на множестве возможных значений аргументов. В работе [8] предложены ряд определений и соответствующая мера неопределенности.

Определение 1. Для задач, пессимистическим вариантом в которых является наименьшая неопределенность, существенная неопределенность H_y сведения о числе y (результате преобразования F) – есть минимальная неопределенность H_y сведения о числе y по множеству значений аргументов:

$$\bar{H}_y(F, \{H_{X_j}\}) = \min_{x_j \in X_j} H_y(F, \{H_{X_j}\}, \{x_j\}),$$

$j = 1, \ldots, J$.

Определение 2. Для задач, пессимистическим вариантом в которых является наибольшая неопределенность, *существенная неопределенность* H_y сведения о числе y (результате преобразования F) – есть максимальная неопределенность H_y сведения о числе y по множеству значений аргументов:

$$\bar{H}_y(F,\{H_{X_j}\}) = \max_{x_j \in X_j} H_y(F,\{H_{X_j}\},\{x_j\}),$$

$j = 1, \ldots, J$.

Предлагаемый аппарат относительно близок к известному аппарату ультраоператоров теории ультрасистем [6], также оперирующих со сведениями в терминах определений неэлементарных сведений (рассмотрены ниже).

Отличие предложенного аппарата заключается в следующем:

– предлагается оперировать не сведениями, а их неопределенностями, не рассматриваемыми в аппарате ультраоператоров;

– в некоторых задачах рассматриваются элементарные сведения, что является частным случаем в аппарате ультраоператоров и облегчает вычисления;

– область применения сужается до чисел (то есть сведения-множества могут быть только числовой природы, компактами, в том числе многомерными);

– оперирование числовыми множествами-сведениями в некоторых случаях исключает необходимость применения в явном виде решетки (и соответствующих шкал) понятий и позволяют оперировать в неявном виде с бесконечными решетками.

Рассмотрим аспект истинности сведений.

Информация, сведения которой характеризовались двумя значениями истинности p – либо истина ($p = 1$), либо ложь ($p = 0$), уже приводилась ранее. В случаях, когда истинность сведения о точке может иметь не только эти два значения, но и некоторые промежуточные ($0 < p < 1$), возникает неэлементарная информация.

В [9] решеткой достоверностей называется произвольная решетка P, в которой максимальный элемент трактуется как истина, а минимальный как ложь. Сравнимые элементы p_1, p_2 решетки достоверностей записывают $p_1 < p_2$. Решетка достоверностей, состоящая только из двух элементов $\{p_{\min} = 0, p_{\max} = 1\}$, называется элементарной, остальные – неэлементарными. Если неэлементарная решетка достоверностей линейно упорядочена, она называется вероятностной, иначе – модальной.

Таким образом, семантика всякого сведения предполагает наличие четырех величин: опорного множества W состояний объекта, семантического указателя W_0 одного из состояний объекта $W_0 \in W$, подмножества δ состояний объектов из W, $\delta \subset W$, и семантической истинности p, которая характеризует истинность выполнения условия $W_0 \in \delta$. При этом неэлементарное сведение обозначается кортежем $<p, \delta(W_0)>$.

В исследованиях иногда полезно применение сведения, неопределенность которого равна максимально возможной, а вероятность, соответственно, единице.

В решаемой задаче введен ряд ограничений, в том числе ограничение на исследуемое исходное множество X_n, $n = \overline{1,m}$, являющееся конечным: $M = \text{card}(X_n) < \infty$. Поскольку имеющаяся неопределенность удовлетворяет всем свойствам, предъявляемым к информационной энтропии (доказательство не приводится из-за объема), и учитывается введенное ограничение, в качестве меры информационной неопределенности в классах с нечеткой структурой предлагается использовать меру неопределенности, предложенную в [8].

Результатом фильтрации на участках f_2, f_3, …, f_{g-1}, а также функционирования предложенной модели в целом является $H(K_p^T) \rightarrow 0$, то есть максимальное снижение информационной энтропии и выделение четких объектов (классов) за счет введения ограничений фильтрации на каждом этапе в зависимости от условий решаемой задачи.

