УДК 004.891.3

Н. И. Омирова, А. В. Тишков
Первый Санкт-Петербургский государственный медицинский университет им. академика И. П. Павлова

Конструирование атрибутов с применением нечеткого вывода при диагностике уролитиаза

Рассматривается задача добавления нового атрибута, построенного с помощью нечеткого вывода, при диагностике уролитиаза методом обучения с учителем. Приведены примеры использования нечеткой логики на порядковых и количественных данных. Проведен анализ уровня точности классификатора согласно кросс-валидации до и после добавления трех дополнительных атрибутов, вычисляемых с использованием нечеткого вывода. После интерпретации полученных результатов обнаружено, что добавление новых атрибутов увеличило точность классификации при помощи дерева решений на 5 %: с 79 до 84 %. Наибольшее повышение точности достигнуто с использованием нечеткой логики на порядковых данных.

Задача классификации, деревья решений, нечеткая логика

Метод классификации, или обучения с учителем, успешно применяется в задачах медицинской диагностики. «Учителем» служит набор атрибутов (медицинских показателей) объекта (пациента), которому уже сопоставлен класс (диагноз) [1], [2]. Нередко к этим атрибутам полезно добавить новые, искусственно полученные на основании существующих. Существует ряд известных методов создания таких атрибутов: сложение существующих, возведение в квадрат и др. В настоящей статье для этой цели предлагается использовать нечеткий вывод.

Множество атрибутов, на которых происходит обучение, неоднородно как по своей структуре, так и по связи с диагнозом. Выделяют номинальные, порядковые и количественные атрибуты [3]. В статье рассматривается задача классификации, в которой каждый объект характеризуется 17 атрибутами, из которых 5 порядковых и 12 количественных.

На основе связи каждого атрибута с классом (диагнозом) множество атрибутов можно упорядочить, пользуясь алгоритмами информативности или методом проверки статистических гипотез. По принципу слабой или сильной связи с диагнозом мы разобьем 12 количественных атрибутов на два набора.

Для выделения новых атрибутов используются три группы: 5 порядковых и две группы по 6 количественных (со слабой и с сильной связью с диагнозом). Данный метод относится к известным способам улучшения точности классификации. Для выделения новых атрибутов будет использован нечеткий вывод.

Описание исходных данных. Исследуемая выборка состояла из двух групп. Группа больных уролитиазом составила 45 человек (22 женщины, 23 мужчины) в возрасте 25–60 лет. Группа контроля была сформирована из 35 практически здоровых добровольцев. Группа контроля сопоставима с группой больных по полу и возрасту.

В качестве атрибутов многомерной информации выступали 17 показателей общего и биохимического анализов крови и мочи, наиболее применимые в практике [4], [5].

Характеристика исходных данных представлена в табл. 1.

Исходные данные представлены в виде как порядковых (холодовая проба, цвет, прозрачность, бактерии, соли), так и количественных (остальные 12 показателей) атрибутов.

Для анализа исходных данных в исследовании применялся статистический пакет Past, для определения информативности «весов» атрибутов – Rapid Miner, для создания нечеткой модели и получения нечеткого выхода – пакет Fuzzy Logic Toolbox Matlab. Для формулировки нечетких правил использовались деревья решений, построенные в Rapid Miner.

