КОМПЬЮТЕРНОЕ ОБЕСПЕЧЕНИЕ И ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА

УДК 004.896

Е. С. Макарова

ИССЛЕДОВАНИЕ ВЛИЯНИЯ ПАРАМЕТРОВ НЕЧЕТКОЙ МОДЕЛИ НА ТОЧНОСТЬ КЛАССИФИКАЦИИ ПРЕЦЕДЕНТОВ

Метод рассуждений на основе прецедентов (Case-Based Reasoning, CBR) используется для представления знаний в социально-экономических системах. Представлена история развития метода рассуждений на основе прецедентов и применение этого метода в различных областях. Рассматривается гибридная модель представления знаний на основе интеграции метода рассуждений на основе прецедентов и нечеткой логики. Рассмотрен алгоритм формирования нечетких правил, где каждая входная лингвистическая переменная может принимать 3, 5 или 7 терм-значений, описываемых треугольными функциями принадлежности. Предлагается новая процедура аккумуляции заключений конкурирующих правил, полученных в результате логического вывода. Исследуется точность классификации полученной гибридной модели на разных наборах данных и с различным набором функций принадлежности. Результаты исследований позволяют утверждать, что разработанный метод машинного обучения на основе нечеткого вывода существенно повышает точность классификации прецедентов.

Ключевые слова: прецедентный подход, прецедент, нечеткая логика, нечеткое множество, нечеткие правила, логический вывод, база знаний, процедура аккумуляции.

Введение

Возникновение в 1970-е гг. экспертных систем и систем, основанных на знаниях (СОЗ), связывают с развитием научной области искусственного интеллекта. К 1992 г. в промышленную эксплуатацию было внедрено около двух тысяч СОЗ [1]. Главными проблемами при разработке СОЗ являются извлечение высококачественных знаний и их представление в модели, адекватной решаемой задаче.

В 1980-е гг. формируется одна из парадигм, которая привлекает все больше и больше исследователей. Метод рассуждений на основе прецедентов (Case-Based Reasoning, CBR) – метод, в котором новые задачи решаются путем адаптации ранее решенных аналогичных проблем. Работа Р. Шенка [2], вышедшая в 1977 г., по распространенному мнению, является первой, где была описана концепция метода рассуждений на основе прецедентов. В этой работе было предложено обобщать знания о ситуациях и записывать их в виде сценариев, на основании которых в дальнейшем можно было бы формировать выводы. Впоследствии Шенк исследовал роль воспоминаний о предыдущих ситуациях (прецедентах) и структуру процесса организации и сохранения предыдущего опыта в виде контейнера знаний, который играет важную роль как в процессе решения задач, так и в процессе обучения [3]. Подчеркнем следующее: уже в [4] можно найти упоминания о том, что естественные понятия предметной области (концепты) чаще всего не могут быть классифицированы по простому набору свойств (признаков), но могут быть описаны с помощью более сложной структуры (прецедентов).

В начале 80-х гг. были разработаны первые приложения, основанные на методе рассуждений на основе прецедентов. Первая система, CYRUS [5], основанная на этом методе, была реализацией модели динамической памяти Шенка. Эта модель в дальнейшем послужила основой для создания других систем, например, MEDIATOR [6] и CHEF [7].

Исследования метода рассуждений на основе прецедентов проводились как в США, так и в Европе [8–10]. В Великобритании метод рассуждений на основе прецедентов исследовался применительно к гражданскому строительству (для диагностики неисправностей, ремонта и ре-

конструкции зданий) [11]; исследования по разработке систем, использующих метод рассуждений на основе прецедентов для интерпретации строительных норм и правил, проводившиеся в Эдинбурге, описаны в [12], исследователи из Уэльса применяли метод рассуждений на основе прецедентов при проектировании мостов автострад [13]. CBR-группы существуют также в Израиле [14] и Японии [15].

Исследования метода рассуждений на основе прецедентов проводятся и в России. В [16] исследуется интеграция Data Mining и метода прецедентов. В [17] предлагаются способы формирования базы знаний на основе правил и прецедентов с использованием онтологического анализа предметной области. В [18, 19] рассмотрены вопросы, связанные с моделированием временных зависимостей в интеллектуальных системах поддержки принятия решений (ИСППР) на основе прецедентов. В [20, 21] исследовались преимущества применения систем на основе темпоральных прецедентов для поддержки принятия решений в динамических предметных областях в условиях наличия неопределенности в исходных данных и достаточно жестких временных ограничений.

В настоящее время существуют готовые программные продукты для облегчения реализации метода рассуждений на основе прецедентов, включая коммерческие пакеты KATE, Spotlight, ESTEEM, а также свободно распространяемые программные продукты, такие как CBR-Works и CASPIAN. Основные достоинства прецедентных систем – легкость их реализации, относительная простота, не требуется также знание явной модели предметной области. Благодаря этому CBR-системы являются хорошим средством для представления знаний, однако таким системам присущ и ряд недостатков: сложность учета динамических факторов; описание прецедентов обычно не учитывает более глубокие знания о предметной области; большая база данных приводит к снижению производительности системы; возникают также трудности с определением хорошего критерия для индексирования и сравнения прецедентов. Довольно часто словари поиска и алгоритмы определения подобия приходится отлаживать вручную. Преодолеть указанные недостатки можно, создавая систем, в которых механизм принятия решений на основе прецедентов объединяют с другими моделями принятия решений.

В [22] в качестве альтернативного понятия подобия прецедентов рассматривается понятие релевантности на основе нечеткой логики. Показана приемлемая точность классификации прецедентов даже для малой обучающей выборки. Таким образом, перспективное решение проблем метода рассуждений на основе прецедентов лежит в области создания гибридных моделей, которые компенсируют недостатки классического CBR-подхода.

Мы предлагаем усовершенствование гибридной модели, которое заключается в новой процедуре аккумуляции заключений конкурирующих правил, полученных в результате логического вывода. Результаты исследований показывают значительное увеличение точности классификации прецедентов с использованием нового метода аккумуляции заключений на различных наборах данных и с разным набором функций принадлежности.

