

УДК 004.85

Метод определения релевантности прецедентов на основе нечетких лингвистических правил^{*}

Т.В. АВДЕЕНКО¹, Е.С. МАКАРОВА²

¹ 630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, доктор технических наук, профессор. E-mail: tavdeenko@mail.ru

² 630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, аспирант. E-mail: KATMC@yandex.ru

Прецедентный подход (*Case-Based Reasoning, CBR*) является полезным инструментом представления знаний в социально-экономических системах, так как использует рассуждения по аналогии, свойственные человеческому интеллекту. Традиционный подход к оценке подобия прецедентов заключается в вычислении взвешенной метрики – расстояния в многомерном пространстве. В статье вместо подобия прецедентов рассматривается их релевантность на основе нечеткой логики. Нечеткие лингвистические правила используются в качестве инструмента для представления знаний о релевантности прецедентов. Рассматривается процедура настройки нечетких правил на основе обучения по выборкам прецедентов. Исследуются треугольные и квадратичные функции принадлежности. Проведенные исследования показывают приемлемую точность классификации прецедентов даже для малой обучающей выборки. При этом более гладкие (квадратичные) функции принадлежности показывают в среднем большую точность предсказания. Введение нечетких правил в качестве критериев оценки подобия приносит следующие существенные преимущества в прецедентную модель: нечеткие правила «если-то» представляют собой более общую и гибкую платформу для описания знаний о сходстве между прецедентами, чем расстояние в многомерном пространстве; сущность базы знаний, содержащей нечеткие правила, позволяет приобретать и интегрировать знания из различных источников; нечеткая база знаний является хорошо понятной за счет интерпретации отдельных языковых правил, следовательно, пользователи могут легко понять, как и почему прецедент был выбран для поиска, содержательно подтвердив нечеткие правила, полученные в результате обучения. Это создает возможность динамического участия человека во взаимодействии с системой, основанной на прецедентах, для адаптации поведения системы к конкретным требованиям.

Ключевые слова: прецедентный подход, база знаний, нечеткая логика, нечеткое множество, лингвистическая переменная, нечеткие правила, рассуждения по аналогии, функция принадлежности

DOI: 10.17212/1814-1196-2016-1-17-34

^{*} Статья получена 14 ноября 2015 г.

ВВЕДЕНИЕ

Методы искусственного интеллекта в последнее время завоевывают все большую популярность при решении задач управления и принятия решений. Это объясняется тем, что традиционно используемые в этой области классические методы оптимизации часто не приводят к желаемому результату, так как не учитывают особенности поведения ЛПР. В этом смысле прецедентный подход (*Case-Based Reasoning, CBR*) является полезным инструментом представления знаний в социально-экономических системах, так как использует рассуждения по аналогии, свойственные человеческому интеллекту.

Прецедентный подход позволяет решить новую задачу, используя или адаптируя решение какой-либо ранее уже известной задачи. Когда рассматривается новая проблема или ситуация, система находит подобный прецедент в базе знаний прецедентов и пытается использовать его решение для текущей проблемы. При необходимости система адаптирует найденный прецедент к новой ситуации. Затем происходит обработка текущего случая, после чего прецедент и его решение заносится в базу прецедентов для его дальнейшего использования. Метод машинного обучения, основанный на выявлении сходства, получил название «обучение по аналогии». Данный метод базируется на эвристике, утверждающей, что аналогичные проблемы имеют аналогичные решения, и имеет целью собрать локальные знания из близких прецедентов и приспособить их для решения новой проблемы.

Область возможного использования прецедентного подхода достаточно широка. Например, в сфере консалтинга часто приходится оказывать помощь пользователям на основе опыта, полученного при решении аналогичных задач в прошлом. Различные консультанты могут сталкиваться с аналогичными проблемами, поэтому использование своего или чужого опыта в поддержке принятия решений на основе метода прецедентов может оказаться весьма эффективным средством, значительно сокращающим драгоценное время, необходимое для принятия решений.

Определение локального сходства между прецедентами является ключевой задачей для систем, основанных на прецедентном подходе. До настоящего времени основное направление исследований работ при создании моделей подобия было ориентировано на построение весовой функции подобия прецедентов, развивающей концепцию расстояния в многомерном пространстве. При этом характеристикам (признакам) прецедентов присваиваются различные веса в соответствии с их значимостью, а глобальная функция подобия определяется как взвешенная сумма локальных значений коэффициентов. В литературе предлагались различные методы автоматического выявления весов [1]. Метод пошагового обучения для модификации весов признаков на основе обратной связи, извлеченных из прецедентов результатов, был предложен в работе [2]. В исследовании [3] рассматривался принцип вероятностного рейтинга для нахождения значений весовой функции. Информация о ранге прецедента была использована в работе [4] для адаптации весов таким образом, чтобы ранг извлеченных прецедентов соответствовал заданному значению. Введение критерия повышения точности представляет собой еще один способ для адаптации алгоритма приращения веса, как предлагалось в [5]. Тем не менее независимо от того, как значения весовых коэффициентов получены, возможность сходства этих методов по своей сущности ограниче-

на весовыми комбинациями сопоставления локальных метрик. Это ограничение затрудняет представление более общих знаний и критериев оценивания прецедентов с глобальной содержательной точки зрения.

Действительно, в реальной ситуации может оказаться, что понастоящему полезные решения располагаются в прецедентах, расположенных достаточно далеко (в смысле расстояния) от текущей ситуации. Выявление возможности подобных прецедентов передавать полезную информацию для решения новой проблемы представляет собой важнейшую задачу для исследований. Кроме того, поиск соседних прецедентов является слишком ограниченным для применения прецедентного подхода в ситуациях, когда никаких случаев, похожих (в смысле близости) на целевую задачу, в базе прецедентов нет. В этом случае необходимо искать решение новой проблемы в удаленных прецедентах, приспособлявая их к текущей ситуации.

В данной статье исследуется прецедентный подход, не ограниченный отношением подобия (близости) для поиска прецедентов. Подход вводит отношение релевантности между двумя произвольными случаями. Два прецедента являются релевантными, если решение в одном случае можно напрямую использовать для другого случая. Релевантность прецедентов не зависит от их близости расположения. При этом используется глобальная информация о разнообразных прецедентах, в том числе и расположенных далеко друг от друга. Использование далеко находящихся друг от друга прецедентов позволяет преодолеть главный недостаток традиционного прецедентного подхода, когда подобный опыт для решения новых проблем не доступен.

Новая модель сходства без функции взвешивания впервые была предложена в работах [6–7] как попытка поиска более мощного инструментария представления знаний для реализации механизма поиска на основе прецедентного подхода. Идея заключалась в том, чтобы закодировать информацию о значимых свойствах в комбинациях локальных метрик таким образом, чтобы функция взвешивания больше была не нужна.

Интеграция теории нечетких множеств с прецедентным подходом рассматривалась некоторыми исследователями. В исследовании [8] разъяснено, что существует тесная связь между нечеткой системой моделирования и теорией прецедентов; в [9] формализована фундаментальная гипотеза о прецедентном подходе в условиях нечетких правил. Авторы создали формальный подход в этом случае на основе логического вывода, который может быть реализован как особый тип нечетких множеств на основе приближенных рассуждений.