Таблица 1

№ п/п	Атрибут	Тип данных	Среднее ± стандартное отклонение для количественных данных. Для номинальных – численность наиболее и наименее объемных групп
1	Диагноз	binominal	mode = Уролитиаз (45), least = Группа контроля (35)
2	Холодовая проба	nominal	mode = 0 (40), least = 1 (40)
3	Соли	nominal	mode = 0 (45), least = 3 (4)
4	Цвет	nominal	mode = 1 (56), least = 3 (1)
5	Прозрачность	nominal	mode = 1 (42), least = 3 (3)
6	Бактерии	nominal	mode = 0 (50), least = 5 (5)
7	Цитраты	real	avg = 1.073 ± 0.952
8	Оксалаты	real	avg = 0.118 ± 0.061
9	Калий, моча	real	avg = 35.347 ± 19.383
10	Альбумин, кровь	real	avg = 38.976 ± 5.395
11	С-реактивный белок	real	avg = 10.147 ± 21.749
12	Белок	real	avg = 0.122 ± 0.265
13	Относительная плотность	integer	avg = 1016.138 ± 5.357
14	Ph	real	avg = 5.866 ± 0.978
15	Фосфор неорганический, кровь	real	avg = 1.068 ± 0.190
16	Осмолярность мочи	real	avg = 649.592 ± 212.241
17	Кальций ионизированный	real	avg = 1.246 ± 0.388
18	Экскреция титруемых аминокислот	real	avg = 38.181 ± 23.675

Таблица 2

Атрибуты	Шапиро–Уилка p (normal) больные	Жарка–Бера p (normal) больные	Согласованность с нормальным распределением	Шапиро–Уилка p (normal) группа контроля	Жарка–Бера p (normal) группа контроля	Согласованность с нормальным распределением
Цитраты	<10–7	<10–21	Нет	<10–4	<10–7	Нет
Оксалаты	0.0642	0.5957	Да	0.8413	0.7404	Да
Калий, моча	0.0512	0.247	Да	0.5536	0.6667	Да
Альбумин, кровь	0.1253	0.6309	Да	0.0002	0.0090	Да
СРБ	<10–10	<10–82	Нет	<10–8	<10–38	Нет
Белок	<10–9	<10–35	Нет	<10–5	0.0386	Нет
Относительная плотность	0.3219	0.502	Да	0.0077	0.1276	Нет
Ph	<10–6	0.0686	Нет	0.0693	0.4463	Да
Фосфор неорганический, кровь	0.9528	0.7524	Да	0.0119	0.9557	Нет
Осмолярность мочи	0.7201	0.6016	Да	0.3055	0.6134	Да
Кальций ионизированный	0.4704	0.586	Да	<10–9	<10–61	Нет
Экскреция титруемых аминокислот	0.0583	0.4699	Да	0.0005	0.0153	Нет

Таблица 3

Атрибуты	Р-значение	Критерий
Цитраты	<10–6	Манна–Уитни
Оксалаты	<10–8	Стьюдента
Альбумин, кровь	0.0001	
Калий, моча	<10–5	
СРБ	<10–3	Манна–Уитни
Белок	0.004	
Ph	0.01	
Экскреция титруемых аминокислот	0.11	
Фосфор неорганический, кровь	0.20	
Относительная плотность	0.34	
Кальций ионизированный	0.90	
Осмолярность мочи	0.99	Стьюдента

На первом этапе исследования все количественные данные были проверены на нормальность распределения. Проверка проводилась с использованием критериев Шапиро–Уилка и Жарка–Бера. Результаты проверки согласованности распределения с нормальным представлены в табл. 2.

Если оба выборочных распределения согласованы с нормальным, то для поиска различий использовался критерий Стьюдента с неравными дисперсиями. Если одна из выборок не согласована, то применялся критерий Манна–Уитни. Исходные 12 количественных атрибутов можно разделить на шесть более и шесть менее информативных в соответствии с полученным Р-значением (табл. 3).

На мощность статистических критериев может повлиять относительно небольшой размер выборки, поэтому дополнительно было проведено упорядочение атрибутов с использованием весов, основанных на измерении информативности по критерию gain ratio (табл. 4).

Таблица 4

Атрибуты	Вес
Ph	0
Относительная плотность	0.016
Осмолярность мочи	0.044
Экскреция титруемых аминокислот	0.056
Кальций ионизированный	0.067
Фосфор неорганический, кровь	0.123
Белок	0.155
СРБ	0.245
Калий, моча	0.251
Оксалаты	0.411
Альбумин, кровь	0.951
Цитраты	1

Оба метода одинаково разделили 12 атрибутов на две группы по шесть с незначительными перестановками внутри групп. Первая группа соответствует более сильной связи с диагнозом, чем вторая.