Гибридная модель представления знаний на основе интеграции метода рассуждений на основе прецедентов и нечеткой логики

Метод рассуждений на основе прецедентов – это метод, в котором решение новых проблем происходит на основе уже известных решений аналогичных проблем. Прецедент представляет собой описание проблемы и её решения. Прецеденты регистрируются в базе прецедентов. Задача метода рассуждений на основе прецедентов состоит в решении текущей проблемы в соответствии со следующими четырьмя этапами (CBR-цикл) [5]:

1. Найти наиболее подходящий (базовый) прецедент для решения текущей проблемы на основе мер сходства между проблемами.
2. Адаптировать решение из базового прецедента.
3. Пересмотреть решение текущей проблемы (полученное решение), если это необходимо.
4. Сохранить текущий случай как новый прецедент в базе прецедентов.

Интеграцию метода рассуждений на основе прецедентов с нечеткой логикой рассмотрим в следующей постановке. Предположим, что имеется выборка прецедентов, характеризующаяся рядом признаков (переменных), описывающих ситуацию, которые могут быть как количественными, так и качественными. В модели нечеткого вывода на рис. 1 эти признаки описываются входными лингвистическими переменными A, B, C, …. Выборка содержит качественный признак, характеризующий решение прецедента. В нашей постановке данный признак будет относить прецедент к определенному классу решений (классообразующий признак). В модели

нечеткого вывода на рис. 1 классообразующий признак описывается выходной лингвистической переменной *y*. Таким образом, мы рассматриваем задачу классификации прецедентов – отнесения прецедентов к тому или иному классу решений.

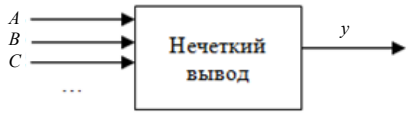

Рис. 1. Модель нечеткого вывода

На рис. 2 представлена гибридная модель, полученная в результате интеграции метода рассуждений на основе прецедентов и нечеткой логики.

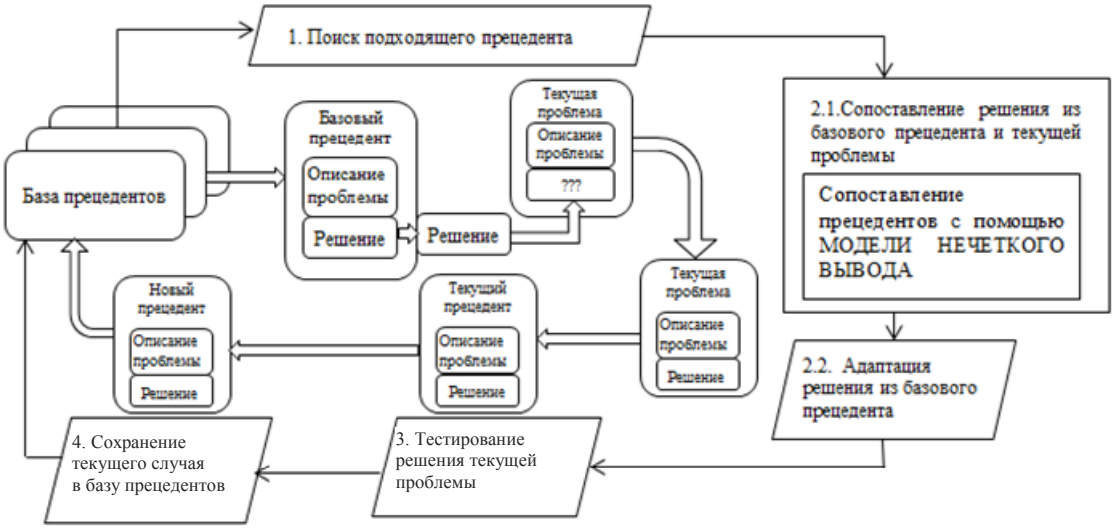

Рис. 2. Интеграция метода рассуждений на основе прецедентов и нечеткой логики

Исследования будем проводить на известных эталонных наборах данных: «Ирис», «Уровень знаний студентов» [23]. Набор данных «Ирис» содержит 150 примеров и характеризуется четырьмя атрибутами: длина и ширина (см) чашелистиков и длина и ширина лепестков. В данный набор входят три класса, в каждом из которых содержится по 50 прецедентов. Классообразующим атрибутом прецедента является класс набора данных «Ирис»: Iris setosa, Iris versicolor или Iris virginica. Набор данных «Уровень знаний студентов» содержит 119 примеров и характеризуется четырьмя атрибутами: SCG (показатель повторений материала студентами); STR (уровень затрат времени на обучение); LPR (уровень понимания связи предмета и цели на экзамене); UNS (уровень знаний студента). В данный набор входят три класса, в каждом из которых содержится разное количество прецедентов. Классообразующим атрибутом прецедента является класс, к которому относится уровень знаний студентов: Low, Middle, High.

Рассмотрим алгоритм формирования нечетких правил. Предположим, что каждая входная лингвистическая переменная принимает 3, 5 или 7 терм-значений, описываемые треугольными функциями принадлежности. Алгоритм формирования нечетких правил рассмотрим на наборе данных «Ирис» [23].

Алгоритм формирования нечетких правил из выборки прецедентов (алгоритм обучения)

Шаг 1. Загрузка исходных данных. На этом шаге загружаем наборы данных в программу.

Шаг 2. Формирование обучающей выборки. Загруженные данные группируем в классы, подсчитываем количество прецедентов в каждом классе и выводим это количество пользователю. Из исходной выборки отбираем прецеденты в обучающую выборку пропорционально распределению по классам. Отбор прецедентов в обучающую выборку происходит случайным образом. Прецеденты, которые не попали в обучающую выборку, формируют тестовую выборку.

Шаг 3. Построение функции принадлежности. Универсальное множество для каждой лингвистической переменной определяется динамически. В качестве универсального множества используем отрезок от минимального до максимального значения. Для задания терм-значений лингвистической переменной каждый полученный интервал (универсум) разделим на количество функций принадлежности равное $(2N + 1)$, $N = 1, 2, 3$, т. е. 3, 5 и 7.

