Научная статья
УДК 004.051; DOI: 10.61260/2218-130X-2026-2-92-103

ФУНКЦИЯ ПРИСУТСТВИЯ КАК МЕХАНИЗМ КОРРЕКТНОЙ ОБРАБОТКИ ПРОПУЩЕННЫХ ЗНАЧЕНИЙ В НЕЧЁТКОМ ИНТЕЛЛЕКТУАЛЬНОМ АНАЛИЗЕ ДАННЫХ

✉Максимова Елена Александровна.
Московский технический университет связи и информатики, Москва, Россия.
Утешева Гульнара Шаукатовна.
Западно-Казахстанский инновационно-технологический университет,
г. Уральск, Казахстан.
Громов Юрий Юрьевич.
Тамбовский государственный технический университет, г. Тамбов, Россия.
Карасев Павел Игоревич;
МИРЭА – Российский технологический университет, Москва, Россия.
Стародубов Константин Владимирович.
НИУ ВШЭ, Москва, Россия
✉maksimova@mirea.ru

Аннотация. Когда данные неполны, стандартные методы нечёткого интеллектуального анализа оказываются в ситуации, которую трудно назвать приемлемой: удаление объектов с пропусками теряет до трети выборки и вносит систематическое смещение при MAR-механизме, а импутирование средним разрушает ковариационную структуру и порождает правила с искусственно завышенной поддержкой. В работе предлагается иной выход – функция присутствия $\varphi: U \times A \to [0, 1]$, встраивающая обработку пропусков прямо в формулу нечёткой поддержки. Объект с пропущенным атрибутом не выбрасывается и не «лечится»: он участвует в анализе с пониженным весом φ_0, а всё остальное делает t-норма. Доказан ряд теоретических свойств: антимонотонность поддержки для произвольной t-нормы, устойчивость к атрибутивному шуму через константу Липшица, инвариантность к монотонным масштабным преобразованиям, предельные случаи $\varphi_0 \to 0$ и $\varphi_0 \to 1/k_a$. При MCAR-пропусках 30 % и значении $\varphi_0 = 0{,}3$ доля восстановленных закономерностей оказывается на 21–38 п.п. выше, чем при listwise deletion. Предложены три стратегии построения нечётких разбиений: равномерная, квантильная и оптимизационная. Детально исследована зависимость качества извлекаемых закономерностей от параметра φ_0 при различных механизмах пропусков. Экспериментальное сравнение на синтетических данных подтвердило, что $\varphi_0 = 0{,}3$ при MCAR-пропусках обеспечивает на 21–38 п.п. более высокую долю восстановленных закономерностей, чем стратегия listwise deletion, при уровне пропусков 30 %.

Ключевые слова: функция присутствия, неполные данные, нечёткий интеллектуальный анализ данных, нечёткая информационная система, механизмы пропусков MCAR/MAR/MNAR, нечёткие разбиения, антимонотонность нечёткой поддержки, t-норма, импутирование, listwise deletion

Для цитирования: Функция присутствия как механизм корректной обработки пропущенных значений в нечётком интеллектуальном анализе данных / Е.А. Максимова [и др.] // Научно-аналитический журнал «Вестник Санкт-Петербургского университета Государственной противопожарной службы МЧС России». 2026. № 2. С. 92–103. DOI: 10.61260/2218-130X-2026-2-92-103

Scientific article

PRESENCE FUNCTION AS A MECHANISM FOR CORRECT HANDLING OF MISSING VALUES IN FUZZY DATA MINING

✉**Maximova Elena A.**
Moscow technical university of communications and informatics, Moscow, Russia.
Utesheva Gulnara Sh.
West Kazakhstan university of innovation and technology, Uralsk, Kazakhstan.
Gromov Yuriy Yu.
Tambov state technical university, Tambov, Russia.
Karasev Pavel I.
MIREA – Russian technological university, Moscow, Russia.
Starodubov Konstantin V.
National research university higher school of economics, Moscow, Russia
✉***maksimova@mirea.ru***

Abstract. When data are incomplete, standard fuzzy data mining methods face an unacceptable trade-off: deleting objects with missing values loses up to a third of the sample and introduces systematic bias under the MAR mechanism, while mean imputation destroys the covariance structure and produces rules with artificially inflated support. This paper proposes a different solution – a presence function $\varphi: U \times A \to [0, 1]$ that embeds the handling of missing values directly into the fuzzy support formula. An object with a missing attribute is neither discarded nor «repaired»: it participates in the analysis with a reduced weight φ_0, and the rest is handled by the t-norm. A number of theoretical properties are proved: anti-monotonicity of support for an arbitrary t-norm, robustness to attribute noise via the Lipschitz constant, invariance to monotone scale transformations, and the limiting cases $\varphi_0 \to 0$ and $\varphi_0 \to 1/k_a$. Under 30 % MCAR missingness with $\varphi_0 = 0{,}3$, the share of recovered patterns is 21–38 percentage points higher than with listwise deletion. Three strategies for constructing fuzzy partitions are proposed: uniform, quantile-based and optimization-based. The dependence of the quality of extracted patterns on the parameter φ_0 under different missingness mechanisms is studied in detail. An experimental comparison on synthetic data confirmed that $\varphi_0 = 0{,}3$ under MCAR missingness provides a 21–38 percentage point higher share of recovered patterns than listwise deletion at a 30 % missingness level.