Нечеткая логика. Правила и графики функции принадлежности. Нечеткие рассуждения основаны на понятии лингвистической переменной, которое будет использоваться для всех атрибутов [6], [7]. Применение нечетких правил к полученным лингвистическим переменным даст нечеткую оценку наличия заболевания для каждого пациента, которая затем дефаззифицируется в выходную переменную с целочисленной областью значений на отрезке от 0 до 10 [8]. В результате генерируется новый атрибут, добавляемый к исходным.

Основной набор правил выбирается из деревьев решений при помощи эксперта и фаззифицируется. Деревья решений строятся при помощи известных алгоритмов классификации [9]. Классов, как и в основной задаче диагностики уролитиаза, два: «группа контроля», «уролитиаз».

Создано три нечетких модели. В нечеткой модели на основе порядковых атрибутов предполагается использовать 5 входных и 1 выходную переменную (табл. 5). В качестве входных переменных используются только порядковые параметры

Таблица 5

Переменная	Наименование переменной	Терм-множество	
		Множество	Символический вид
Входная	Цвет	T1 = {«соломенная», «слабо желтая», «желтая», «бурая»}	T1 = {S, SY, Y, B}
	Прозрачность	T2 = {«прозрачная», «слабо мутная», «мутная»}	T2 = {P, SM, M}
	Бактерии	T3 = {«нет», «бактерии», «дрожжи», «кандида»}	T3 = {A, B, C, D}
	Холодовая проба	T4 = {«осадок присутствует», «осадок отсутствует»,}	T4 = {Y, N}
	Соли	T5 = {«соли присутствуют», «соли присутствуют»,}	T5 = {Y, N}
Выходная	Возможность наличия заболевания	T6 = {«очень низкая», «низкая», «средняя», «высокая», «очень высокая»}	T6 = {ON, N, S, V, OV}

Таблица 6

№	Холодовая проба	Соли	Цвет	Прозрачность	Бактерии	Возможность наличия заболевания
1	Y	Y	B	M	D	OV
2	Y	Y	S	P	C	V
3	N	Y	B	M	B	S
4	N	Y	SY	SM	A	N
5	N	N	S	P	A	ON

Таблица 7

Переменная	Наименование переменной	Терм-множество	
		Множество	Символический вид
Входная	Цитраты	T1 = {«Ниже нормы», «норма»}	T1 = {NN, N}
	Оксалаты	T2 = {«Ниже нормы», «норма»}	T2 = {NN, N}
	Калий, моча	T3 = {«Ниже нормы», «норма»}	T3 = {NN, N}
	Альбумин, кровь	T4 = {«Ниже нормы», «норма»}	T4 = {NN, N}
	СРБ	T5 = {«Норма», «выше нормы»}	T5 = {N, VN}
	Белок	T6 = {«Норма», «выше нормы»}	T6 = {N, VN}
Выходная	Возможность наличия заболевания	T7 = {«Очень низкая», «низкая», «средняя», «высокая», «очень высокая»}	T7 = {ON, N, S, V, OV}

Таблица 8

№	Цитраты	Оксалаты	Калий, моча	Альбумин, кровь	СРБ	Белок	Возможность наличия заболевания
1	NN	NN	NN	NN	VN	VN	OV
2	NN	N	NN	NN	VN	N	V
3	NN	N	N	NN	VN	N	S
4	N	NN	N	N	N	VN	N
5	N	NN	N	N	N	N	ON

Таблица 9

Переменная	Наименование переменной	Терм-множество	
		Множество	Символический вид
Входная	Относительная плотность	T1 = {«Ниже нормы», «норма»}	T1 = {NN, N}
	Ph	T2 = {«Ниже нормы», «норма»}	T2 = {NN, N}
	Фосфор неорганический, кровь	T3 = {«Ниже нормы», «норма»}	T3 = {NN, N}
	Осмолярность мочи	T4 = {«Ниже нормы», «норма»}	T4 = {NN, N}
	Кальций ионизированный	T5 = {«Ниже нормы», «норма»}	T5 = {NN, N}
	Экскреция титруемых АМК	T6 = {«Ниже нормы», «норма»}	T5 = {NN, N}
Выходная	Возможность наличия заболевания	T7 = {«очень низкая», «низкая», «средняя», «высокая», «очень высокая»}	T7 = {ON, N, S, V, OV}

многомерных медицинских данных. В качестве выходной переменной используется возможность обнаружения у пациента наличия мочекаменной болезни.