Определим минимальное и максимальное значения для каждого свойства на обучающей выборке. Например, длина чашелистиков может варьироваться от 4,3 до 7,9 см. Рассмотрим первый элемент данных, который попал в обучающую выборку: {6,5; 2,8; 4,6; 1,5; Iris versicolor}. Определим минимальное и максимальное значения для каждого признака в обучающей выборке, т. е. для признака *A*: минимальное значение 4,3, максимальное – 7,9; для признака *B*: минимальное значение 2,0, максимальное – 4,4; для признака *C*: минимальное значение 1,0, максимальное – 6,9; для признака *D*: минимальное значение 0,1, максимальное – 2,5. На рис. 3 представлены функции принадлежности для двух признаков – *A* и *B*.

Шаг 4. Формирование нечетких правил с помощью обучающей выборки. Определим степень принадлежности обучающих данных для каждой функции принадлежности. Для примера рассмотрим признак *A*.

1. Для трех функций принадлежности.

Выберем первое значение из элемента данных {6,5; 2,8; 4,6;1,5; Iris versicolor}, т. е. $x_1 = 6{,}5$ для признака *A* и трех функций принадлежности (рис. 3, *а*). Значению $x_1 = 6{,}5$ соответствует значение A_1, равное 0, значение функции принадлежности A_2, равное 0,78, значение функции принадлежности A_3, равное 0,22. Максимальное значение равно 0,78 (max {0; 0,78; 0,22} = 0,78) и принадлежит функции принадлежности A_2, поэтому включается в предусловие нечеткого правила как $x_1[\max = 0{,}78 \text{ в } A_2]$, т. е. $x_1 = A_2$. Выберем второе значение из элемента данных {6,5; 2,8; 4,6; 1,5; Iris versicolor}, т. е. $x_2 = 2{,}8$ для признака *B* и трех функций принадлежности (рис. 3, *г*). Значению $x_2 = 2{,}8$ соответствует значение функции принадлежности B_1 равное 0,39, значение функции принадлежности B_2, равное 0,61, значение функции принадлежности B_3, равное 0. Максимальное значение равно 0,61 (max {0,39; 0,61; 0} = 0,61) соответствует функции принадлежности B_2, поэтому включается в предусловие нечеткого правила как $x_2[\max = 0{,}61 \text{ в } B_2]$, т. е. $x_2 = B_2$.

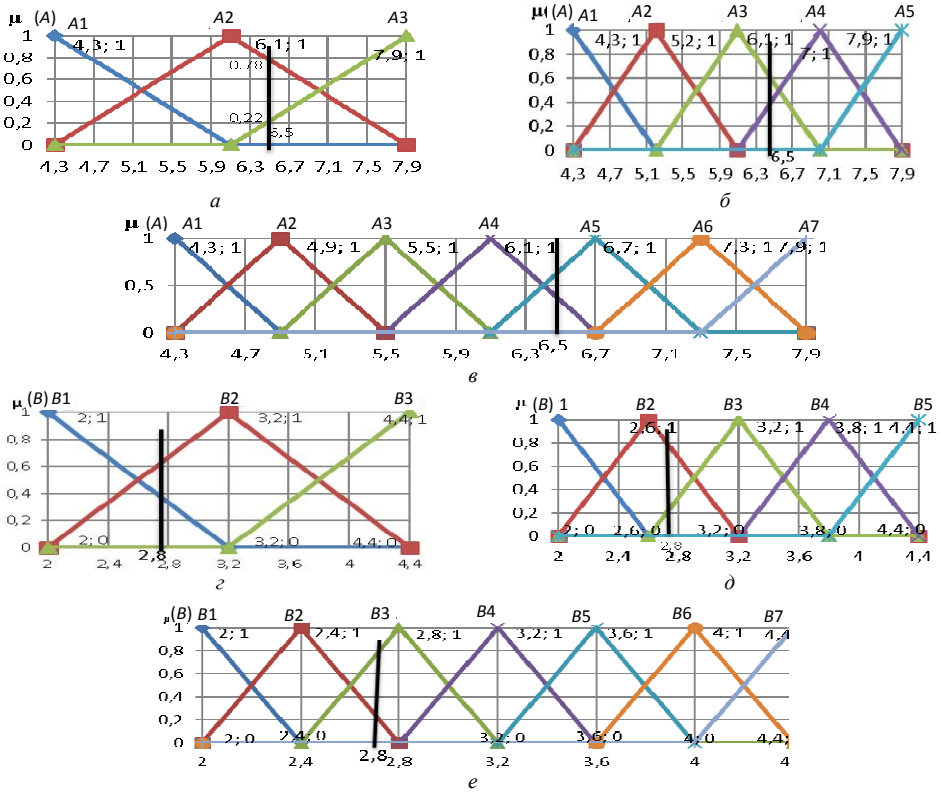

Рис. 3. Функции принадлежности (ФП):
для атрибута *A*: *а* – 3 ФП; *б* – 5 ФП; *в* – 7 ФП;
для атрибута *B*: *г* – 3 ФП; *д* – 5 ФП; *е* – 7 ФП

Формируя предусловия для x_3 и x_4 аналогичным образом, получаем следующие правила для 3 функций принадлежности:

$\{x_1, x_2, x_3, x_4; Z\} \Rightarrow$

$\{x_1[\max = 0,78 \text{ в } A_2], x_2[\max = 0,61 \text{ в } B_2], x_3[\max = 0,45 \text{ в } C_2], x_4[\max = 0,85 \text{ в } D_1];$

Iris setosa$\}$

Правило: $IF(x_1 = A_2, x_2 = B_2, x_3 = C_2, x_4 = D_1)$*Then* Iris setosa

1. Для пяти функций принадлежности.