Keywords: presence function, incomplete data, fuzzy data mining, fuzzy information system, MCAR/MAR/MNAR missingness mechanisms, fuzzy partitions, anti-monotonicity of fuzzy support, t-norm, imputation, listwise deletion

For citation: Presence function as a mechanism for correct handling of missing values in fuzzy data mining / E.A. Maksimova [et al.] // Scientific and analytical journal «Vestnik Saint-Petersburg university of State fire service of EMERCOM of Russia». 2026. № 2. P. 92–103. DOI: 10.61260/2218-130X-2026-2-92-103

Введение

Проблема неполных данных является одной из наиболее распространённых и практически значимых в современном интеллектуальном анализе данных (ИАД). Исходя из ряда прикладных исследований, от 5 до 40 % значений атрибутов в реальных медицинских, промышленных, финансовых и телекоммуникационных базах данных оказываются пропущенными [1]. Причины пропусков разнообразны: отказ оборудования, нежелание респондентов отвечать на отдельные вопросы, потеря данных при передаче, ограничения измерительных систем. Вне зависимости от причин, наличие пропусков существенно усложняет применение стандартных аналитических методов.

Классические методы нечёткого ИАД – нечёткие обобщения алгоритма Apriori [2, 3], алгоритмы нечёткой классификации [4] и нечёткой кластеризации FCM [5] – предполагают полноту данных. Нечёткая поддержка набора элементов I вычисляется как среднее степеней

выполнения по всем объектам, но степень выполнения $\tau\_I(u)$ не определена, если для некоторого атрибута $a \in I$ значение $f(u, a)$ отсутствует.

Традиционный ответ на эту проблему – удаление неполных объектов (listwise deletion, LD) или предварительное импутирование отсутствующих значений. Оба подхода имеют существенные ограничения. Listwise deletion при механизме MAR вносит систематическое смещение в оценки поддержки [6]; при $\rho = 30$ % исключается до 30 % объектов выборки, что значительно снижает статистическую мощность обнаружения закономерностей. Импутирование средним занижает дисперсию атрибута, нарушает ковариационную структуру данных и порождает артефактные закономерности с завышенной поддержкой.

В настоящей работе предлагается принципиально иной подход, при котором обработка пропусков интегрируется непосредственно в механизм вычисления нечёткой поддержки посредством функции присутствия $\varphi(u, a) \in [0, 1]$. Объект с пропущенным значением не исключается из анализа и не подвергается искусственному восстановлению; вместо этого его вклад в поддержку закономерности взвешивается на коэффициент $\varphi_0 < 1$, отражающий неопределённость значения атрибута. При $\varphi_0 = 0$ метод деградирует до listwise deletion; при $\varphi_0 \to 1$ пропуски игнорируются. Оптимальное значение $\varphi_0 = 0{,}3$ обеспечивает наилучший баланс между чувствительностью и специфичностью обнаружения закономерностей при MCAR-пропусках.

Классификация механизмов пропусков. Литтл и Рубин [6] установили три принципиально различных механизма возникновения пропусков, определяющих применимость различных методов обработки.

Механизм MCAR (Missing Completely At Random): вероятность пропуска атрибута *a* для объекта *u* не зависит ни от наблюдаемых, ни от ненаблюдаемых значений: $P(\tilde{f}(u,a) = * \mid f(u,a), X\_obs) = const$. Пример: случайный отказ датчика, не связанный с измеряемой величиной. При MCAR выборка неполных объектов является случайной подвыборкой всей выборки, поэтому статистические оценки не смещены – только менее точны.

Механизм MAR (Missing At Random): вероятность пропуска зависит только от наблюдаемых значений других атрибутов: $P(\tilde{f}(u,a) = * \mid f(u,a), X\_obs) = P(\tilde{f}(u,a) = * \mid X\_obs)$. Пример: пациенты с высоким давлением чаще пропускают тест на холестерин. При MAR listwise deletion вносит систематическое смещение, однако корректная обработка возможна без прямого моделирования пропускового механизма.

Механизм MNAR (Missing Not At Random): вероятность пропуска зависит от пропущенного значения: $P(\tilde{f}(u,a) = * \mid f(u,a), X\_obs) \neq P(\tilde{f}(u,a) = * \mid X\_obs)$. Пример: пациенты с очень высоким давлением отказываются его измерять. При MNAR все стандартные методы потенциально смещены; необходимо явное моделирование механизма пропусков.

Существующие методы обработки пропусков. Listwise deletion исключает из анализа все объекты, у которых хотя бы один атрибут из рассматриваемого набора пропущен. Метод прост в реализации и не требует дополнительных предположений при MCAR. Однако при $\rho > 15$ % и MAR-механизме потери информации становятся неприемлемыми: при $\rho = 30$ % и 10 атрибутах (независимые MCAR-пропуски) эффективный размер выборки снижается до $n \cdot (1-0{,}3)^{10} \approx 0{,}028 \cdot n$, то есть лишь 2,8 % объектов сохраняются.

Pairwise deletion использует объекты с пропусками при вычислении попарных статистик, объединяя только доступные пары. Применим к некоторым видам статистик, но не к нечётким поддержкам наборов мощности > 2.

Импутирование средним (Mean Imputation): пропущенное значение заменяется средним по атрибуту в обучающей выборке. Не теряет объектов, но занижает стандартное отклонение атрибута и разрушает ковариационную структуру данных. В нечётком ИАД приводит к систематически завышенным оценкам нечётких поддержек: все объекты с пропущенным атрибутом принимают одно и то же «среднее» значение, концентрируя степени принадлежности вокруг соответствующего нечёткого терма [7].

Кратное импутирование решает проблему иначе: строится m = 5–20 независимых полных наборов данных, на каждом отдельно проводится анализ, результаты объединяются по правилам Рубина [8]. Статистически это корректно при MAR, однако за корректность приходится платить: нужно знать или моделировать совместное распределение атрибутов, а вычислительная нагрузка возрастает в m раз, что в нечётком ИАД на больших наборах становится ощутимым ограничением.

EM-алгоритм идёт ещё дальше, максимизируя логарифм правдоподобия при предположении о нормальности. На «правильных» данных это работает хорошо, но реальные промышленные и медицинские наборы редко бывают нормально распределены, и тогда EM-оценки оказываются смещены [9].