Нечеткие множества и нечеткая логика используются также для представления прецедентов и сопоставления прецедентов в процессе реализации прецедентной модели. Нечеткие множества были использованы, чтобы изобразить неточные характеристики прецедента в нечетких рассуждениях на основе метода прецедентов [10]. В работе [11] нечеткое подмножество «маленький» было определено на различных атрибутах для численной оценки сходства между прецедентами. Кроме того, нечеткие лингвистические правила были приняты в [12] и [13] в качестве гибких средств для выражения критериев оценки подобия.

В настоящей статье рассматривается метод определения релевантности прецедентов на основе нечетких лингвистических правил. В следующем разделе мы приводим описание гибридной модели, объединяющей базу прецедентов

и нечеткий логический вывод из базы знаний. В разделе 2 приводится метод машинного обучения системы нечетких правил на основе выборки прецедентов. В разделе 3 приводится анализ результатов вычислительных экспериментов, проведенных на тестовой выборке прецедентов, с использованием описываемого в статье математического и программного инструментария.

1. ГИБРИДНАЯ ПРЕЦЕДЕНТНАЯ МОДЕЛЬ ПРЕДСТАВЛЕНИЯ ЗНАНИЙ НА ОСНОВЕ НЕЧЕТКОЙ ЛОГИКИ

В исследованиях по искусственному интеллекту прецедентный подход (*CBR*) является одним из ключевых направлений, которое достаточно быстро развивается. В работе [1] вводится следующее определение прецедента: «прецедент – это описание проблемы или ситуации в совокупности с подробным указанием действий, предпринимаемых в данной ситуации или для решения данной проблемы».

Таким образом, прецедент как единица знания включает в себя:

- описание ситуации;
- решение, которое было принято в этой ситуации;
- результат применения решения.

Описание ситуации содержит всю информацию, которая необходима для достижения цели (выбора наиболее подходящего решения). Например, если целью является быстрое получение ответа на вопрос пользователя, то необходимо получить максимальную информацию о проблеме: в какой области возник вопрос, в каком документе (идентификационная информация); роль пользователя (определяет набор допустимых функций); предполагаемый результат; контактная информация пользователя.

Решение содержит набор операций, которые необходимо выполнить для получения успешного результата; последовательность действий, которую можно осуществить, чтобы избежать отказа, и т. д.

Результат применения решения – это обратная связь, которая возникает при применении решения из прецедента к текущей ситуации. В описание результата могут входить ссылки на другие прецеденты, текстовая и другая информация.

Прецедент может содержать как положительный, так и отрицательный исход. Если текущая проблема не решена при выполнении определенной последовательности действий, которая указана в прецеденте, то необходимо поместить прецедент в архив. Под архивом понимается место, где хранятся прецеденты, которые не актуальны для текущей версии системы. Удалять прецеденты нет необходимости, так как последовательность действий может показать направление дальнейших действий для решения вопроса.

Прецедент можно представить различными способами: от описания множества фактов в виде строк в базе данных (таблица «объект–свойство») до сложной иерархической фреймовой структуры, которая опирается на модель предметной области. Важно понимать, что выбирать представление прецедента необходимо исходя из общих целей системы. Главной проблемой при представлении прецедента является выбор необходимой информации, которую необходимо включить в описание прецедента для осуществления быстрого и удобного поиска.

Механизм принятия решения на основе прецедентов представлен на рис. 1. Сначала выявляется некоторая ситуация, далее эта ситуация сопоставляется с информацией из базы прецедентов, на основании чего принимается решение. Соответствующая информация формирует контейнер, который называется прецедентом и сохраняется в базу прецедентов для использования в последующих ситуациях. Ситуация для которой был сохранен прецедент, называется опорной, или базовой [14].

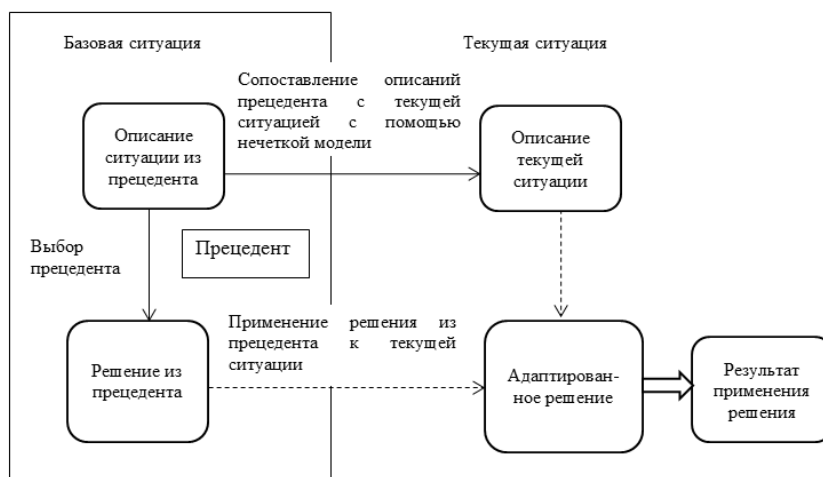


Рис. 1. Механизм принятия решения на основе прецедентов

Таким образом, процесс функционирования прецедентной системы можно представить с помощью следующего CBR-цикла [15]:

- 1) получение (выбор) из хранилища наиболее уместного прецедента или множества прецедентов на основе заданного отношения подобия;
- 2) использование выбранных прецедентов для принятия решения;
- 3) пересмотр и коррекция (адаптация) в случае необходимости принимавшихся ранее решений в выбранных прецедентах;
- 4) сохранение в хранилище принятого решения и сложившейся ситуации в качестве нового прецедента или соответствующее изменение выбранного прецедента, что может быть полезным в дальнейшем при решении аналогичных задач.

Опыт реального внедрения прецедентных систем в медицинской практике выявляет следующие их достоинства [16].

– **Когнитивная адекватность.** В системах, основанных на прецедентах, используется способ, которым специалисты «рассуждают» при решении проблемы.

– **Очевидная настройка под предметную область.** Прецедентная система легко настраивается к особенностям специфичных предметных областей.

– **Двойственность объективных и субъективных знаний.** Системы, основанные на прецедентах, не только используют знания одного или даже нескольких экспертов (как это делается для экспертных систем, например, основанных на правилах), они строятся на основе существующих случаев из практики (которые могут и не быть до конца понятны).

– **Автоматическое получение знаний.** Системы на основе прецедентов обладают способностью к самообучению на основе тех знаний, которые они априорно включают.

Основной задачей, которую необходимо решить в прецедентном подходе, является выбор метода, с помощью которого вычисляется мера сходства (подобия). Метод задается на этапе создания *СВР*-системы. Часто используется метод ближайшего соседа, в основе которого лежит измерение степени совпадения значений свойств, определяющих прецедент. В работах [2–5] предлагались меры, основанные на введении весовой функции, учитывающей значимость каждого из признаков. В настоящей статье рассматривается понятие релевантности, основанное на аппарате нечеткой логики.

Нечеткая логика – это обобщение традиционной логики в случае, когда истинность рассматривается как лингвистическая переменная, принимающая значения типа «очень истинно», «более-менее истинно», «не очень ложно» и т. п. [17]. Указанные лингвистические значения представляются нечеткими множествами.

Под лингвистической переменной будем понимать переменную, которая для своего задания использует лингвистические значения, которые являются качественными характеристиками. Например, температура воды и скорость – это лингвистические переменные; «холодная», «прохладная», «теплая», «горячая» – это лингвистические значения (характеристики); «примерно 45 градусов», «около 60 км/ч» – это значения нечетких чисел.