Выберем первое значение из элемента данных {6,5; 2,8; 4,6; 1,5; Iris versicolor}, т. е. $x_1 = 6,5$ для признака *A* и пяти функций принадлежности (рис. 3, *б*). Значению $x_1 = 6,5$ соответствует значение функции принадлежности A_1, равное 0, значение функции принадлежности A_2, равное 0, значение функции принадлежности A_3, равное 0,58, значение функции принадлежности A_4, равное 0,4, значение функции принадлежности A_5, равное 0. Максимальное значение равно 0,58 (max {0; 0; 0,58; 0,4; 0} = 0,58) и принадлежит функции принадлежности A_3, поэтому включается в предусловие нечеткого правила как $x_1[\max = 0,58 \text{ в } A_3]$, т. е. $x_1 = A_3$. Выберем второе значение из элемента данных {6,5; 2,8; 4,6; 1,5; Iris versicolor}, т. е. $x_2 = 2,8$ для признака *B* и пяти функций принадлежности (рис. 3, *д*). Значению $x_2 = 2,8$ соответствует значение функции принадлежности B_1, равное 0, значение функции принадлежности B_2, равное 0,78, значение функции принадлежности B_3, равное 0,22, значение функции принадлежности B_4, равное 0, значение функции принадлежности B_5, равное 0. Максимальное значение равно 0,78 (max {0; 0,78; 0,22; 0; 0} = 0,78) и принадлежит функции принадлежности B_2, поэтому включается в предусловие нечеткого правила как $x_2[\max = 0,78 \text{ в } B_2]$, т. е. $x_2 = B_2$. Формируя предусловия для x_3 и x_4 аналогичным образом, получаем следующие правила для 5 функций принадлежности:

$\{x_1, x_2, x_3, x_4; Z\} \Rightarrow$

$\{x_1[\max = 0,58 \text{ в } A_3], x_2[\max = 0,78 \text{ в } B_2], x_3[\max = 0,75 \text{ в } C_5], x_4[\max = 0,9 \text{ в } D_3];$

Iris virginica$\}$

Правило: $IF(x_1 = A_3, x_2 = B_2, x_3 = C_5, x_4 = D_3)$*Then* Iris virginica

3. Для семи функций принадлежности.

Выберем первое значение из элемента данных {6,5; 2,8; 4,6; 1,5; Iris versicolor}, т. е. $x_1 = 6,5$ для признака *A* и семи функций принадлежности (рис. 3, *в*). Значению $x_1 = 6,5$ соответствует значение функции принадлежности A_1, равное 0, значение функции принадлежности A_2, равное 0, значение функции принадлежности A_3, равное 0, значение функции принадлежности A_4, равное 0,38, значение функции принадлежности A_5, равное 0,62, значение функции принадлежности A_6, равное 0, значение функции принадлежности A_7, равное 0. Максимальное значение равно 0,62 (max {0; 0; 0; 0,38; 0,62; 0; 0} = 0,62) и принадлежит функции принадлежности A_5, поэтому включается в предусловие нечеткого правила как $x_1[\max = 0,62 \text{ в } A_5]$, т. е. $x_1 = A_5$. Выберем второе значение из элемента данных {6,5; 2,8; 4,6; 1,5; Iris versicolor}, т. е. $x_2 = 2,8$ для признака *B* и семи функций принадлежности (рис. 3, *е*). Значению $x_2 = 2,8$ соответствует значение функции принадлежности B_1, равное 0, значение функции принадлежности B_2, равное 0,22, значение функции принадлежности B_3, равное 0,8, значение функции принадлежности B_4, равное 0, значение функции принадлежности B_5, равное 0, значение функции принадлежности B_6, равное 0, значение функции принадлежности B_7, равное 0. Максимальное значение равно 0,8 (max {0; 0,22; 0,8; 0; 0; 0; 0} = 0,8) и принадлежит функции принадлежности B_3, поэтому включается в предусловие нечеткого правила, как $x_2[\max = 0,8 \text{ в } B_3]$, т. е. $x_2 = B_3$. Формируя предусловия для x_3 и x_4 аналогичным образом, получаем следующие правила для 7 функций принадлежности:

$\{x_1, x_2, x_3, x_4; Z\} \Rightarrow$

$\{x_1[\max = 0,62 \text{ в } A_5], x_2[\max = 0,8 \text{ в } B_3], x_3[\max = 0,5 \text{ в } C_7], x_4[\max = 0,76 \text{ в } D_5];$

Iris versicolor$\}$

Правило: $IF(x_1 = A_5, x_2 = B_3, x_3 = C_7, x_4 = D_5)$*Then* Iris versicolor

Шаг 5. Сопоставление каждому правилу R степени истинности SP(R). Разрешение конфликтов. При переходе от прецедентов к набору нечетких правил может возникнуть конфликтная ситуация, когда множеству правил с одинаковыми предпосылками соответствуют различные заключения. В этом случае важно сопоставить степень истинности правила, т. е. чтобы у большего количества прецедентов класс был определен верно. Для этого исследуем 2 метода разрешения конфликтов.

Первый метод определения степени истинности правила. Для каждого правила введем степень истинности, которая рассчитывается как максимальное значение функций принадлежности для каждого атрибута. Для правила вида: Если (x_1 это A_1, x_2 это A_2), то y это B, степень истинности, обозначим как $SP(R) = \max(A_1, A_2)$. Для правила из примера на рис. 3, *а* для трех функций принадлежности $SP(R) = \max(0{,}78; 0{,}61; 0{,}45, 0{,}85) = 0{,}85$. Для формирования правил используется несколько прецедентов. В процессе формирования нечетких правил мы можем получить целый набор правил, антецеденты которых совпадают, а консеквенты разные. Такой набор правил называется конфликтным. Таким образом, в модели мы приняли допущение, что если антецеденты правил одинаковые, а консеквенты разные, то разрешить конфликт поможет степень истинности. Из множества конфликтующих правил с одинаковыми условиями выбирается правило с максимальной степенью истинности. Рассмотрим *пример*:

$$IF\,(x_1 = A_3\,,\, x_2 = B_2, x_3 = C_4\,,\, x_4 = D_3)\ Then\ \ Z = \text{Iris versicolo r},\ SP_1(R) = 0{,}64$$

$$IF\,(x_1 = A_3\,,\, x_2 = B_2, x_3 = C_4\,,\, x_4 = D_3)\ Then\ \ Z = \text{Iris virginica}\ ,\ SP_2(R) = 0{,}5$$

$$IF\,(x_1 = A_3\,,\, x_2 = B_2, x_3 = C_4\,,\, x_4 = D_3)\ Then\ \ Z = \text{Iris virginica}\ ,\ SP_3(R) = 0{,}55$$

Для этого правила, с учетом допущения,

$$IF\,(x_1 = A_3\,,\, x_2 = B_2, x_3 = C_4\,,\, x_4 = D_3)\ Then\ \ Z = \text{Iris versicolo r},$$

т. к. $SP(R) = \max(0{,}64, 0{,}5, 0{,}55) = 0{,}64$. Мы выбираем класс Iris versicolor, соответственно, 2 случая будут сопоставлены некорректно. В этом случае мы не учитываем количество правил с одинаковыми антецедентами.