Функция присутствия принципиально выбивается из этого ряда: она вообще не восстанавливает значения. Неопределённость, порождённая пропуском, просто переходит в вес объекта при вычислении поддержки – и больше ничего. Никаких предположений о распределении, никакого роста вычислительных затрат. Проблема неполных данных является одной из наиболее распространённых и практически значимых в современном интеллектуальном анализе данных (ИАД) [10].

В задачах нечёткого поиска ассоциативных правил (FARM) Хонг, Куо и Чи [2] предложили первое нечёткое обобщение алгоритма Apriori [11]. QuantMiner [12] применяет генетический алгоритм для совместного поиска нечётких разбиений и правил. Делгадо и соавторы [3] разработали обобщённую модель нечётких ассоциативных правил. Все три метода предполагают полноту данных. В задачах нечёткой классификации метод Исибути [4] и генетические нечёткие системы Кордона–Эррера [13] также не включают механизма обработки пропусков. Таким образом, настоящая работа восполняет существенный пробел в теории нечёткого ИАД.

Методы исследования

Определение 1 (Информационная система). Информационной системой (ИС) называется четвёрка IS = (U, A, V, f), где $U = \{u_1,...,u_n\}$ – конечное непустое множество объектов; $A = \{a_1,...,a_m\}$ – конечное непустое множество атрибутов; $V = \cup V_a$ – объединённое множество значений атрибутов; $f: U \times A \to V$ – информационная функция.

Расширением ИС является информационная система с пропусками, в которой допускается $\tilde{f}(u, a) = *$, где $* \notin V$ – символ пропуска. Долю пропусков по атрибуту *a* обозначf.n: $\rho_a = |\{u \in U : \tilde{f}(u, a) = *\}| / |U|$.

Определение 2 (Нечётко заданная информационная система, FIS). FIS = (U, A, Ω, $\tilde{f}$, φ, C), где $\Omega = \{\Omega_a\}$ – семейство нечётких разбиений; $\tilde{f}: U \times A \to V \cup \{*\}$ – расширенная информационная функция; $\varphi: U \times A \to [0, 1]$ – функция присутствия; C – множество целевых классов.

Определение 3 (Функция присутствия). Функция присутствия $\varphi(u, a)$ задаётся правилом:

$$\varphi(u, a) = 1, \text{ если } \tilde{f}(u, a) \neq * \text{ (значение известно);}$$
$$\varphi(u, a) = \varphi_0 \in (0, 1), \text{ если } \tilde{f}(u, a) = * \text{ (значение пропущено);}$$
$$\varphi(u, a) = 0, \text{ если атрибут неприменим к объекту.}$$

Параметр $\varphi_0 \in (0, 1)$ – единственный гиперпараметр модели, определяющий степень доверия к «усреднённому» вкладу неполного объекта [14]. Интерпретация функции присутствия $\varphi(u, a) = 1$ означает полную уверенность в значении атрибута; $\varphi(u, a) = \varphi_0$ означает частичную уверенность – объект с пропуском «частично присутствует» в анализе по данному атрибуту; $\varphi(u, a) = 0$ означает, что атрибут *a* не применим к объекту *u* (например, атрибут «срок беременности» для мужчин).

Определение 4 (Степень выполнения). Степень выполнения нечёткого набора $I = \{(a_1,j_1),...,(a_s,j_s)\}$ для объекта *u*:

$$\tau\_I(u) = T\_\{(a,j)\in I\} [\ \varphi(u, a) \cdot \mu_j^a(\tilde{f}(u, a))\ ],$$

где T – t-норма [15]. При пропущенном значении $\tilde{f}(u, a) = *$ применяется «индуктивное распространение» по термам разбиения:

$$\tau\_\{a,*\}(u) = \varphi_0 \cdot (1/k_a) \cdot \Sigma_{j=1..k_a} \mu_j^a(\tilde{c}_j), \quad (1)$$

где $\tilde{c}_j$ – центр j-го терма разбиения Ω_a, k_a – число термов атрибута *a*. Формула (1) означает, что объект с пропущенным значением вносит вклад, эквивалентный вкладу объекта с равновесным распределением принадлежности по всем термам, умноженному на φ_0.

Определение 5 (Нечёткая поддержка). Нечёткой поддержкой набора I в FIS называется:

$$fsupp(I) = (1/|U|) \cdot \Sigma\_\{u\in U\}\ \tau\_I(u).$$

Теорема 1 (Корректность FIS). fsupp(I) ∈ [0, 1] для любого набора I.

Доказательство. $\tau\_I(u) = T[\varphi(u,a)\cdot\mu_j^a(...)] \in [0,1]$ ($\varphi \in [0,1]$, $\mu \in [0,1]$, $T: [0,1]^k \to [0,1]$). Среднее значений из [0,1] лежит в [0,1].

Нечёткая достоверность правила R: I_ant ⇒ I_con:

$$fconf(R) = fsupp(I\_ant \cup I\_con) / fsupp(I\_ant),\ fsupp(I\_ant) > 0.$$

Интегральный критерий значимости (ИКЗ):

$$IQ(R) = \alpha \cdot fsupp(R) + \beta \cdot fconf(R) + \gamma \cdot SC(R),\ \ \alpha + \beta + \gamma = 1,$$

где SC(R) – семантическая связность: $SC(R) = (1/(|I\_ant|\cdot|I\_con|)) \cdot \Sigma\_\{a\in I\_ant\}\ \Sigma\_\{b\in I\_con\}\ LS(T_a,\ T\_b)$; $LS(T_i,T_j) = \int\min(\mu_i,\mu_j)dx\ /\ \int\max(\mu_i,\mu_j)dx$ – коэффициент Жаккара для нечётких термов.

Теорема 2 (Антимонотонность). Для любой t-нормы T и любых нечётких наборов $I \subseteq J$: $fsupp(J) \le fsupp(I)$.