После определения содержательной постановки задачи была построена ее нечеткая модель в форме соответствующей системы нечеткого вывода [10], [11]. При построении нечеткой модели оценки возможности обнаружения у пациента наличия уролитиаза была использована шкала в баллах в интервале от 0 до 10.

Следующим этапом построения модели становится построение базы нечетких правил [12]. Для этой цели использовались четкие правила, сгенерированные на основе деревьев решений. Стандартные деревья решений построены на всех атрибутах. На порядковых исходных данных точность кросс-валидации ниже, чем на количественных [13].

Эксперт выбирал правила из деревьев решений и на их основе формулировал собственные нечеткие правила. Всего экспертом была сформулирована 51 нечеткая продукция. В табл. 6 представлены примеры правил.

Во второй нечеткой модели на основе 6 «сильных», «информативных» количественных атрибутов предполагается использовать 6 входных переменных и 1 выходную переменную (табл. 7).

Таблица 10

№	Относительная плотность	Ph	Фосфор неорганический, кровь	Осмолярность мочи	Кальций ионизированный	Экскреция титруемых АМК	Возможность наличия заболевания
1	NN	NN	NN	NN	NN	NN	OV
2	N	N	NN	NN	NN	NN	V
3	N	N	NN	N	NN	NN	S
4	NN	NN	N	NN	N	N	N
5	N	NN	N	N	N	N	ON

Эксперт выбирал правила из деревьев решений и на их основе формулировал собственные нечеткие правила. Всего экспертом была сформулирована 41 нечеткая продукция (табл. 8).

В третьей нечеткой модели на основе 6 «слабых» «не информативных» количественных атрибутов предполагается использовать 6 входных переменных и 1 выходную переменную (табл. 9).

Эксперт выбирал правила из деревьев решений и на их основе формулировал собственные нечеткие правила. Всего экспертом было сформулировано 49 нечетких продукций (табл. 10).

Метод Мамдани использован в качестве схемы нечеткого вывода, методом активации служит MIN. Во всех шагах в качестве логической связки для подусловий применяется только нечеткая конъюнкция (операция «И»), поэтому в качестве метода агрегирования использовалась операция min-конъюнкции [14], [15]. Для аккумуляции заключений правил использовался метод max-дизъюнкции. В качестве метода дефаззификации планируется использовать метод центра тяжести.

Нечеткий выход и анализ результатов. Проведенные вычисления позволили получить три дополнительных показателя: нечеткий выход, основанный на порядковых показателях, нечеткий выход, основанный на информативных относительно диагноза количественных показателях, и нечеткий выход, основанный на малоинформативных количественных показателях (табл. 11).

Значения в табл. 11 являются оценкой наличия уролитиаза по шкале от 0 до 10. При этом большее число соответствует большей возможности наличия заболевания.

Проверка согласованности с нормальным распределением представлена в табл. 12. Согласно таблице все распределения не согласовываются с нормальным, следовательно, для доказательства различий использован непараметрический критерий Манна–Уитни (табл. 13).