Для определения некоторой лингвистической переменной используется лингвистическое терм-множество, которое содержит в себе все лингвистические значения. Под областью значений переменной понимается множество всех числовых значений, которые могут относиться к определенному параметру изучаемой системы, а также множество значений, которые важны с точки зрения решаемой задачи. Лингвистические значения представляются нечеткими множествами (НМ), которые представляют собой совокупность элементов произвольной природы, относительно которых нельзя с полной определенностью утверждать, что какой-либо элемент принадлежит данному множеству [18].

Пусть имеется некоторое обычное (универсальное, или универсум) множество X элементов x . Нечеткое множество A определяется как упорядоченное множество пар вида $\langle x, \mu_A(x) \rangle$, где $x \in X$ – является элементом некоторого универсального множества X (универсума), $\mu_A(x)$ – функция принадлежности $\mu_A: X \rightarrow [0,1]$. При этом $\mu_A(x) = 1$ для некоторого x означает, что элемент x определенно принадлежит нечеткому множеству A , а значение $\mu_A(x) = 0$ означает, что элемент x определенно не принадлежит нечеткому множеству A .

Формально конечное нечеткое множество записывается в виде

$$A = \{ \langle x_1, \mu_A(x_1) \rangle, \langle x_2, \mu_A(x_2) \rangle, \dots, \langle x_n, \mu_A(x_n) \rangle \}.$$

Общее представление гибридной системы, основанной на прецедентном подходе и нечеткой модели, представлено на рис. 2. Работа системы начинается с описания текущей ситуации Q , в которой надо принять решение, с использованием тех же признаков, которые использовались для описания прецедентов C в базе прецедентов CB . Для текущей ситуации система должна найти релевантное решение в базе прецедентов [19]. Релевантность между

текущей ситуацией и прецедентами в базе прецедентов устанавливается с использованием нечеткого логического вывода, использующего множество нечетких правил, которые содержат знания для оценки релевантности между случаями. Выбираются только те прецеденты из базы прецедентов CB , степень релевантности которых выше заданного порогового уровня α .

Таким образом, множество прецедентов, извлеченных из базы, формулируются как $M = \{C \in CB \mid \text{Relev}(C, Q) \geq \alpha\}$, где $\text{Relev}(C, Q)$ – функция сходства прецедента C с текущей ситуацией Q , вычисляемая на основе нечетких рассуждений. Выбранные прецеденты в M передаются на следующий блок «слияния решений», где решения полученных прецедентов объединяются и адаптируются для разрешения текущей ситуации.

Целью шага «слияния решений» является нахождение новых решений для текущей проблемы путем модификации и обобщения известных решений, полученных из прецедентов. Подобие прецедентов определяется с помощью показателей полезности или целесообразности выбранных прецедентов для решения новой проблемы. Прецеденты с высоким значением мощности будут иметь больше влияния при определении окончательного решения.

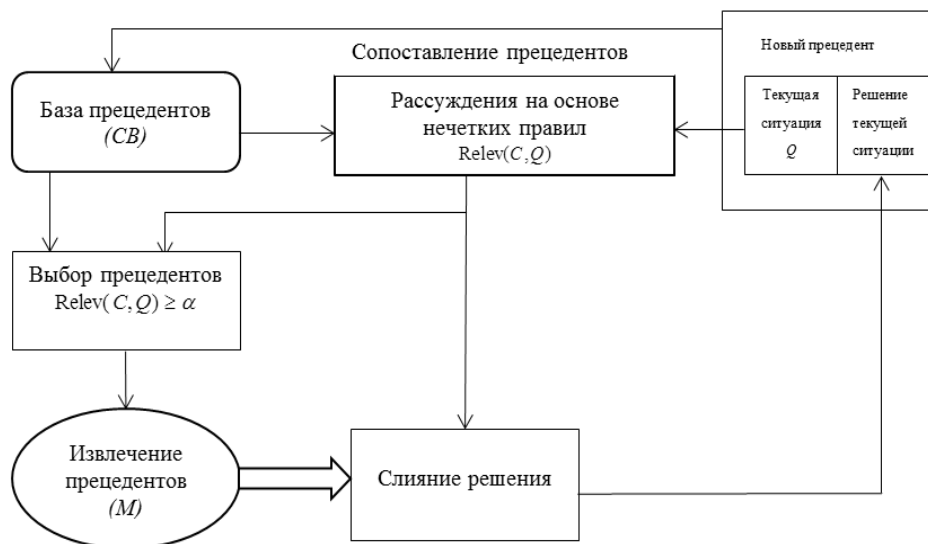


Рис. 2. Схема работы гибридной системы на основе нечеткой логики

Обобщенная схема работы нечеткого модуля представлена на рис. 3. Первоначально на вход системы поступает выборка прецедентов из базы прецедентов. Каждый из прецедентов характеризуется набором признаков. Для проверки работы алгоритма исходная выборка случайным образом делится на две подвыборки – обучающую и тестовую. На первом этапе происходит настройка базового множества нечетких правил под имеющиеся объективные данные обучающей выборки прецедентов. На следующем этапе проводится проверка точности классификации прецедента на имеющейся тестовой выборке. Для этого каждый элемент тестовой выборки рассматривается как текущая ситуация, для которой на основе нечеткого вывода находится соответствующий прецедент из обучающей выборки. Далее полученное в ре-

зультате такой процедуры решение сравнивается с решением прецедента, после чего делаются выводы о точности предсказания.

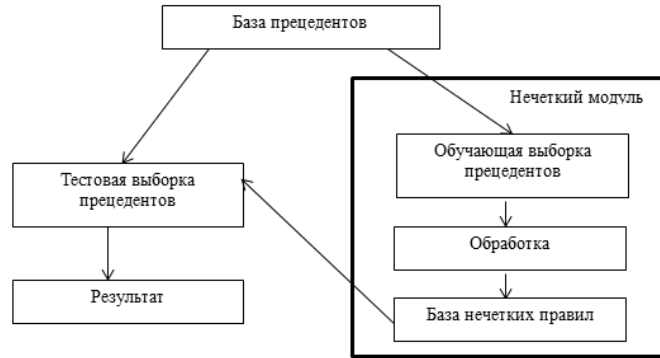


Рис. 3. Обобщенная схема работы нечеткого модуля

В следующем разделе описывается формальная процедура настройки системы нечетких правил на основе обучающей выборки из базы прецедентов.

2. НАСТРОЙКА СИСТЕМЫ НЕЧЕТКИХ ПРАВИЛ НА ОСНОВЕ ОБУЧАЮЩЕЙ ВЫБОРКИ ПРЕЦЕДЕНТОВ

Рассмотрим структуру нечетких правил, которые используются в гибридной модели для определения релевантности прецедентов. Предположим, что имеется n свойств (атрибутов) для каждой проблемы (прецедента) в предметной области. Таким образом, прецедент C в базе прецедентов индексируется как $(n+1)$ -мерный кортеж $C = (x_1, x_2, x_3, \dots, x_n, S)$, где $x_1, x_2, x_3, \dots, x_n$ – значения атрибутов прецедента; S – решение проблемы, определенное в прецеденте.