Второй метод определения степени истинности правила. Находим общее количество правил с одинаковыми антецедентами, для нашего примера это будет 3. Далее рассчитываем новую степень истинности для выбора класса. В табл. 1 представлен механизм расчета степени истинности с учетом количества одинаковых правил для выбора класса.

Таблица 1

Выбор класса с учетом количества одинаковых правил

Класс	Значение $SP(R)$	Расчет $SP(R)$ для выбора класса
Iris versicolor	0,213	$SP(R) = 1/3 \cdot SP_1 = 1/3 \cdot 0{,}64 = 0{,}213$
Iris virginica	0,35	$SP(R) = 1/3 \cdot SP_2 + 1/3 \cdot SP_3 = 1/3 \cdot 0{,}5 + 1/3 \cdot 0{,}55 = 0{,}35$

В общем случае получаем следующую формулу для определения класса:

$$SP^*(R) = \max_{k \in I_{class}} SP_k\,(R),\ class = \arg\{\max_{j=1,n} \sum\nolimits_{i \in I_j} \frac{1}{m} SP_i(R)\},$$

где m – количество конфликтующих правил ($i = 1, \ldots, m$); $SP_i(R)$ – степень истинности i-го правила; n – количество классов, множество I разбивается на j-подмножеств, в соответствии с разбиением конкурирующих правил по классам. Определив по формуле класс, правилу ставим в соответствие степень истинности $SP^*(R)$. Для примера, представленного выше, значение $SP^* = 0{,}35$ (табл. 1), и это будет класс Iris virginica. В таблицу правил для правила $IF(x_1 = A_3, x_2 = B_2, \quad x_3 = C_4\ ,\ x_4 = D_3)$ установим класс Iris virginica со степенью истинности 0,55. Таким образом, двум наборам данных из трех, представленных в примере, будет правильно сопоставлен класс.

Шаг 6. Создание базы нечетких правил. База нечетких правил представляется списком правил:

$IF(x_1 = A_2,\ x_2 = B_2,\ x_3 = C_2,\ x_4 = D_1)\ Then$ Iris versicolor

$IF(x_1 = A_1,\ x_2 = B_3,\ x_3 = C_1,\ x_4 = D_3)\ Then$ Iris virginica

$IF(x_1 = A_2,\ x_2 = B_3,\ x_3 = C_2,\ x_4 = D_2)\ Then$ Iris versicolor и т. д.

Экспериментальные исследования гибридной модели

В этом разделе будут показаны результаты оценки на известных эталонных наборах данных: «Ирис», «Уровень знаний студентов» [23].

Обучение базы прецедентов. В экспериментах с набором данных «Ирис» случайным образом была сформирована обучающая выборка, а остальные данные попали в тестовую выборку. Нечеткие правила были применены к тестовой выборке. Число экспериментов было 10 для каждого вида испытаний. Для каждого атрибута было определено минимальное и максимальное значение на обучающей выборке. Далее каждый полученный интервал был разделен на 3, 5 или 7 функций принадлежности, т. е. количество лингвистических значений, например, для параметра A (длина чашелистика), в правилах изменялось от трех (A_1, A_2, A_3) до семи (A_1, A_2, A_3, A_4, A_5, A_6, A_7). Для обучения было выбрано 120 прецедентов. В табл. 2 показана точность классификации на тестовой выборке для набора данных «Ирис». Используем первый метод для определения степени истинности правила, т. е. если антецеденты правил одинаковые, а консеквенты разные, то устранить конфликт поможет максимальная степень истинности.

Таблица 2

Тест на точность для набора данных «Ирис»

Номер испытания	Точность классификации		
	3 ФП	5 ФП	7 ФП
1	0,90	0,83	0,83
2	0,87	0,80	0,63
3	0,97	0,77	0,73
4	0,80	0,63	0,63
5	0,83	0,73	0,57
6	0,87	0,83	0,60
7	0,80	0,70	0,63
8	0,90	0,77	0,70
9	0,93	0,83	0,73
10	0,87	0,67	0,80
Среднее	**0,874**	**0,756**	**0,685**

Аналогичные эксперименты были проведены для набора данных «Уровень знаний студентов». Для обучения было выбрано 90 прецедентов. В табл. 3 показана точность классификации на тестовой выборке для набора данных «Уровень знаний студентов».

Таблица 3

Тест на точность для набора данных «Уровень знаний студентов»

Номер испытания	Точность классификации		
	3 ФП	5 ФП	7 ФП
1	0,76	0,65	0,59
2	0,72	0,62	0,59
3	0,72	0,62	0,55
4	0,76	0,59	0,52
5	0,72	0,65	0,55
6	0,68	0,65	0,55
7	0,76	0,59	0,52
8	0,70	0,68	0,59
9	0,68	0,62	0,55
10	0,76	0,59	0,52
Среднее	**0,726**	**0,626**	**0,553**

Согласно данным табл. 3, с помощью предложенной гибридной модели была достигнута хорошая классификационная точность для трех значений лингвистических переменных, при увеличении числа значений лингвистических переменных классификационная точность снижается, т. к. количество сочетаний становится больше и не все данные попадают в обучающую выборку. Из этого можно сделать вывод, что наибольшую классификационную точность можно получить, используя только три значения лингвистической переменной на этих наборах данных.