Таблица 11

Группа	Нечеткий выход 1	Нечеткий выход 2	Нечеткий выход 3
Контроля	5.00	5.00	5.00
	1.13	6.75	5.00
	1.38	5.00	1.28
	1.09	5.00	5.00
	6.67	6.75	1.28
	1.13	5.00	5.00
	1.09	5.00	5.00
	1.07	8.58	5.00
	1.09	6.75	5.00
	1.13	6.76	1.42
	6.67	5.00	5.00
	1.22	6.75	5.00
	1.13	5.00	1.42
	1.13	5.00	5.00
	5.00	6.75	5.00
	1.13	6.75	5.00
	1.13	5.00	1.42
	1.13	6.75	5.00
	1.07	5.00	5.00
	1.22	6.75	5.00
	1.22	6.75	5.00
	5.00	5.00	5.00
	1.13	6.76	5.00
	1.09	5.00	1.42
	5.00	6.75	3.23
	5.00	6.75	5.00
	1.13	6.75	5.00
	1.09	6.75	5.00
	1.22	6.75	5.00
	1.13	6.76	5.00
	1.09	5.00	5.00
	1.13	8.72	5.00
	1.13	5.00	5.00
	1.22	5.00	5.00
	5.00	5.00	1.28
Уролитиаз	1.38	8.58	5.00
	2.60	8.58	3.23
	5.00	5.00	3.23
	1.13	8.84	5.00
	5.00	8.72	5.00
	1.13	8.58	1.42
	6.76	5.00	5.00
	1.13	5.00	5.00
	2.67	6.75	1.42
	1.22	8.58	5.00
	8.91	8.58	5.00
	8.97	8.58	1.42

Окончание табл. 11

Группа	Нечеткий выход 1	Нечеткий выход 2	Нечеткий выход 3
Уролитиаз	1.13	5.00	5.00
	8.91	8.58	5.00
	5.00	8.58	5.00
	1.13	8.84	3.23
	1.09	5.00	1.42
	2.67	8.58	5.00
	2.67	8.84	5.00
	5.00	5.00	5.00
	6.76	1.42	5.00
	2.67	5.00	5.00
	2.68	6.75	5.00
	6.76	5.00	5.00
	5.00	6.75	5.00
	6.76	5.00	5.00
	2.68	6.75	1.42
	1.13	6.75	5.00
	1.22	8.58	1.28
	5.00	5.00	3.20
	1.13	8.72	5.00
	6.76	5.00	5.00
	8.78	8.58	3.23
	8.78	8.58	5.00
	6.76	5.00	5.00
	1.13	8.84	5.00
	1.38	6.75	5.00
	5.00	5.00	1.42
	1.13	8.58	1.42
	1.07	8.72	5.00
	5.00	5.00	5.00
	1.13	5.00	5.00
	1.13	5.00	3.23
	6.76	8.72	5.00
	6.76	5.00	5.00

Таблица 12

Группа	Нечеткий выход № 1	Нечеткий выход № 2	Нечеткий выход № 3
Уролитиаз	<10–8	<10–5	<10–8
Контроля	<10–4	<10–5	<10–8

Таблица 13

Группа	Манна–Уитни, p
Нечеткий выход № 1	0.0003587
Нечеткий выход № 2	0.04237
Нечеткий выход № 3	0.6254

В табл. 13 первые два из трех нечетких выхода статистически значимо разделяют выборку по диагнозам.

Таким образом, с использованием нечеткой логики было выделено три новых признака. В табл. 14 представлена точность классификаторов (деревьев решений) согласно кроссвалидации с использованием и без использования этих признаков.

Таблица 14

Исходные данные	Точность классификатора, %
Данные без добавления дополнительных столбцов по нечетким выходам	78.75 ± 12.56
Данные с добавлением 1 столбца нечеткого выхода по порядковым данным	83.75 ± 13.75
Данные с добавлением 1 столбца нечеткого выхода по первому набору количественных данных	80.00 ± 11.46
Данные с добавлением 1 столбца нечеткого выхода по второму набору количественных данных	78.75 ± 12.56
Данные с добавлением 2 столбцов нечеткого выхода по порядковым и первому набору количественных данных	83.75 ± 13.75
Данные с добавлением 2 столбцов нечеткого выхода по порядковым и второму набору количественных данных	83.75 ± 13.75
Данные с добавлением 2 столбцов нечеткого выхода по первому и второму набору количественных данных	80.00 ± 11.46
Данные с добавлением 3 столбцов нечеткого выхода по порядковым данным, первому и второму набору количественных данных	83.75 ± 13.75