На основании изложенного можно сделать следующий вывод. Предлагаемый подход и представленные математическая модель и мера информационной неопределенности являются составной частью разрабатываемого метода классификации и кластеризации состояний сложных систем на основе теоретико-множественного подхода и позволяют рассматривать процесс получения четких классов с точки зрения снижения информационной энтропии с использованием каскадного фильтра. Дальнейшее развитие метода предполагает определение ограничений фильтрации, а также синтез алгебраических конструкций для решения задачи кластеризации.

Публикация выполнена в рамках государственного задания ФГУ ФНЦ НИИСИ РАН (фундаментальные научные исследования 47 ГП) по теме № 0065-2019-0001 «Математическое обеспечение и инструментальные средства для моделирования, проектирования и разработки элементов сложных технических систем, программных комплексов и телекоммуникационных сетей в различных проблемно-ориентированных областях» (АААА-А19-119011790077-1).

Литература

1. Русаков К.Д., Селиверстов Д.Е., Смирнов А.Д. Разработка теоретических основ классификации и кластеризации нечетких признаков на основе теории категорий // XI Междунар. конф.: Управление развитием крупномасштабных систем. 2018. Т. 1. С. 320–322.
2. Курош А.Г., Лившиц А.Х., Шульгейфер Е.Г. Основы теории категорий // УМН. 1960. Т. 15. № 6. С. 3–52. DOI: 10.1070/RM1960v015n06ABEH001116.
3. Омельченко В.В. Общая теория классификации. Ч. 2: Теоретико-множественные основания. М.: Кн. мир, 2010. 295 с.
4. Стинрод Н., Эйленберг С. Основания алгебраической топологии; [пер. с англ.]. М.: Физматгиз, 1958. 403 с.
5. Eilenberg S., MacLane S. Relations between homology and homotopy groups of spaces. Ann. Math., 1945, vol. 46, pp. 480–509.
6. Хиль С.Ш., Решетников В.Н., Савилкин С.Б. Применение нечеткой меры достоверности для анализа технического состояния летательных аппаратов в условиях нестохастической неопределенности // Программные продукты, системы и алгоритмы. 2019. № 3. С. 8–13. DOI: 10.15827/2311-6749.19.3.2.
7. Генов А.А., Русаков К.Д., Хиль С.Ш. Идентификация состояния сложной технической системы в условиях неопределенности измерительной информации // Программные продукты и системы. 2017. Т. 30. № 3. С. 373–377. DOI: 10.15827/0236-235x.119.373-377.
8. Князев В.В. Особенности обеспечения достоверности перерабатываемой логической информации в АСУ специального назначения // Вопросы защиты информации. 2009. № 2. С. 22–29. URL: http://izdat.ntckompas.ru/editions/for_readers/archive/article_detail.php?SECTION_ID=155;&ELEMENT_ID=12634 (дата обращения: 20.09.2020).
9. Соболева Т.С., Чечкин А.В. Дискретная математика. М.: Академия, 2006. 256 с.

Software & Systems
DOI: 10.15827/0236-235X.132.599-604

Received 24.09.20
2020, vol. 33, no. 4, pp. 599–604

Development of theoretical bases for classification and clusterization of fuzzy features based on the theory of categories

K.D. Rusakov [1], *Research Associate, rusakov.msk@yandex.ru*
D.E. Seliverstov [2], *Ph.D. (Engineering), Leading Researcher, seliverstov_dmitriyy@rambler.ru*
S.Kh. Khil [3], *Ph.D. (Engineering), Associate Professor, skhill@mail.ru*
S.B. Savilkin [4], *Ph.D. (Physics and Mathematics), Associate Professor, Senior Researcher, Savilkin@mail.ru*

[1] *V.A. Trapeznikov Institute of Control Sciences of RAS, 117997, Russian Federation*
[2] *Plekhanov Russian University of Economics, Moscow, 117997, Russian Federation*
[3] *Moscow Aviation Institute (National Research University), Moscow, 125993, Russian Federation*
[4] *Center of Visualization and Satellite Information Technologies SRISA, Moscow, 117218, Russian Federation*

Abstract. The paper provides a rationale for choosing the measure of uncertainty of information. It describes a modern approach based on the application of fundamental algebraic constructions of category theory. A feature of the set of equivalence relations is the direct establishment of an equivalence relation between an object and a class.