В табл. 4 показана точность классификации на тестовых данных «Ирис», полученная с использованием первого и второго метода для определения степени истинности правил.

Таблица 4

Тест на точность для набора данных «Ирис»

Номер испытания	Точность классификации 3 ФП		Точность классификации 5 ФП	
	Первый метод	Второй метод	Первый метод	Второй метод
1	0,80	0,90	0,83	0,86
2	0,83	0,90	0,70	0,79
3	0,90	0,87	0,73	0,83
4	0,80	0,87	0,93	0,93
5	0,87	0,93	0,83	0,93
6	0,83	0,93	0,86	0,79
7	0,93	0,90	0,80	0,87
8	0,87	0,93	0,76	0,80
9	0,87	0,83	0,76	0,83
10	0,90	0,87	0,83	0,86
Среднее	**0,860**	**0,893**	**0,803**	**0,849**

Согласно данным табл. 4, второй метод показывает лучшие результаты. Из этого можно сделать вывод, что наибольшую классификационную точность можно получить, используя 3 значения лингвистической переменной и второй метод определения степени истинности правил на наборе данных «Ирис».

Эффективность гибридной модели при сокращении прецедентов в обучающей выборке. Проведем эксперименты, в которых попробуем сократить количество прецедентов в обучающей выборке на наборе данных «Ирис». На основании предыдущего испытания мы выяснили, что оптимальное число значений лингвистических переменных равно 3. Значит, будем использовать 3 переменных и второй метод для определения степени истинности правил, т. к. он показывает лучшие результаты. Результаты экспериментов представлены на рис. 4.

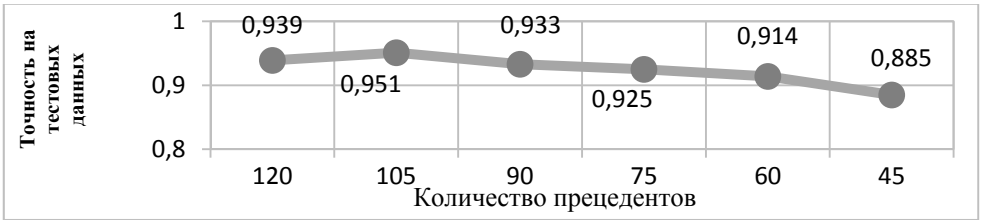

Рис. 4. Точность классификации прецедентов

На рис. 4 видно, что при значительном сокращении тестовых данных (до 45 прецедентов из 150) точность остается достаточно высокой. Это подтверждает гипотезу о том, что при небольшом количестве исходных данных с помощью предложенного алгоритма можно делать прогнозы с достаточной степенью точности.

Сравнение предлагаемого метода с традиционными методами

Предлагаемый нами метод мы сравнивали с некоторыми другими методами машинного обучения с точки зрения точности классификации (на тестовых данных) и числа случаев, которые используются для обучения. Результаты сравнительного анализа представлены в табл. 5.

Таблица 5

Сравнение предлагаемого метода с другими методами на тестовом наборе данных «Ирис»

Метод		Размер обучающей выборки	Размер тестовой выборки	Точность на тестовых данных
Предлагаемый метод		**120**	**30**	**0,939**
Метод кластеризации	*k*-средних с количеством кластеров 3	120	30	0,933
	k-средних с количеством кластеров 5	120	30	0,866
	k-means с количеством кластеров 7	120	30	0,866
	g-means	120	30	0,933
Предлагаемый метод		**105**	**45**	**0,951**
Метод кластеризации	*k*-средних с количеством кластеров 3	105	45	0,950
	k-средних с количеством кластеров 5	105	45	0,800
	k-means с количеством кластеров 7	105	45	0,950
	g-means	105	45	0,933
Предлагаемый метод		**90**	**60**	**0,933**
Метод кластеризации	*k*-средних с количеством кластеров 3	90	60	0,816
	k-средних с количеством кластеров 5	90	60	0,833
	k-means с количеством кластеров 7	90	60	0,916
	g-means	90	60	0,900
Предлагаемый метод		**75**	**75**	**0,925**
Метод кластеризации	*k*-средних с количеством кластеров 3	75	75	0,880
	k-средних с количеством кластеров 5	75	75	0,840
	k-means с количеством кластеров 7	75	75	0,920
	g-means	75	75	0,880
Предлагаемый метод		**60**	**90**	**0,914**
Метод кластеризации	*k*-средних с количеством кластеров 3	60	90	0,900
	k-средних с количеством кластеров 5	60	90	0,877
	k-means с количеством кластеров 7	60	90	0,900
	g-means	60	90	0,877
Предлагаемый метод		**45**	**105**	**0,885**
Метод кластеризации	*k*-средних с количеством кластеров 3	45	105	0,876
	k-средних с количеством кластеров 5	45	105	0,800
	k-means с количеством кластеров 7	45	105	0,876
	g-means	45	105	0,876

Согласно данным табл. 5, точность классификации предложенного метода близка к лучшему результату, полученному при использовании других методов машинного обучения. При сокращении размера выборки до минимально возможных размеров (45 прецедентов из 150) точность по сравнению с другими методами остается максимальной. Значит, используя этот метод, можно формировать правила на небольшом количестве прецедентов из базы знаний без потери степени точности.

Заключение

Таким образом, в ходе исследований получены следующие результаты:

– рассмотрена гибридная модель представления и извлечения знаний, представленных в виде прецедентов;

– предложена новая процедура аккумуляции заключений конкурирующих правил, полученных в результате преобразования множества прецедентов в систему нечетких правил;

– проведено исследование полученной гибридной модели на разных наборах данных и с разным количеством функций принадлежности;

– выявлена более высокая точность классификации прецедентов усовершенствованной гибридной модели по сравнению с ее предыдущей модификацией;

– показано, что при значительном сокращении тестовых данных предлагаемая модель сохраняет высокую степень точности классификации;

– сравнение предложенного метода классификации прецедентов, разработанного на основе формирования множества нечетких лингвистических правил, с другими методами машинного обучения, подтвердило его преимущества.