Исходная выборка была увеличена на 3 дополнительных атрибута, полученных с помощью нечеткого вывода. Проверка кроссвалидацией показывает, что добавление новых атрибутов увеличивает точность классификации с 78.75 ± 12.56 % до 83.75 ± 13.75 %. Причем также удалось выявить, что максимальное увеличение точности достигается с использованием нечеткой логики, на порядковых данных. Количественные показатели разделились на две группы: «сильные» – информативные, и «слабые» – неинформативные.

Применение нечеткой логики на неинформативных количественных атрибутах не приводит к увеличению точности классификатора согласно кроссвалидации, в то время как использование этого математического аппарата на информативных количественных атрибутах увеличивает точность на 1.25 %. Связи с нормальностью распределения найти не удалось, есть зависимость от уровня информативности количественных показателей.

СПИСОК ЛИТЕРАТУРЫ

1. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Berlin: Springer Verlag, 2009. 746 p.

2. Mitchell T. Machine Learning. McGraw-Hill Science/Engineering/Math, 1997. 432 p.

3. Колесников А. В. Гибридные интеллектуальные системы: Теория и технология разработки / под ред. А. М. Яшина; СПбГТУ. СПб., 2001. 711 с.

4. Эмануэль В. Л. Пособие для семейного врача по лабораторным технологиям и интерпретации исследования мочи: учеб. пособие. СПб.: Триада, 2007. 128 с.

5. Simerville J. A., Maxted W. C., Pahira J. J. Urinalysis (review) // Amer. Fam. Physician. 2005. Vol. 71, № 6. P. 1153–1162.

6. Леоненков А. В. Нечеткое моделирование в среде Matlab и fuzzyTech. СПб.: БХВ-Петербург, 2003. 736 с.

7. Нурманова Е. В. Архитектура системы нечеткого вывода, пример реализации // Вестн. МГУПИ. М., 2004. 100 с.

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

9. Нурматова Е. В. Подход к решению задачи классификации технических состояний в нечеткой логической системе // Изв. ТулГУ. Техн. науки. 2010. № 1. С. 170–174.

10. Новак В., Перфильева И., Мочкрож И. Математические принципы нечеткой логики = Mathematical Principles of Fuzzy Logic. М.: Физматлит, 2006. 352 с.

11. Рутковский Л. Искусственные нейронные сети. Теория и практика. М.: Горячая линия – Телеком, 2010. 520 с.

12. Усков А. А., Кузьмин А. В. Интеллектуальные технологии управления. Искусственные нейронные сети и нечеткая логика. М.: Горячая Линия – Телеком, 2004. 143 с.

13. Загоруйко Н. Г. Прикладные методы анализа данных и знаний. Новосибирск: ИМ СО РАН, 1999. 270 с.

14. Круглов В. В., Дли М. И., Голунов Р. Ю. Нечеткая логика и искусственные нейронные сети. М.: Физматлит, 2001. 221 с.

15. Журавлев Ю. И., Рязанов В. В., Сенько О. В. Распознавание. Математические методы. Программная система. Практические применения. М.: Фазис, 2006. 159 с.

N. I. Omirova, A. V. Tishkov
Pavlov First Saint Petersburg State Medical University

FEATURE CONSTRUCTION WITH FUZZY DERIVATION IN DIAGNOSTICS OF UROLITHIASIS

Supervized learning as method of urolithiasis diagnostics is applied. Construction of new attributes utilizing fuzzy derivation is considered. Fuzzy derivation is applied to ordinal and real data. Cross-validation accuracy analysis is performed before and after addition of three fuzzy-based attributes. As a result of feature construction, accuracy of the decision tree classifier increased by 5 %, from 79 to 84 %. The maximum increase of accuracy was achieved on ordinal attributes.

Problem of classification, decision trees, fuzzy logic