УДК 629.735.33

МЕТОДЫ КЛАССИФИКАЦИИ В ДИАГНОСТИКЕ УРОЛИТИАЗА С ПРИМЕНЕНИЕМ НЕЧЕТКОЙ ЛОГИКИ ДЛЯ ПРЕДОБРАБОТКИ ДАННЫХ

***Н. И. Эюбова*[1],**
аспирант
Санкт-Петербургский государственный университет аэрокосмического приборостроения

Рассматривается задача построения классификатора для диагностики уролитиаза с предварительной обработкой данных, включающей снижение размерности и применение нечеткого вывода. Классифицирующие правила выбираются из деревьев решений, фаззифицируются и валидируются медиком-экспертом. Полученные результаты показывают, что предлагаемые методы предварительной обработки данных улучшают точность диагностики уролитиаза.

Ключевые слова *— задача классификации, деревья решений, нечеткая логика, медицинская диагностика, уролитиаз.*

Введение

Выявление и адаптация приемлемого подхода для анализа многомерной медицинской информации — непростая задача. Рассматривается диагностика уролитиаза по данным общего и биохимического анализа мочи и крови. Уролитиаз является распространенным заболеванием, им страдают более 38 % пациентов уролитических стационаров. Несмотря на современные эффективные и дорогостоящие методы лечения, очень часто возникает рецидив заболевания. Стандарты уролитической помощи не содержат методов ранней диагностики урологических заболеваний, и настоящее исследование направлено на заполнение этого пробела.

В работе использовались методы классификации, доказавшие свою состоятельность при построении систем поддержки принятия решений в медицине и биологии [1, 2]. Предобработка данных заключалась в снижении размерности методом главных компонент и конструировании атрибута с применением нечеткой логики.

Исходные данные состоят из количественных и порядковых атрибутов (шкалу порядка см. в работе [3]). Множество значений порядкового атрибута не отражает результат подсчетов или измерений, но на нем введен порядок, соответствующий «уровням», «степеням», «стадиям» и т. п.

Формирование области значений порядкового атрибута связано с загрублением соответствующей характеристики исследуемого объекта. Например, атрибут «холодовая проба» имеет два значения: 0 — не выпал осадок, 1 — выпал осадок. В пограничных ситуациях специалист вынужден выбрать одно из этих значений, хотя вербально он бы сформулировал «мутная жидкость», «небольшой осадок» и т. п. Введение лингвистических переменных позволяет описать пограничные состояния, увеличив количество информации по сравнению с порядковой шкалой. В настоящем исследовании подтверждается, что применение лингвистических и нечетких правил повышает качество диагностики уролитиаза.

Материалы и методы

В качестве атрибутов многомерной медицинской информации выступали 18 показателей общего и биохимического анализов крови и мочи, наиболее применимые в медицинской практике [4, 5]: относительная плотность, Ph, фосфор неорганический (0,83—1,48) ммоль/л крови, ос-

[1] Научный руководитель — кандидат физико-математических наук, доцент кафедры бизнес-информатики Санкт-Петербургского государственного университета аэрокосмического приборостроения *А. В. Тишков*.

молярность мочи (500—900), цитрат, оксалат, холодовая проба, калий в моче, альбумин в крови, кальций ионизированный (1,06—1,32 ммоль/л), экскреция тируемых кислот [ммоль/л], мочевина мочи, соли, С-реактивный белок (0,00—7,5 мг/л), цвет, прозрачность, белок [г/л] , бактерии.

Исходные данные представлены в виде как порядковых (холодовая проба, цвет, прозрачность, бактерии), так и количественных (остальные 14 показателей) атрибутов.

В исследовании применялись пять классификаторов: метод опорных векторов, метод *k*-ближайших соседей, наивный байесовский классификатор, нейронные сети и метод деревьев решений. В задаче диагностики рассматривается два класса: здоров и болен уролитиазом. Для предобработки данных использовались метод главных компонент, нечеткие правила и выводы.

Исследуемая выборка состояла из двух групп. Группа больных уролитиазом составила 45 чел. (22 женщины, 23 мужчины) в возрасте от 25 до 60 лет. При обучении классификатора эта группа была сопоставлена с классом «здоровые». Группа контроля была сформирована из 35 практически здоровых добровольцев, сопоставленная с классом «больные». Группа контроля сопоставима с группой больных по полу и возрасту.

