

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ КЛАССИФИКАЦИИ ПРЕЦЕДЕНТОВ В АВТОМАТИЗИРОВАННЫХ СИСТЕМАХ ТЕХНИЧЕСКОЙ ДИАГНОСТИКИ

УДК 004.02:519.7

КОВАЛЕНКО Игорь Иванович

д.т.н., профессор кафедры программного обеспечения автоматизированных систем
Национального университета кораблестроения им. Макарова, г. Николаев

Научные интересы: методы анализа данных, прикладной системный анализ,
теория оптимальных решений, системы поддержки принятия решений.

e-mail: igor.kovalenko@nuos.edu.ua

МЕЛЬНИК Антон Витальевич

аспирант, старший лаборант кафедры программного обеспечения автоматизированных систем
Национального университета кораблестроения им. Макарова, г. Николаев

Научные интересы: управление проектами, системы поддержки принятия решений, прикладной системный анализ.

e-mail: antonmelniknuos@gmail.com

ВВЕДЕНИЕ

Для определения режима (класса) эксплуатации различных устройств, механизмов и конструкций проводится техническая диагностика таких объектов. Современные методы проведения технической диагностики (например, металлоконструкций порталных кранов (МПК)) позволяют получать данные, как в количественной форме (числовые), так и в качественной (в виде экспертных оценок (ЭО)). Как правило, знания о диагностируемых объектах (ДО) охватывают широкий круг областей знаний и, достаточно часто, носят описательный характер. В связи с этим только специалист, обладающий большим опытом работы в конкретной области, может обосновать принятие решения по конкретному ДО, как правило, находя такое решение «по аналогии». Такой подход, основанный на эффективном использовании экспертного опыта, был развит и формализован в рамках современного научного направления – метод рассуждения по прецедентам (Case-Based Reasoning (CBR)) [4 и др.]. Прецедент – это структурированное представление накопленного опыта в виде данных и знаний, обеспечивающее их последующую автоматизированную обработку при помощи специализированных программных средств. В

большинстве случаев достаточно простого параметрического представления прецедента:

$$CASE = (x_1, x_2, \dots, x_n, D),$$

где x_1, \dots, x_n – параметры, описывающие данный прецедент, D – рекомендации лицу, принимающему решение (ЛПР); n – количество параметров прецедента.

В рамках данного метода решаются такие задачи: поиск (классификация) прецедентов; формирование рекомендаций ЛПР, а в случае необходимости их адаптация; сохранение результатов принятого решения в базе прецедентов (БП). С учётом форм представленных данных, задачи классификации могут формулироваться по разному, и могут нести в себе ряд неопределённостей. Правильный выбор математического аппарата для решения задачи классификации зависит от анализа этих неопределённостей. Поэтому актуальным является сравнительный анализ математических методов классификации.

АНАЛИЗ ПУБЛИКАЦИЙ И ПОСЛЕДНИХ ДОСТИЖЕНИЙ

В настоящее время изучен и формализован ряд неопределённостей, которые моделируются различными математическими методами. Так, например, для форма-

лизации неопределённости широко используются вероятностные методы [2, 5 и др.], для нечёткости, в основном используется теория нечётких множеств (ТНМ) [1, 2, 3, 7, 8 и др.]. Кроме этого, в последние годы активно развивается и получает широкое распространение в различных научных областях теория грубых множеств (ТГМ) [1, 2, 6, 7 и др.], применение которой позволяет формализовать неточность при решении задачи классификации прецедентов и последующего их поиска.

Перечисленные методы формируют различные классификаторы, однако, рекомендации по их применению с учётом указанных видов неопределённостей (НЕ-факторов) [1, 2], разработаны явно недостаточно.

ПОСТАНОВКА ПРОБЛЕМЫ

Целью работы является сравнительный анализ методов классификации прецедентов с учётом некоторых наиболее изученных видов неопределённостей.

ИЗЛОЖЕНИЕ ОСНОВНОГО МАТЕРИАЛА

Проведём анализ современных методов классификации в соответствии со структурной схемой, представленной на рисунке 1.