Аналогичным образом n -мерный кортеж $(y_1, y_2, y_3, \dots, y_n)$ используется для представления значений соответствующих атрибутов текущей проблемы Q . Обозначим FS_j – множество терм-значений лингвистической переменной, используемой для описания j -го атрибута текущей проблемы.

Тогда нечеткое правило, используемое для оценки релевантности прецедента, может быть сформулировано следующим образом:

$$\begin{aligned} & IF [(x_1 = A_1) \wedge (y_1 = B_1)] \text{ and } [(x_2 = A_2) \wedge (y_2 = B_2)] \text{ and } \dots \text{ and} \\ & [(x_n = A_n) \wedge (y_n = B_n)] \text{ then } Relevance = V \end{aligned} \quad (1)$$

с $A_j \in FS_j$, $B_j \in FS_j$ для $j=1 \dots n$ и $V \in \{1.0, 0\}$.

Особым свойством правила в выражении (1) является то, что оно учитывает пару значений атрибутов в каждой части правила, заключенной в квадратные скобки.

Введем в рассмотрение мощность правила, формируемого на основе соответствующего элемента обучающей выборки (строки таблицы «объект–свойство»):

$$t = \min_{j=1 \dots n} (\mu_{A_j}(x_j) * \mu_{B_j}(y_j)).$$

Чем больше мощность правила, тем большее соответствие мы имеем с видом (1).

С набором нечетких правил вида (1) степень релевантности между прецедентом C и текущей проблемой Q может быть рассчитана следующим образом:

$$Relev(C, Q) = \frac{\sum_{\forall k} t_k(C, Q) * V_k}{\sum_{\forall k} t_k(C, Q)},$$

где V_k – одиночное заключение (0 или 1) для правила R_k ; t_k – мощность правила R_k .

Ниже рассмотрен алгоритм формирования нечетких лингвистических правил на основе обучающей выборки, состоящей из прецедентов.

АЛГОРИТМ ФОРМИРОВАНИЯ НЕЧЕТКИХ ПРАВИЛ

Шаг 1. Задание типа функции принадлежности. На этом шаге может быть выбран тип функции. Для представления нечетких множеств используются различные типы функций принадлежности. В рамках поставленной задачи были рассмотрены треугольная и s -образная функции принадлежности (ФП). Треугольную ФП будем называть ФП типа T , а s -образную, соответственно, ФП типа S .

Треугольная и s -образная функции принадлежности задаются аналитически следующими выражениями [20]:

$$f_{\Delta}(x; a, b, c) = \begin{cases} 0, & x \geq a \\ \frac{x-a}{b-a}, & a \leq x \leq b \\ \frac{c-x}{c-b}, & b \leq x \leq c \\ 0, & c \leq x \end{cases}, \quad f_S(x; a, b, c) = \begin{cases} 0, & x \leq a \\ 2\left(\frac{x-a}{b-a}\right)^2, & a < x \leq \frac{a+b}{2} \\ 1-2\left(\frac{b-x}{b-a}\right)^2, & \frac{a+b}{2} < x < b \\ 1, & b \leq x. \end{cases}$$

Шаг 2. Загрузка исходных данных. На этом шаге загружаем наборы обучающих данных в программу. Предусмотрена возможность загрузки файла в txt-формате, или в формате excel. Тип формата файла не влияет на скорость загрузки данных.

Шаг 3. Разделение исходных данных на множества, соответствующие классам. Перебираем все загруженные записи, анализируем колонку «Результат». Группируем множества в зависимости от класса, подсчитываем количество записей в каждом множестве, выводим результат пользователю.

Шаг 4. Отбор тестовых данных случайным образом. Пользователь вводит количество записей, которое будет отобрано случайным образом в обучающую выборку.

Шаг 5. Определение и разбиение пространства параметров на области. Определим минимальное и максимальное значения для каждого параметра на обучающей выборке. Каждый полученный интервал разделим на $(2N + 1)$ областей. Для первого атрибута каждую область обозначим следующим образом: M_{1n} (Малый 1n), ..., M_{11} (Малый 11), S_1 (Средний1), B_{11} (Большой 11), ..., B_{1n} (Большой 1n). Для второго атрибута будет обозначение: M_{2n} (Малый 2n), ..., M_{21} (Малый 21), S_2 (Средний 2), B_{21} (Большой 21), ..., B_{2n} (Большой 2n). И для каждого из них определяем функцию принадлежности. На рис. 4, а представлен пример такого разделения, где область определения параметра разделена на три области ($N = 1$) и функции принадлежности типа T , а на рис. 4, б представлен пример ФП для типа S .

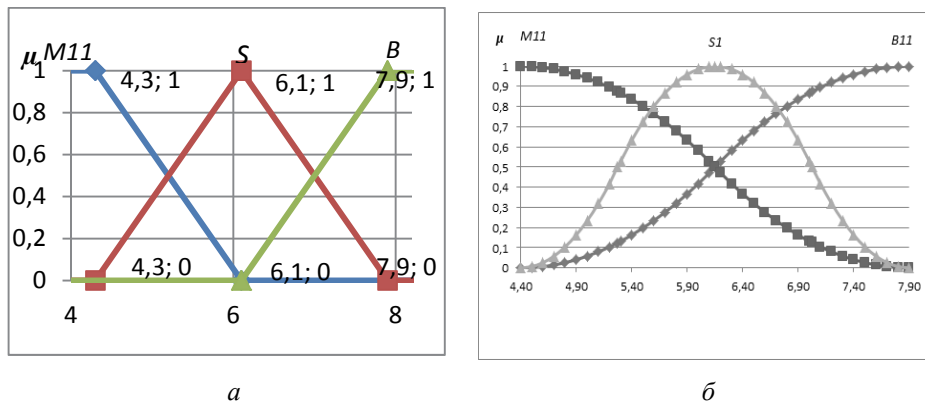


Рис. 4. Функции принадлежности:

а – тип T ; б – тип S

Шаг 6. Формирование нечетких правил на основе обучающей выборки. Для каждой области, определенной на шаге 5, определим степени принадлежности обучающих данных. Для каждой группы данных степень принадлежности будет выражаться значениями функций принадлежности соответствующих НМ. Например, для первого параметра и ФП типа T значение $x_1 = 5,4$ имеет степень принадлежности 0,4 к M_{11} ; 0,6 к S_1 ; 0 к B_{11} . А для параметра A и ФП типа S значение $x_1 = 5,4$ имеет степень принадлежности 0,16 к M_{11} ; 0,84 к S_1 ; 0,64 к B_{11} . Проведем сопоставление обучающей выборки областям, в которых степени принадлежности имеют максимальные значения. Окончательно для ФП типа T можно записать правило:

$$(x_1, x_2, x_3, x_4; R) \Rightarrow$$

$$x_1[\max 0,6 \text{ в } S_1], x_2[\max 0,6 \text{ в } M_{21}], x_3[\max 0,5 \text{ в } B_{31}], x_4[\max 0,7 \text{ в } S_4];$$

$$R[\max 0,9 \text{ в } R_1]$$

$$\text{Правило 1. IF}(x_1 = S_1, x_2 = M_{21}, x_3 = B_{31}, x_4 = S_4) \text{ Then } R = R_1$$

А для ФП типа S можно записать правило:

$$(x_1, x_2, x_3, x_4; R) \Rightarrow$$

$$x_1[\max 0,84 \text{ в } S_1], x_2[\max 0,75 \text{ в } B_{21}], x_3[\max 0,46 \text{ в } M_{31}], x_4[\max 0,83 \text{ в } S_4];$$

$$R[\max 0,89 \text{ в } R_1]$$

$$\text{Правило 2. IF}(x_1 = S_1, x_2 = B_{21}, x_3 = M_{31}, x_4 = S_4) \text{ Then } R = R_1$$

Шаг 7. Каждому правилу присвоим степени истинности. Учитывая большое количество обучающих данных, по которым может быть сформулировано такое же большое количество правил, существует высокая вероятность противоречивости правил. Если антецеденты правил одинаковые, а консеквенты разные, то решить противоречие поможет степень истинности. Из пары правил с одинаковыми условиями выбирается правило с максимальной степенью истинности. Благодаря этому условию разрешается проблема противоречивости в базе правил и сильно уменьшается их количество.