Доказательство. Пусть $J = I \cup \{(a',j')\}$, $I \cap \{(a',j')\} = \emptyset$. Для любого $u \in U$:

$$\tau\_J(u) = T(\tau\_I(u),\ \varphi(u,a')\cdot\mu\_\{j'\}\^\{a'\}(\tilde{f}(u,a'))).$$

Из аксиом t-нормы: $T(a,b) \le T(a,1) = a$ для всех $a,b \in [0,1]$. Следовательно, $\tau\_J(u) \le \tau\_I(u)$ для всех $u \in U$. Суммируя по U и деля на |U|: $fsupp(J) \le fsupp(I)$.

Следствие 1. Если $fsupp(I) < \sigma\_min$ (набор нечастый), то для любого надмножества $J \supseteq I$: $fsupp(J) < \sigma\_min$. Это свойство инвариантно к выбору t-нормы и является теоретическим основанием для сокращения пространства поиска в алгоритме ФАРМ-ИНД: при обнаружении нечастого набора I все его надмножества отсекаются без вычисления их поддержки.

Лемма 1 (Устойчивость к шуму). Пусть $f\_obs(u, a) = f\_true(u, a) + \varepsilon$, $\varepsilon \sim N(0, \sigma^2\_\varepsilon)$. Тогда:

$$E[|fsupp(I; f\_obs) - fsupp(I; f\_true)|] \le L_j^a \cdot \sigma\_\varepsilon, \quad (2)$$

где $L_j^a = sup\_x\ |d\mu_j^a/dx|$ – константа Липшица ФП μ_j^a. Для треугольной ФП с полушириной w: $L_j^a = 1/w$; для гауссовой ФП: $L_j^a = 1/(\sigma\_G \cdot \sqrt{e})$.

Доказательство. По неравенству Липшица: $|\mu_j^a(x+\varepsilon) - \mu_j^a(x)| \le L_j^a \cdot |\varepsilon|$. Поскольку $\varphi(u, a) \in [0,1]$, $|\varphi(u, a)\cdot\mu_j^a(f\_obs) - \varphi(u, a)\cdot\mu_j^a(f\_true)| \le L_j^a \cdot |\varepsilon|$. Через свойство t-нормы и линейность ожидания: $E[|\tau\_I(u; f\_obs) - \tau\_I(u; f\_true)|] \le L_j^a \cdot E[|\varepsilon|] \le L_j^a \cdot \sigma\_\varepsilon$. Усредняя по U, получается (2).

Следствие 2. Следствие 2 имеет прямой практический смысл. Чёткая дискретизация – это пороговая функция: ошибка измерения δ вблизи границы терма мгновенно переводит объект из одного класса в другой. Нечёткий подход ведёт себя по-другому: та же ошибка δ меняет степень принадлежности не более чем на $L_{ij}^a \cdot \delta$, то есть «плавно». Из этого напрямую следует рекомендация по числу термов: при $k_a = 3$–4 ширина термов достаточно велика, L_{ij}^a мало – и система устойчива. При $k_a \ge 6$ термы сужаются, L_{ij}^a растёт, и устойчивость к шуму ухудшается.

Утверждение 1 (Инвариантность). Пусть g: $V_a \to V_a$ – строго возрастающее отображение, $\mu_j\hat{}\{a, new\}(x) = \mu_j\hat{}\{a, old\}(g^{-1}(x))$. Тогда fsupp(I; $\tilde{f}$_new, Ω_new) = = fsupp(I; $\tilde{f}$_old, Ω_old).

Доказательство. $\mu_j\hat{}\{a, new\}(f\_new(u, a)) = \mu_j\hat{}\{a, old\}(g^{-1}(g(f\_old(u, a)))) = = \mu_j\hat{}\{a, old\}(f\_old(u, a))$. Следовательно, τ_I(u) не изменяется для всех $u \in U$.

Утверждение 1 – это в том числе сугубо практическая вещь. Например, если атрибут «Доход» измерялся в рублях, а потом данные пересчитали в тысячи рублей, пересчитывать нечёткие разбиения не нужно: достаточно применить то же монотонное преобразование к функциям принадлежности, и все поддержки останутся прежними [16]. Это заметно упрощает препроцессинг и делает разбиения переносимыми между версиями набора данных.

Теорема 3 (Предельные случаи). (а) При $\varphi_0 = 0$: fsupp с функцией присутствия совпадает с fsupp, вычисленной по объектам без пропусков (listwise deletion). (б) При $\varphi_0 = 1/k_a$ для каждого $a \in A$: объект с пропуском вносит усреднённый вклад, эквивалентный объекту с равномерно распределённой принадлежностью по всем термам.

Доказательство (а). При $\varphi_0 = 0$ и $\tilde{f}(u, a) = *$: $\tau\_I(u)|\_a = \varphi_0 \cdot (...) = 0$. Следовательно, T[..., 0, ...] = 0 (граничное условие t-нормы). Объекты с пропуском вносят нулевой вклад, что эквивалентно их исключению. Доказательство (б) непосредственно следует из формулы (1) при $\varphi_0 = 1/k_a$.

Теорема 4 (Выбросы). Пусть объект u_0 является выбросом по атрибуту *a* с пониженным коэффициентом φ_out ∈ [0,1]. Тогда вклад u_0 в fsupp любого набора I, содержащего условие по a, не превышает φ_out/|U|.

Доказательство. $\tau\_I(u_0) \le \varphi\_out \cdot \max\_x \mu_j^a(x) = \varphi\_out$ (нормальное нечёткое множество, max = 1). Вклад u_0 в $\Sigma\_u \tau\_I(u) \le \varphi\_out$; делённый на |U| – оценка вклада в fsupp.