Значения, которые могут принимать параметры прецедента, обычно представляются тремя типами: количественные, качественные и шкалированные [2]. В случае представления параметров прецедентов количественными значениями, которые являются результатами

измерения физических величин (например, вес, длина, температура и др.), используются вероятностно-статистические методы классификации. В настоящее время для классификации прецедентов широкое распространение получил метод «ближайшего соседа» и его модификации. В основе данного метода лежит мера сходства, в качестве которой используется, например, взвешенная евклидова метрика:

$$\text{sim}(X_k, X_j) = \sqrt{\sum_{i=1}^N \omega_i (X_i^k - X_i^j)^2},$$

где X_i^k и X_i^j – значение i -го признака для k -го и j -го прецедента соответственно; N – общее количество параметров прецедентов; ω_i – вес i -го признака. Ближайшим является прецедент, метрика которого окажется минимальной. Однако, следует отметить, что эффективность метода зависит от выбора метрики (меры сходства), которую весьма затруднительно определить для качественных и шкалированных типов значений, что является одним из недостатков данного метода. Кроме этого, этот метод не даёт четких рекомендаций ЛПР в случае, наличия нескольких равноудалённых прецедентов, а значения весовых коэффициентов параметров прецедентов назначаются на усмотрение эксперта, что может привести к получению ошибочных результатов.

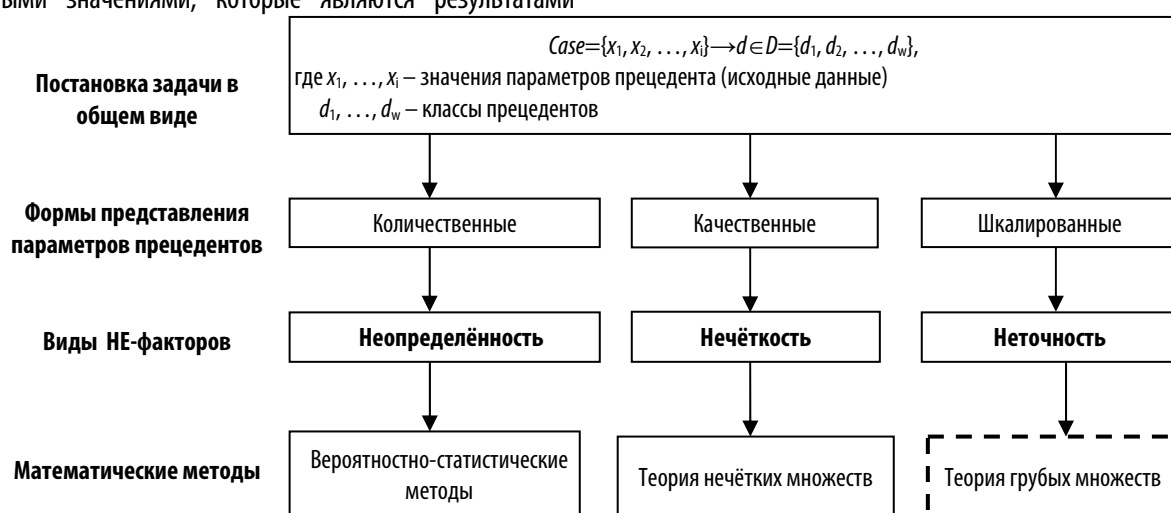


Рис. 1. Структурная схема анализа методов классификации прецедентов.

Указанные недостатки метода «ближайшего соседа» и его модификации вносят неопределённости при обработ-

ке количественных данных, которые могут быть промоделированы вероятностными методами классификации,

например, байесовским классификатором. Такая неопределённость, связанная с тем, с какими шансами может произойти каждое случайное событие из полной группы таких событий. Для этого необходимо выполнение двух условий: рассматриваются все возможные в данной ситуации события (вероятность принадлежности прецедентов классам); реализоваться может только одно из событий (классифицируемый прецедент принадлежит только одному классу). Общая идея байесовского классификатора заключается в определении вероятности принадлежности прецедентов к заданным классам и условные вероятности значений признаков параметров прецедентов, связанных с каждым из классов. Для реализации данного метода вводятся следующие понятия и обозначения: $p(d_w)$ – априорная вероятность класса прецедентов; $p(CASE^*/d_w)$ – условная вероятность принадлежности прецедента с параметрами $CASE^* = (x_1, x_2, \dots, x_n)$, классу d_w ; $p(d_w/CASE^*)$ – апостериорная вероятность принадлежности прецедента $CASE^*$ классу d_w , которая вычисляется выражением:

$$p(d_w / CASE^*) = \frac{p(d_w) \cdot p(CASE^* / d_w)}{\sum_{i=1}^m p(d_i) \cdot p(CASE^* / d_i)}.$$