Правило. Если x_1 это A_1 , x_2 это A_2 , то y это B , степень истинности обозначим как $SP(R) = \max(\mu A_1, \mu A_2, \mu B)$. Правило 1 из примера имеет $SP(R) = \max(0,6, 0,6, 0,5, 0,7) = 0,7$. Для правила 2 степень истинности $SP(R) = \max(0,84, 0,75, 0,46, 0,83) = 0,84$.

Шаг 8. Создание базы нечетких правил. База нечетких правил представляется списком правил:

$$1. \text{ IF}(x_1 = B_{11}, x_2 = M_{21}, x_3 = M_{31}, x_4 = B_{41}) \text{ Then } R = R_1.$$

$$2. \text{ IF}(x_1 = B_{12}, x_2 = S_2, x_3 = B_{31}, x_4 = B_{41}) \text{ Then } R = R_2.$$

$$3. \text{ IF}(x_1 = M_{11}, x_2 = B_{21}, x_3 = S_3, x_4 = B_{41}) \text{ Then } R = R_1.$$

$$4. \text{ IF}(x_1 = S_1, x_2 = M_{21}, x_3 = M_{31}, x_4 = M_{41}) \text{ Then } R = R_1.$$

$$5. \text{ IF}(x_1 = S_1, x_2 = B_{21}, x_3 = M_{31}, x_4 = B_{41}) \text{ Then } R = R_2.$$

Обобщим рассмотренный выше метод на случай нечеткой системы с произвольным количеством входов и выходов. На рис. 5 представлена блок-схема алгоритма создания базы правил, которая является основой для программной реализации.

3. ЭКСПЕРИМЕНТАЛЬНЫЕ ИССЛЕДОВАНИЯ МЕТОДА

В данном разделе будут показаны результаты экспериментальных исследований рассмотренного метода обучения системы нечетких правил на известном эталонном наборе данных «Ирис», который может быть загружен с веб-страницы <http://archive.ics.uci.edu/ml/>. Была рассмотрена выборка, содержащая 150 объектов (прецедентов), которые характеризуются четырьмя атрибутами. Набор данных изначально был классифицирован по трем классам: «Ирис» вида *Setosa*, «Ирис» *Versicolour*, «Ирис» *Virginica*.

3.1. ОБУЧЕНИЕ НЕБОЛЬШОЙ БАЗЫ ПРЕЦЕДЕНТОВ

В экспериментах из набора данных «Ирис» были случайным образом выбраны данные для обучения (небольшая база прецедентов). Остальные данные были использованы в качестве тестовых данных. Нечеткие правила, полученные в результате обучения извлеченных из базы прецедентов, были

применены к прецедентам для классификации случаев на тестовых данных. Компьютерный эксперимент для этого набора данных был проведен 25 раз. При увеличении числа испытаний от 5 до 25 среднее значение точности не изменялось. Таблица 1 показывает точность классификации на тестовых задачах для данных «Ирис».

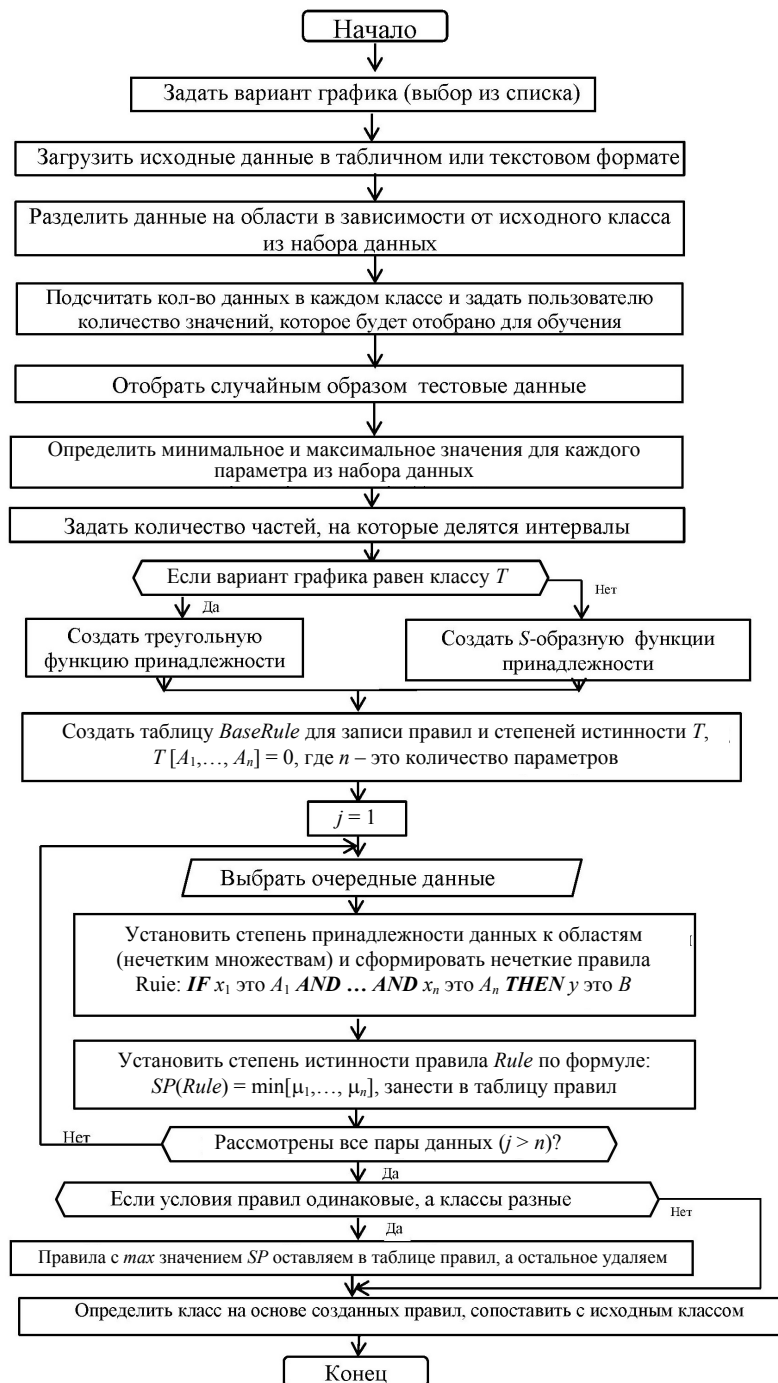


Рис. 5. Блок-схема построения базы правил

Таблица 1

Тест на точность набора данных «Ирис»