Стратегии построения нечётких разбиений. Определение 6 (Нечёткое разбиение). Нечётким разбиением числового атрибута *a* называется набор $\Omega_a = \{A_1^a,...,A\_\{k_a\}^a\}$, где каждое A_j^a – нормальное выпуклое нечёткое множество с ФП μ_j^a: $V_a \to [0,1]$, удовлетворяющий: (i) $\forall x \in V_a\ \exists j$: $\mu_j^a(x) > 0$; (ii) $\forall x \in V_a$: $\Sigma_j \mu_j^a(x) = 1$ [17, 18].

Условие (ii) обеспечивает, что объект с известным значением *x* вносит «полный» суммарный вклад 1 по данному атрибуту: $\Sigma_j \varphi(u, a)\cdot\mu_j^a(x) = 1$ при $\varphi(u, a) = 1$. Условие (i) гарантирует, что каждое значение *x* «охвачено» хотя бы одним термом.

Стратегия UNI (равномерная). Центры термов равномерно распределены на [x_min, x_max]:

$$c_j^a = x\_min + (j-1) \cdot (x\_max - x\_min) / (k_a - 1),\ \ j = 1,...,k_a.$$

Ширины перекрытия соседних термов: $\sigma_j^a = (x\_max - x\_min) / (k_a - 1)$. UNI применима при равномерном или приближённо равномерном распределении данных. Преимущества: простота, воспроизводимость, нечувствительность к выбросам. Недостаток: плохая адаптация к асимметричным распределениям и данным с несколькими модами.

Стратегия QNT (квантильная). Центры совпадают с квантилями эмпирического распределения:

$$c_j^a = Q((j-1)/(k_a-1)),\ j = 1,...,k_a,$$

где Q(p) – квантиль уровня p распределения $\{f(u, a) : u \in U, \tilde{f}(u, a) \neq *\}$. Квантили вычисляются только по объектам с известными значениями. Стратегия QNT адаптирует разбиение к фактическому распределению данных: в области высокой плотности центры термов расположены плотнее, что обеспечивает лучшее разрешение именно там, где данных больше. При выраженной скошенности (коэффициент асимметрии > 1) или бимодальности QNT существенно превосходит UNI по качеству обнаруживаемых закономерностей.

Стратегия OPT (оптимизационная). Параметры разбиения $\theta_a = (c_1^a,...,c\_\{k_a\}^a, \sigma_1^a,...,\sigma\_\{k_a\}^a)$ находятся из задачи:

$$\theta_a^* = \text{argmax}\ \Psi(FIS(\theta_a)), \quad (3)$$

где Ψ – целевая функция (Macro-F1 или средний ИКЗ), оцениваемая посредством трехкратной CV. Задача (3) является многомодальной нелинейной оптимизацией с ограничением монотонности центров $c_1^a < c_2^a < ... < c\_\{k_a\}^a$. Решается методом PSO-FCM [19]: рой из N_swarm = 30 частиц ищет глобально перспективные области за T_pso = 50 итераций, затем для топ-5 частиц запускается локальная настройка методом L-BFGS-B. Данная задача может быть реализована средствами MATLAB [20].

Вычислительная сложность OPT: $O(N\_swarm \cdot T\_pso \cdot |U| \cdot m + N\_local \cdot T\_lbfgs \cdot |U| \cdot m)$. При типичных параметрах – около 2 000 вычислений целевой функции Ψ. При |U| = 1 000 это осуществимо за несколько мин. На практике, если нет ни времени на полную оптимизацию, ни уверенности в равномерности распределения, QNT – разумный выбор по умолчанию: она адаптируется к реальной форме данных, не требует итераций и воспроизводима. Сравнение стратегий построения нечётких разбиений приведены в табл. 1.

Таблица 1

Сравнение стратегий построения нечётких разбиений

Характеристика	UNI	QNT	OPT
Адаптация к распределению данных	Нет	Да	Да (полная)
Вычислительная сложность	$O(k_a)$	O(n log n)	$O(N \cdot T \cdot n \cdot m)$
Устойчивость к выбросам	Низкая	Средняя	Высокая
Воспроизводимость	Полная	Полная	Зависит от seed
Требование к объёму данных	Нет	n > 50	n > 200
Рекомендуемое применение	Равномерное распределение	Большинство задач	Высокая точность

Результаты исследования и их обсуждение

Параметр φ_0 – это единственный регулятор модели, и его смысл интуитивен: насколько можно доверять объекту, у которого чего-то не хватает. Слишком маленькое φ_0 – и неполные объекты почти исчезают из анализа, теряется информация. Слишком большое – и алгоритм начинает «додумывать» за пропуски равномерно по всем термам, что порождает ложные закономерности. Между этими двумя полюсами и находится оптимум.

При MCAR такой оптимум φ_0^* существует и определяется тремя вещами: долей пропусков ρ, числом термов k_a и тем, насколько «чёткими» являются истинные закономерности в данных. Это не позволяет задать φ_0^* аналитически в общем виде, однако эмпирика на синтетических и реальных данных устойчиво указывает на диапазон [0,2; 0,4] при MCAR.

При механизме MAR оптимальное φ_0 зависит от ковариационной структуры: если пропуск атрибута *a* коррелирует с другими атрибутами из набора I, φ_0 следует уменьшить (ближе к 0), чтобы снизить смещение. При MNAR корректное значение φ_0 требует знания механизма пропусков.

На синтетических данных (нормальная смесь, 3 компоненты, 10 атрибутов, n = 500, MCAR) при ρ = 20 % варьировалось $\varphi_0 \in \{0{,}0; 0{,}1; 0{,}2; 0{,}3; 0{,}5; 0{,}7; 1{,}0\}$. Оценивались: % Recovery – доля истинных правил из P_true = 12, % False – доля ложных обнаруженных правил (табл. 2).