СПИСОК ЛИТЕРАТУРЫ

1. *DTI.* Knowledge-Based Systems Survey of UK Applications. Department of Trade & Industry, 1992, UK. 153 p.

2. *Schank R. C., Abelson R. P.* Scripts, plans, goals and understanding. Hillsdale, NJ: Erlbaum Associates, 1977. 256 p.

3. *Schank R.* Dynamic Memory: A theory of reminding and learning in computers and people. Cambridge University Press, 1982. 234 p.

4. *Wittgenstein L.* Philosophical Investigations. Blackwell, 1953. P. 31–34.

5. *Aamodt A., Plaza E.* Case-based reasoning: foundational issues, methodological variations, and system approaches // AI Communications. 1994. Vol. 7, no. 1. P. 39–59.

6. *Simpson R. L.* A Computer Model of Case-Based Reasoning in Problem Solving: An Investigation in the Domain of Dispute Mediation // Technical Report GIT-ICS-85/18, Georgia Institute of Technology, School of Information and Computer Science, 1985. P. 410–415.

7. *Hammond K. J.* CHEF: A model of case-based planning. In: Proc. American Association for Artificial Intelligence, AAAI-86, Philadelphia, PA, 1986. P. 267–271.

8. *Sharma S., Sleeman D.* REFINER: A case-based differential diagnosis aide for knowledge acquisition and knowledge refinement. In: EWSL 88; Proc. European Working Session on Learning, 1988. P. 201–210.

9. *Keane M.* Where's the beef? The absence of pragmatic factors in theories of analogy. In: ECAI, 1988. P. 327–332.

10. *Richter A. M., Weiss S.* Similarity, uncertainty and case-based reasoning in PATDEX. In: RS Boyer (ed.) Automated Reasoning, Essays in Honour of Woody Bledsoe. Kluwer, 1991. P. 249–265.

11. *Watson I. D., Abdullah S.* Developing case-based reasoning systems: a case study in diagnosing building defects. In: Proc. IEE Colloquium on Case-Based Reasoning: Prospects for Applications. Digest No: 1994/057. P. 1/1–1/3.

12. *Yang S., Robertson D.* A case-based reasoning system for regulatory information. In: Proc. IEE Colloquium on Case-Based Reasoning: Prospects for Applications. Digest No: 1994/057 P. 3/1–3/3.

13. *Moore C. J., Lehane M. S., Proce C. J.* Case-based reasoning for decision support in engineering design. In: Proc. IEE Colloquium on Case-Based Reasoning: Prospects for Applications. Digest No: 1994/057. P. 4/1–4/4.

14. *Oxman R. E.* PRECEDENTS: Memory structure in design case libraries. In: CAAD Futures, 1993. Elsevier. 152 p.

15. *Kitano H.* Challenges for massive parallelism. In: Proc. 13th. Conference on Artificial Intelligence, UCAI-93, 1993. P. 813–834.

16. *Юдин В. Н., Карпов Л. Е., Ватазин А. В.* Методы интеллектуального анализа данных и вывода по прецедентам в программной системе поддержки врачебных решений // Альманах клинической медицины. 2008, т. 17. Ч. 1. С. 266–269.

17. *Васильев В. И., Белков Н. В.* Система поддержки принятия решений по обеспечению безопасности персональных данных // Вестн. Уфим. гос. авиац. техн. ун-та. 2011. Т. 15, № 5 (45). С. 54–65.

18. *Варшавский П. Р., Еремеев А. П.* Методы правдоподобных рассуждений на основе аналогий и прецедентов для интеллектуальных систем поддержки принятия решений // Новости искусственного интеллекта. 2006. № 3. С. 39–62.

19. *Варшавский П. Р., Еремеев А. П.* Моделирование рассуждений на основе прецедентов в интеллектуальных системах поддержки принятия решений // Искусственный интеллект и принятие решений. 2009. № 2. С. 45–57.

20. *Еремеев А. П., Куриленко И. Е.* Реализация механизма временных рассуждений в современных интеллектуальных системах // Изв. РАН. Теория и системы управления. 2007. № 2. С. 120–136.

21. *Еремеев А. П., Куриленко И. Е.* Расширение возможностей моделирования временных зависимостей в интеллектуальных системах на основе применения темпоральных прецедентов // Интеллектуальные системы. Коллектив. моногр. Вып. 6; под ред. В. М. Курейчика. М.: Физматлит, 2013. С. 89–118.

22. *Авдеенко Т. В., Макарова Е. С.* Метод определения релевантности прецедентов на основе нечетких лингвистических правил // Науч. вестн. Новосибирск. гос. техн. ун-та. 2016. Т. 62, № 1. С. 17–34.

23. *Репозиторий* машинного обучения. URL: http://archive.ics.uci.edu/ml (дата обращения: 24.06.2016).

Статья поступила в редакцию 29.07.2016

ИНФОРМАЦИЯ ОБ АВТОРЕ

Макарова Екатерина Сергеевна – Россия, 630073, Новосибирск; Новосибирский государственный технический университет; аспирант кафедры экономической информатики; KATMC@yandex.ru.

E. S. Makarova

RESEARCH OF INFLUENCE OF PARAMETERS OF FUZZY CONTROL MODEL ON THE ACCURACY OF CLASSIFICATION OF PRECEDENTS

Abstract. Case-Based Reasoning (CBR) is used for the knowledge representation in the social and economic systems. The paper presents the literature review of case-based reasoning method and its application in various fields. The hybrid model of integrated case library and fuzzy logic inference is considered. The algorithm for generating fuzzy rules, where each linguistic input variable can take 3, 5 or 7 term-values, described by triangular membership functions, is examined. A new procedure for accumulating conclusions of conflicting rules obtained as a result of logical inference, is proposed. The classification accuracy of the proposed hybrid model has been studied for different data samples with different membership functions. The results of the experiments draw to the conclusion that the developed method of automatic training based on fuzzy output significantly increases the accuracy of the classification of precedents.

Key words: Case-Based Reasoning (CBR), precedent, fuzzy logic, fuzzy set, fuzzy rules, inference, knowledge base, process of accumulation.