Номер испытания	Исходная выборка	Обучающая выборка	Точность классификации	
			ФП типа T	ФП типа S
1	150	120	0,90	0,97
2	150	120	0,87	0,97
3	150	120	0,93	0,93
4	150	120	0,87	0,97
5	150	120	0,87	0,83
6	150	120	0,97	0,93
7	150	120	0,93	0,89
8	150	120	0,97	0,91
9	150	120	0,97	0,89
10	150	120	0,89	0,97
11	150	120	0,84	0,91
12	150	120	0,82	0,87
13	150	120	0,84	0,91
14	150	120	0,89	0,91
15	150	120	0,80	0,91
16	150	120	0,89	0,89
17	150	120	0,87	0,93
18	150	120	0,87	0,97
19	150	120	0,93	0,91
20	150	120	0,89	0,91
21	150	120	0,87	0,93
22	150	120	0,96	0,96
23	150	120	0,87	0,96
24	150	120	0,91	0,98
25	150	120	0,84	0,91
Среднее	150	120	0,893	0,928

Из табл. 1 видно, что с помощью рассматриваемой модели была достигнута хорошая классификационная точность для обоих типов функции принадлежности. Под достаточной классификационной точностью будем понимать уровень точности более 0,800. Среднее значение точности для функций принадлежности типа S оказалось выше, чем для типа T . Это можно объяснить тем, что функции принадлежности типа S более гладкие (не содержат разрывов производных) по сравнению с функциями принадлежности типа T .

3.2. УСТОЙЧИВОСТЬ ПРОТИВ РЕЗКОГО СОКРАЩЕНИЯ ПРЕЦЕДЕНТОВ

Интересное явление наблюдалось в экспериментах, в которых после обучения нечетких правил оригинальная база прецедентов могла быть уменьшена до малых размеров и при этом предоставляла высокую точность классификации для прецедентной системы. Подтверждение результатов можно увидеть в табл. 2–5. Исходный размер базы прецедентов для всех испытаний 150.

Таблица 2

**Точность на наборе данных
«Ирис» при обучающей выборке
90 прецедентов**

Номер испытаний	Обучаю- щая выборка	Точность на тестовых данных	
		ФП типа T	ФП типа S
1	90	0,860	0,920
2	90	0,910	0,930
3	90	0,810	0,890
4	90	0,820	0,880
5	90	0,830	0,800
6	90	0,900	0,950
7	90	0,880	0,820
8	90	0,920	0,950
9	90	0,800	0,970
10	90	0,850	0,920
11	90	0,830	0,900
12	90	0,850	0,850
13	90	0,870	0,920
14	90	0,880	0,920
15	90	0,920	0,880
16	90	0,880	0,830
17	90	0,820	0,900
18	90	0,900	0,880
19	90	0,900	0,900
20	90	0,870	0,950
21	90	0,930	0,930
22	90	0,870	0,920
23	90	0,920	0,930
24	90	0,870	0,830
25	90	0,830	0,880
Среднее	90	0,868	0,898

Таблица 3

**Точность на наборе данных
«Ирис» при обучающей выборке
60 прецедентов**

Номер испытаний	Обучаю- щая выборка	Точность на тестовых данных	
		ФП типа T	ФП типа S
1	60	0,930	0,880
2	60	0,860	0,910
3	60	0,880	0,910
4	60	0,880	0,870
5	60	0,880	0,870
6	60	0,920	0,940
7	60	0,860	0,860
8	60	0,790	0,820
9	60	0,860	0,870
10	60	0,840	0,960
11	60	0,920	0,880
12	60	0,880	0,890
13	60	0,830	0,890
14	60	0,840	0,900
15	60	0,890	0,900
16	60	0,920	0,720
17	60	0,830	0,870
18	60	0,820	0,880
19	60	0,810	0,890
20	60	0,810	0,900
21	60	0,880	0,810
22	60	0,880	0,890
23	60	0,890	0,890
24	60	0,860	0,830
25	60	0,900	0,910
Среднее	60	0,866	0,878

Таблица 4

**Точность на наборе данных
«Ирис» при обучающей выборке
45 прецедентов**

Номер испытаний	Обучаю- щая выборка	Точность на тестовых данных	
		ФП типа T	ФП типа S
1	45	0,940	0,850
2	45	0,760	0,860
3	45	0,840	0,850
4	45	0,910	0,850
5	45	0,880	0,900
6	45	0,730	0,810
7	45	0,930	0,880
8	45	0,870	0,790
9	45	0,880	0,860
10	45	0,920	0,880
11	45	0,900	0,820
12	45	0,860	0,830
13	45	0,850	0,820
14	45	0,920	0,890
15	45	0,870	0,840
16	45	0,930	0,930
17	45	0,840	0,870
18	45	0,800	0,850
19	45	0,840	0,880
20	45	0,830	0,900
21	45	0,870	0,890
22	45	0,880	0,890
23	45	0,820	0,880
24	45	0,820	0,890
25	45	0,810	0,840
Среднее	45	0,860	0,862

Таблица 5

**Точность на наборе данных
«Ирис» при обучающей выборке
30 прецедентов**

Номер испытаний	Обучаю- щая выборка	Точность на тестовых данных	
		ФП типа T	ФП типа S
1	30	0,880	0,900
2	30	0,830	0,880
3	30	0,690	0,670
4	30	0,830	0,810
5	30	0,680	0,830
6	30	0,790	0,820
7	30	0,750	0,830
8	30	0,810	0,780
9	30	0,760	0,780
10	30	0,900	0,790
11	30	0,780	0,710
12	30	0,800	0,770
13	30	0,800	0,860
14	30	0,800	0,870
15	30	0,830	0,770
16	30	0,810	0,820
17	30	0,840	0,800
18	30	0,750	0,680
19	30	0,660	0,870
20	30	0,750	0,870
21	30	0,910	0,850
22	30	0,770	0,820
23	30	0,820	0,820
24	30	0,670	0,830
25	30	0,810	0,870
Среднее	30	0,788	0,812

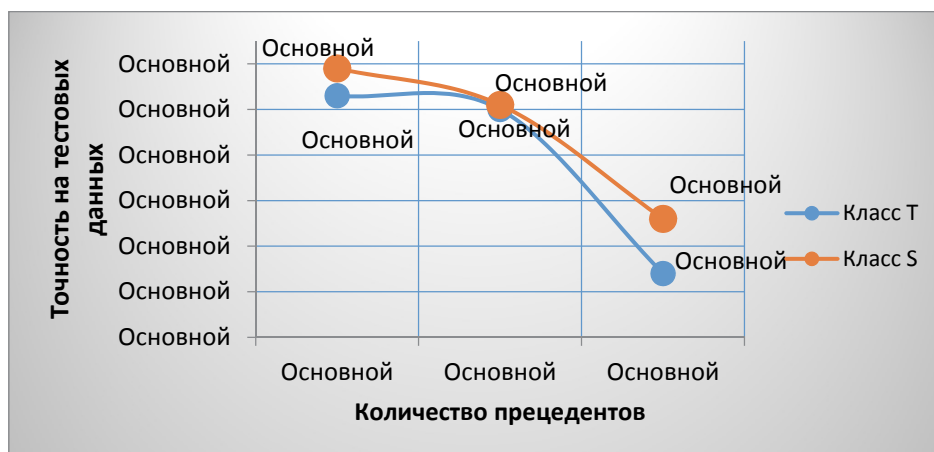


Рис. 6. Точность на наборе данных «Ирис»

Из табл. 2–5 и рис. 6 видно, что при значительном сокращении тестовых данных (до 30 прецедентов из 150) точность поиска остается достаточно высокой, это подтверждает гипотезу о том, что при небольшом количестве исходных данных с помощью предложенного алгоритма можно делать прогнозы с достаточным уровнем точности.