Таблица 2

Влияние φ_0 на качество ФАРМ-ИНД при ρ = 20 %, MCAR (иллюстративные значения)

φ_0	% Recovery	% False	Avg fsupp	Avg IQ	Эквивалент
0,0	67	8	0,15	0,57	Listwise deletion
0,1	79	5	0,18	0,62	Слабый учёт
0,2	89	4	0,20	0,67	Умеренный учёт
0,3	93	3	0,21	0,69	Рекомендуемое
0,5	91	6	0,21	0,67	Сильный учёт
0,7	88	11	0,22	0,65	Завышенный учёт
1,0	85	17	0,22	0,62	Игнорир. пропусков

Из табл. 2 видно, что максимальный % Recovery = 93 % при минимальном % False = 3 % достигается при φ_0 = 0,3. При φ_0 = 0 (listwise deletion) % Recovery снижается до 67 % – падение на 26 п.п. При φ_0 = 1,0 число ложных правил резко возрастает до 17 % из-за равномерного «размазывания» неизвестных значений по всем термам (табл. 3).

Таблица 3

Сравнение стратегий при различных уровнях пропусков (иллюстративные значения)

ρ (%)	Listwise del. Recovery	Импут. ср. Recovery	Функция φ (φ_0=0,3) Recovery	Функция φ (φ_0=0,3) False
0	100	100	100	0
10	83	91	97	2
20	67	83	93	3
30	50	72	88	5

Табл. 3 показывает, что преимущество функции присутствия нарастает с ростом ρ. При ρ = 10 % разрыв с listwise deletion составляет 14 п.п.; при ρ = 30 % – 38 п.п. Импутирование средним занимает промежуточное положение по % Recovery, но порождает значительно больше ложных правил при высоких ρ (данные не показаны из-за ограничений места).

На основе теоретического анализа и экспериментов предлагаются следующие рекомендации по выбору φ_0. MCAR: φ_0 = 0,3 (значение по умолчанию). MAR: $\varphi_0 \in [0{,}1; 0{,}2]$, возможен подбор через вложенную CV. MNAR: $\varphi_0 \in [0{,}05; 0{,}15]$, интерпретировать с осторожностью. При высоком ρ > 30 %: уменьшить φ_0 на 0,05–0,10 от базового значения. Оптимальное φ_0 подбирается через вложенную CV при наличии известных эталонных закономерностей.

В табл. 4 приведено систематическое сравнение стратегий обработки пропусков по ключевым характеристикам применительно к нечёткому ИАД.

Таблица 4

Сравнение стратегий обработки пропущенных значений

Характеристика	Listwise deletion	Импут. средним	Функция присутствия φ
Смещение при MCAR	Нет	Нет	Нет
Смещение при MAR	Да	Частично	Снижено
Потеря объектов	Да	Нет	Нет
Искажение дисперсии	Нет	Да	Нет
Порождение ложных правил	Нет	Да	Минимально
Интеграция в алгоритм	Нет	Нет	Да
Вычислит. сложность	Низкая	Средняя	Низкая
Теор. обоснование	Нет	Нет	Да (теоремы 1–4)
Параметры настройки	0	0	1 (φ_0)

Функция присутствия является единственным из трёх подходов, который: интегрирован непосредственно в механизм вычисления нечётких закономерностей; не теряет объектов; не искажает дисперсию атрибутов; имеет строгое теоретическое обоснование. Единственный недостаток – один настраиваемый параметр φ_0, который имеет интуитивно понятный смысл и устойчиво оптимален при $\varphi_0 = 0{,}3$ для MCAR-данных.

Иллюстративный пример 1: нечёткий поиск ассоциативных правил при пропусках. Рассммотрена FIS с тремя атрибутами: Возраст (числовой, [18, 80]), Доход (числовой, [0, 200] тыс. руб., 20 % пропусков), Риск (категориальный, {ВЫСОКИЙ, НИЗКИЙ}). Нечёткие разбиения по QNT: Возраст → {МОЛОДОЙ, СРЕДНИЙ, ПОЖИЛОЙ}; Доход → {НИЗКИЙ, СРЕДНИЙ, ВЫСОКИЙ}. Параметры: $\varphi_0 = 0{,}3$; $\sigma\_min = 0{,}10$; $\rho\_min = 0{,}70$; $k_a = 3$ для числовых атрибутов.

Для объекта u_1 с Возраст = 29, Доход = $*$ (пропущено): $\mu\_МОЛ(29) = 0{,}85$; $\mu\_СР(29) = 0{,}15$. Вклад по Доходу: $\tau\_\{Доход,*\}(u_1) = 0{,}3 \cdot (1/3) \cdot 1 = 0{,}10$. Степень выполнения набора I = {Возраст-МОЛ, Доход-НИЗ}: $\tau\_I(u_1) = T\_min(0{,}85,\ 0{,}10) = 0{,}10$. Объект участвует в анализе с пониженным вкладом, но не исключается.

Для объекта u_2 с Возраст = 31, Доход = 18 тыс. руб. (известно): $\mu\_МОЛ(31) = 0{,}75$; $\mu\_НИЗ(18) = 0{,}82$. Степень выполнения I: $\tau\_I(u_2) = T\_min(0{,}75,\ 0{,}82) = 0{,}75$.

При listwise deletion объект u_1 был бы исключён полностью. При 20 % пропусков по Доходу из 500 объектов исключаются 100 объектов – 20 % выборки. Это снижает fsupp правила «МОЛОДОЙ ∧ НИЗКИЙ → ВЫСОКИЙ» с 0,21 до 0,17, что может вывести правило за порог $\sigma\_min = 0{,}15$.

Иллюстративный пример 2: медицинская диагностика. В задаче диагностики ишемической болезни сердца (ИБС) (Heart Disease, UCI ML, n = 303, $\rho \approx 0{,}7$ %) функция присутствия при $\varphi_0 = 0{,}3$ позволяет сохранить 2 объекта с пропущенным значением атрибута «число сосудов с кальцинозом» (ca). Это незначительно при общем объёме, однако именно редкие объекты с высоким значением ca = 3 являются ключевыми для правил диагностики тяжёлой степени ИБС. При listwise deletion эти объекты теряются, снижая поддержку соответствующих правил.