The paper shows that at present, there are a number of actual applied problems in the classification field that require a different approach to establishing the equivalence relationship – the use of a cascade filter model with intermediate states. To justify the measure of uncertainty about an object, the authors proposed to use theoretical propositions based on the mathematical apparatus of the theory of ultra-operators. The proposed device also operates with information in terms of definitions of non-elementary information.

The characteristics of the proposed device include: the suggestion operate not with information, but with their uncertainties, not considered in the device of ultra-operators; some problems are considered basic information which is a special case in the device of ultra-operators and facilitates the calculations; the scope is narrowed to numbers (i.e., data – sets can only be of numeric nature, compacts, including multidimensional); operating with numeric sets-information in some cases eliminates the need to explicitly use the grid (and the corresponding scales) concepts, and allow to operate implicitly with infinite lattices.

The approach proposed by the authors and the presented mathematical model and measure of information uncertainty is an integral part of the developed "Method of classification and clustering of States of complex systems based on the set-theoretic approach" and allows us to consider the process of obtaining clear classes from the point of view of reducing information entropy using a cascade filter.

Keywords: classification, clustering, category, functor, information entropy.

Acknowledgements. *The reported study was funded by the framework of the state assignment of FSI FRC RISS RAS (implementation of fundamental research 47 GP) on topic no. 0065-2019-0001 "Software and tools for modeling, design, and development of elements of complex technical systems, software systems, and telecommunication networks problem-oriented areas" (AAAA-A19-119011790077-1).*

References

1. Rusakov K.D., Seliverstov D.E., Smirnov A.D. Development of the theoretical foundations for fuzzy feature classification and clustering based on the theory of categories. *Proc. Intern. Conf. MLSD*, 2018, vol. 1, pp. 320–322 (in Russ.).
2. Kurosh A.G., Livshits A.Kh., Shulgeifer E.G. Foundations of the theory of categories. *Russian Math. Surveys*, 1960, vol. 15, no. 6, pp. 3–52 (in Russ.). DOI: 10.1070/RM1960v015n06ABEH001116.
3. Omelchenko V.V. *General Classification Theory. P. 2: Set-Theoretical Foundations*. Moscow, 2010, 295 p. (in Russ.).
4. Steenrod N., Eilenberg S. *Foundations of Algebraic Topology*. Princeton, 1952, 323 p. (Rus. ed.: Moscow, 1958, 403 p.).
5. Eilenberg S., MacLane S. Relations between homology and homotopy groups of spaces. *Ann. Math*., 1945, vol. 46, pp. 480–509.
6. Hill S.Sh., Reshetnikov V.N., Savilkin S.B. Fuzzy measure of reliability to analyze aircraft technical condition in the context of non-stochastic uncertainty. *Software Journal: Theory and Applications*, 2019, no. 3, pp. 8–13 (in Russ.). DOI: 10.15827/2311-6749.19.3.2.
7. Genov A.A., Rusakov K.D., Hill S.Sh. Identification of a complex technical system functional state under conditions of measurement data ambiguity. *Software & Systems*, 2017, vol. 30, no. 3. pp. 373–377 (in Russ.). DOI: 10.15827/0236-235x.119.373-377.
8. Knyazev V.V. The features of ensuring the reliability of processed logical information in a special purpose ACS. *Information Security Issues*, 2009, no. 2, pp. 22–29. Available at: http://izdat.ntckompas.ru/editions/for_readers/archive/article_detail.php?SECTION_ID=155;&ELEMENT_ID=12634 (accessed September 20, 2020).
9. Soboleva T.S., Chechkin A.V. *Discrete Math*. Moscow, 2006, 256 p. (in Russ.).

Для цитирования

Русаков К.Д., Селиверстов Д.Е., Хиль С.Ш., Савилкин С.Б. Разработка теоретических основ классификации и кластеризации нечетких признаков на основе теории категорий // Программные продукты и системы. 2020. Т. 33. № 4. С. 599–604. DOI: 10.15827/0236-235X.132.599-604.

For citation

Rusakov K.D., Seliverstov D.E., Khil S.Kh., Savilkin S.B. Development of theoretical bases for classification and clusterization of fuzzy features based on the theory of categories. *Software & Systems*, 2020, vol. 33, no. 4, pp. 599–604 (in Russ.). DOI: 10.15827/0236-235X.132.599-604.