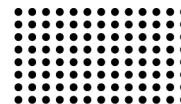
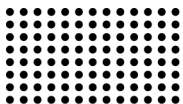
При классификации новых прецедентов рассчитываются условные вероятности его принадлежности к каждому из классов. Прецедент будет отнесен к тому классу, для которого рассчитанная условная вероятность будет максимальна по всему множеству классов. Как и другие методы, байесовский классификатор не лишён недостатков, к которым относится: статистическая информация, необходимая для применения данного метода, как правило, отсутствует, а её сбор и обработка связана с организационными и вычислительными сложностями. Кроме этого возникают трудности при пополнении БП новой информацией, что обуславливается необходимостью перерасчёта всех вероятностей. Также, использование данного метода базируется на допущении о том, что для каждого класса прецедентов свойственны определённые значения параметров прецедентов, которые не пересекаются. Однако, на практике достаточно часто возникают ситуации, при которых одинаковые значения свойственны нескольким различным классам.

В ряде случаев, результаты измерения экспертами некоторых параметров, представлены в вербальной

шкале, значения которых относятся к качественному типу данных (например, трещины в МПК: «длинная трещина», «нормальная трещина», «короткая трещина» и др.). При обработке таких типов данных возникает нечёткость, которая связана со степенью принадлежности некоторого элемента к некоторым классам (множествам), поскольку эти классы (множества) являются нечеткими, плохо определёнными понятиями. Для формализации нечёткости широко применяются методы (например: алгоритмы Гитмана-Левина; Распини; Думитреску и др. [3]), в основе которых лежит теория нечётких множеств. Нечётким множеством \tilde{A} множества X называется множество упорядоченных пар, составленных из элемента и его значения принадлежности: $\tilde{A} = \{(x / \mu_{\tilde{A}}(x))\}$, где $x \in X$, $\mu_{\tilde{A}}(x) \in [0, 1]$.

Функция $\mu_{\tilde{A}} : X \rightarrow [0, 1]$ называется функцией принадлежности, характеризующейся субъективной мерой нечёткости, и определяемая в результате опроса экспертов. Стандартная нечёткая логика опирается на такие допущения о природе принадлежности, как полнота (любой элемент либо принадлежит, либо не принадлежит множеству) и различимость элементов (любые два элемента являются, различимы на шкале принадлежности). Основываясь на значении функции принадлежности, такие качественные признаки можно упорядочить друг относительно друга, значения которых образуют ранговую или порядковую шкалу. Например: параметр прецедента x_1 = «длинная трещина» со значением $\mu_1 = 0,7$, принадлежит 1-му классу эксплуатации, в то же время этот параметр со значением $\mu_2 = 0,3$ принадлежит 2-му классу, то такой параметр следует отнести к 1-му классу, т. к. $\mu_1 > \mu_2$ и соответственно 1-й класс \succ 2-го класса. Данный пример подтверждается тот факт, что при классификации прецедентов с использованием ТНМ один и тот же параметр может принадлежать нескольким классам, но с разными значениями функций принадлежности.

При классификации новых прецедентов, рассчитываются функции принадлежности параметров прецедентов к каждому из классов, например, используя алгоритм [5]. Прецедент будет отнесен к тому классу, для которого рассчитанная функция принадлежности будет максимальна по всему множеству классов. Кроме этого, как и всем вышеописанным методам классификации прецедентов, потребуются дополнительный алгоритм опреде-



ления весовых коэффициентов, значения которых, могут задаваться на усмотрение экспертов.

Вместе с тем, задача эксперта при реализации технической диагностики больше заключается не в измерении параметров ДО в той или иной форме, а в оценивании их состояния, например: «хорошо», «нормально», «плохо» и др. Такая ситуация приводит к получению множества неупорядоченных данных в виде ЭО, и исключает возможность применения описанных методов. Это объясняется тем, что отсутствует информация, которая позволит связать между собой значения таких данных, которые относятся к шкалированному типу. К такой информации относятся: значения функций принадлежности, как в случае с ТНМ; значения о вероятностях, как в случае байесовского классификатора; количественные значения параметров для которых можно задать метрику, используемую в методе «ближайшего соседа». Например: параметр прецедента x_1 = «нормально» принадлежит 2-му и 3-классу прецедентов, однако, невозможно точно утверждать к какому классу следует отнести новый прецедент со значением такого параметра, тем самым моделируется неточность принадлежности некоторых элементов множеству (классу), для формализации которой используется теория грубых множеств [1, 2, 6, 7 и др.], в основе которой лежит понятие неразличимости (эквивалентности).