REFERENCES

1. *DTI. Knowledge-Based Systems Survey of UK Applications*. Department of Trade & Industry, 1992, UK. 153 p.
2. Schank R. C., Abelson R. P. *Scripts, plans, goals and understanding*. Hillsdale, NJ: Erlbaum Associates, 1977. 256 p.
3. Schank R. *Dynamic Memory: A theory of reminding and learning in computers and people*. Cambridge University Press, 1982. 234 p.
4. Wittgenstein L. *Philosophical Investigations*. Blackwell, 1953, pp. 31–34.
5. Aamodt A., Plaza E. Case-based reasoning: foundational issues, methodological variations, and system approaches. *AI Communications*, 1994, vol. 7, no. 1, pp. 39–59.
6. Simpson R. L. *A Computer Model of Case-Based Reasoning in Problem Solving: An Investigation in the Domain of Dispute Mediation.* Technical Report GIT-ICS-85/18, Georgia Institute of Technology, School of Information and Computer Science, 1985, pp. 410–415.
7. Hammond K. J. *CHEF: A model of case-based planning.* In: Proc. American Association for Artificial In-telligence, *AAAI-86,* Philadelphia, PA, 1986, pp. 267–271.
8. Sharma S., Sleeman D. *REFINER: A case-based differential diagnosis aide for knowledge acquisition and knowledge refinement*. In: EWSL 88; Proc. European Working Session on Learning, 1988, pp. 201–210.
9. Keane M. *Where's the beef? The absence of pragmatic factors in theories of analogy*. In: ECAI, 1988, pp. 327–332.
10. Richter A. M., Weiss S. *Similarity, uncertainty and case-based reasoning in PATDEX.* In: RS Boyer (ed.) Automated Reasoning, Essays in Honour of Woody Bledsoe. Kluwer, 1991, pp. 249–265.
11. Watson I. D, Abdullah S. *Developing case-based reasoning systems: a case study in diagnosing building defects.* In: Proc. IEE Colloquium on Case-Based Reasoning: Prospects for Applications. Digest No: 1994/057, pp. 1/1–1/3.
12. Yang S., Robertson D. *A case-based reasoning system for regulatory information.* In: Proc. IEE Colloquium on Case-Based Reasoning: Prospects for Applications. Digest No: 1994/057, pp. 3/1–3/3.
13. Moore C. J., Lehane M. S., Proce C. J. *Case-based reasoning for decision support in engineering design*. In: Proc. IEE Colloquium on Case-Based Reasoning: Prospects for Applications. Digest No: 1994/057, pp. 4/1–4/4.
14. Oxman R. E. *PRECEDENTS: Memory structure in design case libraries*. In: CAAD Futures, 1993. Else-vier. 152 p.
15. Kitano H. *Challenges for massive parallelism*. In: Proc. 13th. Conference on Artificial Intelligence, UCAI-93, 1993, pp. 813–834.
16. Iudin V. N., Karpov L. E., Vatazin A. V. Metody intellektual'nogo analiza dannykh i vyvoda po pretsedentam v programmnoi sisteme podderzhki vrachebnykh reshenii [Data mining methods and conclusion on the precedents in the software support system of medical solutions]. *Al'manakh klinicheskoi meditsiny*, 2008, vol. 17, part 1, pp. 266–269.
17. Vasil'ev V. I., Belkov N. V. Sistema podderzhki priniatiia reshenii po obespecheniiu bezopasnosti personal'nykh dannykh [System of support of decision-making to ensure the security of personal data]. *Vestnik Ufimskogo gosudarstvennogo aviatsionnogo tekhnicheskogo universiteta*, 2011, vol. 15, no. 5 (45), pp. 54–65.

18. Varshavskii P. R., Eremeev A. P. Metody pravdopodobnykh rassuzhdenii na osnove analogii i pretsedentov dlia intellektual'nykh sistem podderzhki priniatiia reshenii [Methods of plausible reasoning on the basis of analogies and precedents for intellectual systems of support of decision-making]. *Novosti iskusstvennogo intellekta*, 2006, no. 3, pp. 39–62.

19. Varshavskii P. R., Eremeev A. P. Modelirovanie rassuzhdenii na osnove pretsedentov v intellektual'nykh sistemakh podderzhki priniatiia reshenii [Modeling of reasoning on the basis of precedents in intellectual systems of decision-making support]. *Iskusstvennyi intellekt i priniatie reshenii*, 2009, no. 2, pp. 45–57.

20. Eremeev A. P., Kurilenko I. E. Realizatsiia mekhanizma vremennykh rassuzhdenii v sovremennykh intellektual'nykh sistemakh [Implementation of temporal reasoning in over-time intelligence systems]. *Izvestiia RAN. Teoriia i sistemy upravleniia*, 2007, no. 2, pp. 120–136.

21. Eremeev A. P., Kurilenko I. E. Rasshirenie vozmozhnostei modelirovaniia vremennykh zavisimostei v intellektual'nykh sistemakh na osnove primeneniia temporal'nykh pretsedentov [Empowerment of modeling temporal relationships in intelligent systems through the application of temporal precedents]. *Intellektual'nye sistemy*. Kollektivnaia monografiia. Vypusk 6. Pod red. V. M. Kureichika. Moscow, Fizmatlit, 2013. P. 89–118.

22. Avdeenko T. V., Makarova E. S. Metod opredeleniia relevantnosti pretsedentov na osnove nechetkikh lingvisticheskikh pravil [Method of determining the relevance of precedents based on fuzzy linguistic rules]. *Nauchnyi vestnik Novosibirskogo gosudarstvennogo tekhnicheskogo universiteta*, 2016, vol. 62, no. 1, pp. 17–34.

23. *Repozitorii mashinnogo obucheniia* [Repository of machine learning]. Available at: http://archive.ics.uci.edu/ml (accessed: 24.06.2016).

The article submitted to the editors 29.07.2016

INFORMATION ABOUT THE AUTHOR

Makarova Ekaterina Sergeevna – Russia, 630073, Novosibirsk; Novosibirsk State Technical University; Postgraduate Student of the Department of Economic Informatics; KATMC@yandex.ru.