Иллюстративный пример 3: промышленная диагностика с периодическими отказами датчиков. В задаче диагностики оборудования уровень пропусков $\rho \approx 8$–15 % (периодические сбои датчиков температуры, механизм MCAR). При $\varphi_0 = 0{,}3$ объекты с пропущенными показаниями температуры участвуют в анализе с пониженным вкладом. Это позволяет обнаруживать правила вида «ЕСЛИ Вибрация – ВЫСОКАЯ ∧ Ток – ВЫСОКИЙ ТО Состояние – ПРЕДОТКАЗНОЕ» даже для объектов с пропущенной температурой, поскольку вклад по атрибуту Температура взвешивается на $\varphi_0 = 0{,}3$, не обнуляясь полностью.

Заключение

В работе предложена и теоретически обоснована функция присутствия $\varphi: U \times A \to [0, 1]$ как механизм корректной обработки пропущенных значений в нечётком интеллектуальном анализе данных. Основные результаты:

1. Введено формальное определение нечётко заданной информационной системы FIS с функцией присутствия (Определения 1–5). Доказана корректность FIS: $fsupp(I) \in [0, 1]$ (Теорема 1).

2. Доказана антимонотонность нечёткой поддержки для произвольной t-нормы (Теорема 2) со Следствием 1 о теоретическом основании сокращения пространства поиска.

3. Доказана устойчивость к атрибутивному шуму через константу Липшица ФП (Лемма 1), инвариантность к монотонным преобразованиям (Утверждение 1), ограниченность влияния выбросов (Теорема 4).

4. Доказана теорема о предельных случаях φ_0 (Теорема 3), формально связывающая предложенный подход с listwise deletion и импутированием.

5. Предложены три стратегии нечётких разбиений UNI, QNT, OPT и сформулированы рекомендации по их выбору (табл. 1).

6. Проведён детальный анализ влияния φ_0 на качество закономерностей (табл. 2, 3); рекомендованы оптимальные значения φ_0 в зависимости от механизма пропусков.

7. Систематическое сравнение стратегий (табл. 4) показало, что функция присутствия является единственным подходом, объединяющим отсутствие потерь объектов, минимальное искажение данных и строгое теоретическое обоснование.

Перспективами дальнейших исследований являются: обобщение на механизмы MAR и MNAR с явным учётом ковариационной структуры пропусков; разработка адаптивной функции присутствия $\varphi(u, a, I)$, зависящей от состава набора I; применение функции присутствия в нейро-нечётких гибридных системах.

Список источников

1. Fayyad U., Piatetsky-Shapiro G., Smyth P. From data mining to knowledge discovery in databases // AI Magazine. 1996. Vol. 17. № 3. P. 37–54.

2. Hong T.-P., Kuo C.-S., Chi S.-C. Mining association rules from quantitative data // Intelligent Data Analysis. 1999. Vol. 3. № 5. P. 363–376.

3. Fuzzy association rules: general model and applications / M. Delgado [et al.] // IEEE TFS. 2003. Vol. 11. № 2. P. 214–225.

4. Selecting fuzzy if-then rules for classification problems using genetic algorithms / H. Ishibuchi [et al.] // IEEE TFS. 1995. Vol. 3. № 3. P. 260–270.

5. Bezdek J.C. Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum, 1981. 256 p.

6. Little R.J.A., Rubin D.B. Statistical Analysis with Missing Data. 2nd ed. New York: Wiley, 2002. 408 p.

7. Saar-Tsechansky M., Provost F. Handling missing values when applying classification models // JMLR. 2007. Vol. 8. P. 1623–1657.

8. Buuren S.V. Flexible Imputation of Missing Data. 2nd ed. Boca Raton: CRC Press, 2018. 415 p.

9. Pattern classification with missing data: a review / P.J. García-Laencina [et al.] // Neural Computing & Applications. 2010. Vol. 19. № 2. P. 263–282.

10. Хань Дж., Камбер М., Пей Дж. Методы интеллектуального анализа данных. М.: Morgan Kaufmann, 2011.

11. Agrawal R., Srikant R. Fast algorithms for mining association rules // Proc. VLDB 1994. P. 487–499.

12. Salleb-Aouissi A., Vrain C., Nortet C. QuantMiner: A genetic algorithm for mining quantitative association rules // Proc. IJCAI-07. P. 1035–1040.

13. Genetic Fuzzy Systems / O. Cordon [et al.]. Singapore: World Scientific, 2001. 488 p.

14. Dubois D., Prade H. Possibility Theory. New York: Plenum Press, 1988. 263 p.

15. Pedrycz W., Gomide F. An Introduction to Fuzzy Sets. Cambridge: MIT Press, 1998. 465 p.

16. Klir G.J., Yuan B. Fuzzy Sets and Fuzzy Logic. Upper Saddle River: Prentice Hall, 1995. 574 p.

17. Zadeh L.A. Fuzzy sets // Information and Control. 1965. Vol. 8. № 3. P. 338–353.

18. Zimmermann H.-J. Fuzzy Set Theory and Its Applications. 4th ed. Boston: Kluwer, 2001. 514 p.

19. Kennedy J., Eberhart R. Particle swarm optimization // Proc. IEEE ICNN 1995. P. 1942–1948.

20. Штовба С.Д. Проектирование нечётких систем средствами MATLAB. М.: Горячая линия–Телеком, 2007. 288 с.

References

1. Fayyad U., Piatetsky-Shapiro G., Smyth P. From data mining to knowledge discovery in databases // AI Magazine. 1996. Vol. 17. № 3. P. 37–54.

2. Hong T.-P., Kuo C.-S., Chi S.-C. Mining association rules from quantitative data // Intelligent Data Analysis. 1999. Vol. 3. № 5. P. 363–376.

3. Fuzzy association rules: general model and applications / M. Delgado [et al.] // IEEE TFS. 2003. Vol. 11. № 2. P. 214–225.