Концептуальные основы теории грубых множеств [6, 7] заключаются в том, что неточные знания (понятия), могут быть определены в рамках заданного обучающего множества с использованием понятий верхнего и нижнего приближений. Нижнее приближение включает те элементы обучающей выборки, которые наверняка принадлежат понятию, верхнее приближение включает все элементы, которые возможно принадлежат понятию.

Разница между двумя этими приближениями образует граничную область и содержит элементы, которые не могут быть классифицированы наверняка на основе имеющейся информации.

Под базой знаний понимается БП, которая определяется в данной теории как $K=(U, R)$, где U – универсум элементов (прецеденты, представленные ЭО) и R – отношение эквивалентности (значения параметров прецедентов), на основе которого могут быть выделены классы эквивалентности (категории) элементов U (обозначаются $IND(R)$). В каждую категорию включаются элементы, которые обладают одинаковыми значениями классификационных признаков

(атрибутов). Внутри каждой категории элементы считаются неразличимыми или эквивалентными.

Пусть элементы универсума классифицированы по категориям на основе отношения эквивалентности R . Если мы возьмем целевое множество элементов $X \subseteq U$, то относительно классификации $IND(R)$ могут быть рассмотрены следующие ситуации:

1. Множество X является объединением некоторых категорий из $IND(R)$. В этом случае множество X называется R -точным.

2. Множество X не может быть выражено как объединение некоторых категорий $IND(R)$. В этом случае X называется R -точным. В этом случае множество X называется R -неточным или R -грубым.

R -нижней аппроксимацией грубого множества X называется подмножество таких его элементов, которые могут быть классифицированы как принадлежащие X на основе классификации $IND(R)$:

$$\underline{R}X = \cup \{Y \in IND(R) : Y \subseteq X\}.$$

R -верхней аппроксимацией грубого множества X называется подмножество таких его элементов, которые возможно принадлежат X :

$$\overline{R}X = \cup \{Y \in IND(R) : Y \cap X \neq \emptyset\}.$$

R -нижнюю аппроксимацию целевого множества X называют R -положительной областью X : $POS_R(X) = \underline{R}X$.

R -отрицательной областью X называется подмножество элементов универсума, которые с определенностью не принадлежат X :

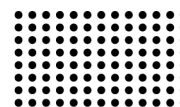
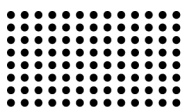
$$NEG_R(X) = U - \overline{R}X.$$

R -граничной областью целевого множества X называется подмножество его элементов, которые принадлежат R -верхней аппроксимации, но не принадлежат R -нижней аппроксимации:

$$BN_R(X) = \overline{R}X - \underline{R}X.$$

Рассмотрим простой пример, иллюстрирующий вышеприведенные понятия [6].

Имеется БП $K=(U, R)$, $U = \{x_i \mid i = \overline{1,10}\}$ – универсум элементов, R –отношение эквивалентности на основе которого выделены следующие классы эквивалентности (категории) на U :



$$U / IND(R) = \{\{x_1, x_2\}, \{x_3, x_7, x_{10}\}, \{x_4\}, \{x_5\}, \{x_6\}, \{x_8\}, \{x_9\}\}.$$

Заданы целевые подмножества (классы прецедентов) $X_1 = \{x_1, x_2, x_4, x_5\}$ и $X_2 = \{x_1, x_2, x_3, x_4\}$. Необходимо определить аппроксимации, отрицательные и граничные области для этих множеств.