В дальнейшем планируется провести исследования на других наборах данных, увеличить количество значений лингвистических переменных, рассмотреть другие функции принадлежности.

ЗАКЛЮЧЕНИЕ

Модель на основе прецедентов с использованием нечетких правил решает традиционную проблему сопоставления метрик в гибридной прецедентной системе. Каждый прецедент из базы оценивается с помощью нечетких правил, как и насколько он похож на текущую ситуацию. Введение нечетких правил в качестве критериев оценки подобия приносит следующие существенные преимущества.

Во-первых, нечеткие правила «если-то» представляют собой более общую и гибкую платформу для представления знаний о сходстве между прецедентами, чем расстояние в многомерном пространстве. Благодаря полным знаниям мы получаем более точное сходство.

Во-вторых, сущность базы знаний, содержащей нечеткие правила, позволяет приобретать и интегрировать знания из различных источников. Квалифицированные нечеткие правила могут быть получены от одного или нескольких специалистов ПО с помощью подходов по извлечению знаний. Они также могут быть извлечены из существующих данных с использованием машинного обучения и интеллектуального анализа данных. Однако стоит отметить, что нечеткие базы правил, сгенерированные из базы прецедентов, могут быть дополнены специальными правилами, определенными для достижения более полных знаний.

В-третьих, нечеткая база знаний является хорошо понятной благодаря интерпретации отдельных языковых правил. Следовательно, пользователи могут легко понять, как и почему прецедент был выбран для поиска, содер-

жательно подтвердив нечеткие правила, полученные в результате обучения. Это создает возможность динамического участия человека во взаимодействии с системой, основанной на прецедентах, для адаптации поведения системы к конкретным требованиям.

СПИСОК ЛИТЕРАТУРЫ

1. Карпов Л.Е., Юдин В.Н. Методы добычи данных при построении локальной метрики в системах вывода по прецедентам. – М.: Питер, 2007.
2. Bonzano A., Cunningham P., Smith B. Using introspective learning to improve retrieval in CBR: a case study in air traffic control // Case-based reasoning: research and development: Second International Conference on Case-Based Reasoning, ICCBR-97, Providence, RI, USA, 25–27 July 1997: Proceedings. – Berlin; New York: Springer, 1997. – P. 291–302.
3. Cercone N., An A., Chan C. Rule-induction and case-based reasoning: hybrid architectures appear advantageous // IEEE Transactions on Knowledge and Data Engineering. – 1999. – Vol. 11. – P. 166–174.
4. Coyle L., Cunningham P. Improving recommendation ranking by learning personal feature weights // Advances in Case-Based Reasoning: 7th European Conference, ECCBR 2004, Madrid, Spain, 30 August – 2 September 2004: Proceedings. – Berlin; New York: Springer, 2004. – P. 560–572.
5. Jarmulak J., Craw S., Rowe R. Genetic algorithms to optimize CBR retrieval // Advances in Case-Based Reasoning: 5th European Workshop, EWCBR 2000, Trento, Italy, 6–9 September 2000: Proceedings. – Berlin; Heidelberg: Springer, 2000. – P. 136–147.
6. Xiong N., Funk P. Building similarity metrics reflecting utility in case-based reasoning // Intelligent & Fuzzy Systems. – 2006. – Vol. 17. – P. 407–416.
7. Xiong N., Funk P. Combined feature selection and similarity modeling in case-based reasoning using hierarchical memetic algorithm // Proceedings IEEE World Congress on Computational Intelligence, 2010. – Barcelona, Spain, 2010. – P. 1537–1542.
8. Нечеткие множества и теория возможностей: последние достижения / под ред. Р.Р. Ягера; пер. с англ. В.Б. Кузьмина; под ред. С.И. Травкина. – М.: Радио и связь, 1986. – 408 с.
9. Dubois D., Prade H. Fuzzy sets and systems theory and applications. – London: Academic Press, 1980. – 411 p.
10. Marques V., Farinha J.T., Brito A. Case-based reasoning and fuzzy logic in fault diagnosis // WSEAS Transactions on Computers. – 2009. – Vol. 8. – P. 1408–1417.
11. Xiong N., Funk P. Construction of fuzzy knowledge bases incorporating feature selection // Soft Computing. – 2006. – Vol. 10. – P. 796–804.
12. Xiong N. Fuzzy rule-based similarity model enables learning from small case bases // Applied Soft Computing. – 2013. – Vol. 13. – P. 2057–2064.
13. Xiong N. Learning fuzzy rules for similarity assessment in case-based reasoning // Expert Systems and Applications. – 2011. – Vol. 38. – P. 10780–10786.
14. Нечипоренко О.А. Использование технологии Case-Based Reasoning в проектировании программных систем // Перспективные информационные технологии и информационные среды. – 2002. – № 3. – С. 27–32.
15. Aamodt A., Plaza E. Case-Based Reasoning: foundational issues, methodological variations, and system approaches // AI Communications. – 1994. – Vol. 7, N 1. – P. 39–59.
16. Nilsson M., Sollenborn M. Advancements and trends in medical case-based reasoning: an overview of systems and system development // Proceedings of the 17th International FLAIRS Conference. – Menlo Park, California, 2004. – P. 178–183.
17. Штовба С.Д. Введение в теорию нечетких множеств и нечеткую логику [Электронный ресурс]. – URL: <http://matlab.exponenta.ru/fuzzylogic/book1> (дата обращения: 14.04.2016).
18. Kablan A. Adaptive neuro fuzzy inference systems for high frequency financial trading and forecasting // Third International Conference on Advanced Engineering Computing and Applications in Sciences. – Los Alamitos, California: IEEE Computer Society, 2009. – P. 105–110.
19. Massie S., Craw S., Wiratunga N. When similar problems don't have similar solutions // Case-Based Reasoning Research And Development: 7th International Conference on Case-Based Reasoning, ICCBR 2007, Belfast, Northern Ireland, 13–16 August 2007: proceedings. – Berlin: Springer, 2007. – P. 92–106.

20. Рутковская Д., Пилиньский М., Рутковский Л. Нейронные сети, генетические алгоритмы и нечеткие системы / пер. с пол. И.Д. Рудинского. – 2-е изд., стер. – М.: Горячая линия–Телеком, 2013. – 384 с.