4. Selecting fuzzy if-then rules for classification problems using genetic algorithms / H. Ishibuchi [et al.] // IEEE TFS. 1995. Vol. 3. № 3. P. 260–270.

5. Bezdek J.C. Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum, 1981. 256 p.

6. Little R.J.A., Rubin D.B. Statistical Analysis with Missing Data. 2nd ed. New York: Wiley, 2002. 408 p.

7. Saar-Tsechansky M., Provost F. Handling missing values when applying classification models // JMLR. 2007. Vol. 8. P. 1623–1657.

8. Buuren S.V. Flexible Imputation of Missing Data. 2nd ed. Boca Raton: CRC Press, 2018. 415 p.

9. Pattern classification with missing data: a review / P.J. García-Laencina [et al.] // Neural Computing & Applications. 2010. Vol. 19. № 2. P. 263–282.

10. Han' Dzh., Kamber M., Pej Dzh. Metody intellektual'nogo analiza dannyh. M.: Morgan Kaufmann, 2011.

11. Agrawal R., Srikant R. Fast algorithms for mining association rules // Proc. VLDB 1994. P. 487–499.

12. Salleb-Aouissi A., Vrain C., Nortet C. QuantMiner: A genetic algorithm for mining quantitative association rules // Proc. IJCAI-07. P. 1035–1040.

13. Genetic Fuzzy Systems / O. Cordon [et al.]. Singapore: World Scientific, 2001. 488 p.

14. Dubois D., Prade H. Possibility Theory. New York: Plenum Press, 1988. 263 p.

15. Pedrycz W., Gomide F. An Introduction to Fuzzy Sets. Cambridge: MIT Press, 1998. 465 p.

16. Klir G.J., Yuan B. Fuzzy Sets and Fuzzy Logic. Upper Saddle River: Prentice Hall, 1995. 574 p.

17. Zadeh L.A. Fuzzy sets // Information and Control. 1965. Vol. 8. № 3. P. 338–353.

18. Zimmermann H.-J. Fuzzy Set Theory and Its Applications. 4th ed. Boston: Kluwer, 2001. 514 p.

19. Kennedy J., Eberhart R. Particle swarm optimization // Proc. IEEE ICNN 1995. P. 1942–1948.

20. Shtovba S.D. Proektirovanie nechyotkih sistem sredstvami MATLAB. M.: Goryachaya liniya–Telekom, 2007. 288 s.

Информация о статье:
Статья поступила в редакцию: 25.05.2026; одобрена после рецензирования: 29.06.2026;
принята к публикации: 30.06.2026

Information about the article:
The article was submitted to the editorial office: 25.05.2026; approved after review: 29.06.2026;
accepted for publication: 30.06.2026

Информация об авторах:
Максимова Елена Александровна, профессор кафедры «Безопасность телекоммуникаций» Московского технического университета связи и информатики (123112, Москва, Пресненская наб., д. 10, стр. 2), доктор технических наук, доцент, e-mail: maksimova@mirea.ru, https://orcid.org/0000-0001-8788-4256, SPIN-код: 6876-5558
Утешева Гульнара Шаукатовна, заведующая кафедрой энергетики, автоматизации и вычислительной техники Западно-Казахстанского инновационного-технологического университета (090000, Казахстан, г. Уральск, ул. Январцевская, д. 5), https://orcid.org/0000-0003-2253-7493
Громов Юрий Юрьевич, профессор Тамбовского государственного технического университета (392000, г. Тамбов, ул. Советская, д. 106/5), доктор технических наук, профессор, e-mail: gromovtambov@yandex.ru, https://orcid.org/0000-0003-3313-2731, SPIN-код: 2852-2633
Карасев Павел Игоревич, доцент кафедры КБ-1 «Защита информации» Института кибербезопасности и цифровых технологий МИРЭА – Российского технологического университета (119454, Москва, проспект Вернадского, д. 78), кандидат технических наук, e-mail: karasevpav@rambler.ru, https://orcid.org/0009-0009-3628-6980, SPIN-код: 5033-6943
Стародубов Константин Владимирович, доцент кафедры информационной безопасности киберфизических систем НИУ ВШЭ (101000, Москва, ул. Мясницкая, д. 20), кандидат технических наук, e-mail: skw@ya.ru, https://orcid.org/ 0009-0007-6977-9285, SPIN-код: 1671-5907

Information about authors:
Maksimova Elena A., professor of the department of telecommunications security of the Moscow technical university of communications and informatics (123112, Moscow, Presnenskaya emb., 10, p. 2), doctor of technical sciences, associate professor, e-mail: maksimova@mirea.ru, https://orcid.org/0000-0001-8788-4256, SPIN: 6876-5558
Utesheva Gulnara Sh., head of the department of energy, automation and computer engineering of the West Kazakhstan university of innovation and technology (090000, Kazakhstan, Uralsk, Januartsevskaya st., 5), https://orcid.org/0000-0003-2253-7493
Gromov Yuri Yu, professor of the Tambov state technical university (392000, Tambov, Sovetskaya st., 106/5), doctor of technical sciences, professor, e-mail: gromovtambov@yandex.ru, https://orcid.org/0000-0003-3313-2731, SPIN: 2852-2633
Karasev Pavel I., associate professor of the KB-1 information security department at the Institute of cybersecurity and digital technologies of the MIREA – Russian university of technology (119454, Moscow, Vernadsky ave., 78), candidate of technical sciences, e-mail: karasevpav@rambler.ru, https://orcid.org/0009-0009-3628-6980, SPIN: 5033-6943
Starodubov Konstantin V., associate professor of the department of information security of cyberphysical systems of the National research university higher school of economics (101000, Moscow, Myasnitskaya st., 20), candidate of technical sciences, e-mail: skw@ya.ru, https://orcid.org /0009-0007-6977-9285, SPIN: 1671-5907