Имеем:

$$\underline{R}X_1 = \{x_1, x_2, x_4, x_5\};$$

$$\overline{R}X_1 = \emptyset;$$

$$NEG_R(X_1) = \{x_3, x_6, x_7, x_8, x_9, x_{10}\};$$

$$BN_R(X_1) = \emptyset;$$

$$\underline{R}X_2 = \{x_1, x_2, x_4\};$$

$$\overline{R}X_2 = \{x_1, x_2, x_3, x_4, x_7, x_{10}\};$$

$$NEG_R(X_2) = \{x_5, x_6, x_8, x_9\};$$

$$BN_R(X_2) = \{x_3, x_7, x_{10}\}.$$

Философия грубых множеств такова, что выделение релевантных категорий элементов на универсуме, характеристика целевых множеств и операция над этими множествами производятся только и только на основе существующих знаний.

Метод, основанный на ТГМ, позволяет корректно обращаться с неточностью принадлежности прецедентов, моделируя их в виде граничной области, что предоставляет возможность дальнейшего анализа и точного отнесения таких прецедентов, значения параметров которых представлены ЭО, к заданным классам.

ВЫВОДЫ

В работе проанализированы неопределённости, возникающие в процессе решения задачи классификации прецедентов, параметры которых представлены различными типами. Установлена содержательная характери-

стика таких неопределённостей, а в результате анализа математических методов, несмотря на некоторые кажущиеся аналогии между ними, были определены принципиальные и существенные отличия к подходам моделирования таких неопределённостей. В основе каждого из таких методов лежит специфический математический аппарат и он предназначен для моделирования различных неопределённостей.

При обработке количественных данных используются вероятностно-статистические методы, а неопределённость моделируется в виде вероятности осуществления случайных событий. Для качественных типов данных применяются методы, основанные на ТНМ, которая моделирует нечёткость функцией принадлежности. Сходство между этими методами заключается лишь в способах получения исходной информации. Оценки вероятностей могут быть получены как объективным, так и субъективным (экспертным) путем. В ТНМ, для оценки степеней принадлежности элементов данному нечеткому множеству, также используются субъективные оценки.

При обработке шкалированных типов данных целесообразно использовать ТГМ, которая моделирует неточность в виде граничной области, при этом она не требует никакой предварительной или дополнительной информации о данных (информации о вероятностях или о степени принадлежности элемента множеству), как в вероятностных методах или ТНМ.

Проведенный анализ выдвигает условия детального анализа неопределённостей (НЕ-факторов), что обеспечивает правильный выбор методов моделирования, представляемых рассмотренными теориями.

ЛИТЕРАТУРА:

1. Batyirshin I. Z. Nechetkie gibridnye sistemy. Teoriya i praktika / Batyirshin I. Z., Nedosekin A. O., Stetsko A. A., Yarushkina N. G. Pod red. N. G. Yarushkinoy – M.: FIZMATLIT, 2007. – 208 s.
2. Vagin V. N. Dostoveriyiy i pravdopodobnyiy vyivod v intellektualnykh sistemah / Vagin V. N., Golovina E. Yu., Zagoryanskaya A. A., Fomina M. V., Pod red. V. N. Vagina, D. A. Pospelova. – M.: FIZMATLIT, 2004. – 704 s.
3. Vyatchenin D. A. Nechetkie metodi avtomaticheskoy klassifikatsii: Monografiya / D. A. Vyatchenin – Mn.: UP «Tehnoprint», 2004. – 219 s.
4. Klimchuk S. A. Primenenie pretsedentov dlya diagnostiki kranov mostovogo tipa / S. A. Klimchuk // Sistemi doslidzhennya ta Informatsiyni tehnologii. – 2012. – # 4. – S. 17-22. – Rezhim dostupa: <http://journal.iasa.kpi.ua/arhiv/2012/No4/2012-n4-klimchuk-text>.
5. Pankevich. O. D. Diagnostuvannya trischin budivelnih konstruksiy za dopomogoyu nechitkih baz znan'. Monografiya / Pankevich O. D., Shtovba S. D. – Vinnitsya: UNIVERSUM-Vinnitsya, 2005. – 108s.
6. Pawlak Z. Rough Sets Theoretical Aspects of Reasoning about Data [Text] // Boston; London: Academic Publishers, 1991. – 229 p.
7. Uzga-Rebrovs O. Nenoteiktibu parvaldisana [Text] / O. Uzga-Rebrovs. – Rezekne: RA Izdevnieciba. 2010. – vol. 3. – 560 lpp.
8. Zadeh L. A. Fuzzy sets [Text] // Information and Control. 1965. V. 8. P. 338-353.