Авдеенко Татьяна Владимировна, доктор технических наук, профессор, заведующий кафедрой Экономической информатики Новосибирского государственного технического университета. Основное направление научных исследований – математическое моделирование, интеллектуальные технологии, анализ данных. Имеет около 140 научных публикаций. E-mail: tavdeenko@mail.ru

Макарова Екатерина Сергеевна, магистр прикладной информатики, аспирантка кафедры Экономической информатики Новосибирского государственного технического университета. Основное направление научных исследований – управление в социально-экономических системах, прецедентный подход. Имеет более 15 публикаций E-mail: KATMC@yandex.ru

Method for determining the relevance of cases based on fuzzy linguistic rules*

T.V. AVDEENKO¹, Y.S. MAKAROVA²

¹ *Novosibirsk State Technical University, 20, K. Marx Prospekt, Novosibirsk, 630073, Russian Federation D.Sc. (Eng.), professor. E-mail: tavdeenko@mail.ru*

² *Novosibirsk State Technical University, 20, K. Marx Prospekt, Novosibirsk, 630073, Russian Federation, postgraduate student. E-mail: KATMC@yandex.ru*

The Case-Based Reasoning approach (CBR) is a useful tool of knowledge representation in socio-economic systems because it uses reasoning by analogy inherent in the human intellect. The traditional approach to the assessment of the similarity of precedents is to calculate the weighted metrics - distance in a multidimensional space. In the article, instead of precedent similarity we consider their relevance based on fuzzy logic. Fuzzy linguistic rules are used as a tool for representing knowledge of the relevance of precedents. We consider the procedure for setting up fuzzy rules based on training the precedent samples. We investigate both triangular and quadratic membership functions. The results obtained showed an acceptable accuracy of the precedent classification even for small training samples. Moreover, smoother (quadratic) membership functions showed on average greater prediction accuracy. The introduction of fuzzy rules as criteria for evaluating the similarity gives significant advantages to the case model: fuzzy rules "if-then" are a more general and flexible platform for the representation of knowledge about the similarities between the cases than distance in a multidimensional space. The essence of the knowledge base containing fuzzy rules allows us to acquire and integrate knowledge from different sources. In addition, the fuzzy knowledge base is well understood due to the interpretation of certain language rules, so users can easily understand how and why the precedent was selected for the search substantially confirming the fuzzy rules obtained as a result of learning. This creates a possibility of dynamic human intervention in the interaction with a case-based system in order to adapt the system's behavior to specific requirements.

Keywords: Case-based reasoning approach, knowledge base, fuzzy logic, fuzzy set, linguistic variable, fuzzy rules, reasoning by analogy, membership function

DOI: 10.17212/1814-1196-2016-1-17-34

* Received 14 November 2015.

REFERENCES

1. Karpov L.E., Yudin V.N. *Metody dobychi dannykh pri postroenii lokal'noi metriki v sistemakh vyvoda po pretsedentam* [Methods of data mining in the construction of a local metric in output by precedents systems]. Moscow, Piter Publ., 2007.
2. Bonzano A., Cunningham P., Smith B. Using introspective learning to improve retrieval in CBR: a case study in air traffic control. *Case-based reasoning: research and development: Second International Conference on Case-Based Reasoning, ICCBR-97: Proceedings*, Providence, RI, USA, 25–27 July 1997, pp. 291–302.
3. Cercone N., An A., Chan C. Rule-induction and case-based reasoning: hybrid architectures appear advantageous. *IEEE Transactions on Knowledge and Data Engineering*, 1999, vol. 11, pp. 166–174.
4. Coyle L., Cunningham P. Improving recommendation ranking by learning personal feature weights. *Advances in Case-Based Reasoning: 7th European Conference, ECCBR 2004: Proceedings*, Madrid, Spain, 30 August – 2 September 2004, pp. 560–572.
5. Jarmulak J., Craw S., Rowe R. Genetic algorithms to optimize CBR retrieval. *Advances in Case-Based Reasoning: 5th European Workshop, EWCBR 2000: Proceedings*, Trento, Italy, 6–9 September 2000, pp. 136–147.
6. Xiong N., Funk P. Building similarity metrics reflecting utility in case-based reasoning. *Intelligent & Fuzzy Systems*, 2006, vol. 17, pp. 407–416.
7. Xiong N., Funk P. Combined feature selection and similarity modeling in case-based reasoning using hierarchical memetic algorithm. *Proceedings IEEE World Congress on Computational Intelligence, 2010*, Barcelona, Spain, 2010, pp. 1537–1542.
8. *Fuzzy set and possibility theory: recent developments*. Ed. by R.R. Yager. New York, Pergamon, 1982. 633 p. (Russ. ed.: *Nechetkie mnozhestva i teoriya vozmozhnostei: poslednie dostizheniya*. Translated from English V.B. Kuz'min. Moscow, Radio i svyaz' Publ., 1986. 408 p.).
9. Dubois D., Prade H. *Fuzzy sets and systems theory and applications*. London, Academic Press, 1980. 411 p.
10. Marques V., Farinha J.T., Brito A. Case-based reasoning and fuzzy logic in fault diagnosis. *WSEAS Transactions on Computers*, 2009, vol. 8, pp. 1408–1417.
11. Xiong N., Funk P. Construction of fuzzy knowledge bases incorporating feature selection. *Soft Computing*, 2006, vol. 10, pp. 796–804.
12. Xiong N. Fuzzy rule-based similarity model enables learning from small case bases. *Applied Soft Computing*, 2013, vol. 13, pp. 2057–2064.
13. Xiong N. Learning fuzzy rules for similarity assessment in case-based reasoning. *Expert Systems and Applications*, 2011, vol. 38, pp. 10780–10786.
14. Nechiporenko O.A. Ispol'zovanie tekhnologii Case-Based Reasoning v proektirovanii programnykh sistem [Usage of technology of Case-Based Reasoning in the design of software systems]. *Perspektivnye informatsionnye tekhnologii i informatsionnye sredy – Advanced information technologies and the information environment*, 2002, no. 3, pp. 27–32.
15. Aamodt A., Plaza E. Case-Based Reasoning: foundational issues, methodological variations, and system approaches. *AI Communications*, 1994, vol. 7, no. 1, pp. 39–59.
16. Nilsson M., Sollenborn M. Advancements and trends in medical case-based reasoning: an overview of systems and system development. *Proceedings of the 17th International FLAIRS Conference*, Menlo Park, California, 2004, pp. 178–183.
17. Shtovba S.D. *Vvedenie v teoriyu nechetkikh mnozhestv i nechetkuyu logiku* [Introduction to the theory of fuzzy sets and fuzzy logic]. Available at: <http://matlab.exponenta.ru/fuzzylogic/book1> (accessed 14.04.2016)
18. Kablan A. Adaptive neuro fuzzy inference systems for high frequency financial trading and forecasting. *Third International Conference on Advanced Engineering Computing and Applications in Sciences*. Los Alamitos, California, IEEE Computer Society, 2009, pp. 105–110.
19. Massie S., Craw S., Wiratunga N. When similar problems don't have similar solutions. *Case-Based Reasoning Research And Development: 7th International Conference on Case-Based Reasoning, ICCBR 2007: Proceedings*, Belfast, Northern Ireland, 13–16 August 2007, pp. 92–106.
20. Rutkowska D., Piliński M., Rutkowski L. *Sieci neuronowe, algorytmy genetyczne i systemy rozmyte*. Warszawa, Łódź, Wydawnictwo Naukowe PWN, 2004 [Neural networks, genetic algorithms and fuzzy systems] (Russ. ed.: Rutkovskaya D., Piliński M., Rutkovskii L. *Neironnye seti, geneticheskie algoritmy i nechetkie sistemy*. Translated from Polish I.D. Rudinskii. 2nd ed. Moscow, Goryachaya liniya–Telekom Publ., 2013. 384 p.).