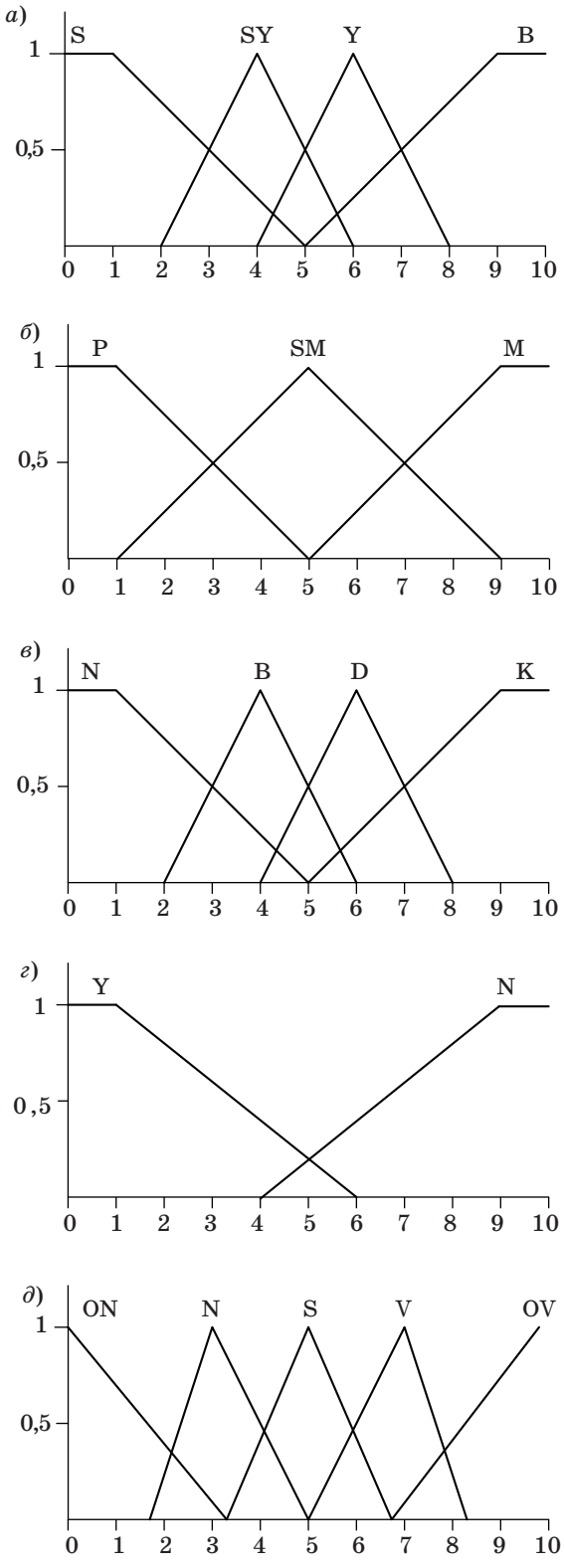
Классификаторы были построены с помощью программного обеспечения Rapid Miner. Для вычислений методом главных компонент использовалась программа MatLab (версия 5). Разработка нечетких правил осуществлялась в пакете Fuzzy Logic Toolbox.

Нечеткие правила

Нечеткие рассуждения основаны на понятии лингвистической переменной, которое будет использоваться для всех порядковых атрибутов. 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. В результате генерируется новый атрибут, добавляемый к исходным.

Основной набор правил выбирается из деревьев решений при помощи эксперта и фаззифицируется. Деревья решений строятся на порядковых атрибутах при помощи известных алгоритмов классификации. Классов, как и в основной задаче диагностики уролитиаза, два: «здоров», «болен».

В нечеткой модели предполагается использовать 4 входные переменные и одну выходную переменную (рис. 1, *а*—*д*, табл. 1). В качестве



■ ***Рис. 1.*** Функции принадлежности термов входных переменных «Цвет» (*а*), «Прозрачность» (*б*), «Бактерии» (*в*), «Холодовая проба» (*г*) и выходной переменной «Вероятность наличия заболевания» (*д*)

■ ***Таблица 1***. Терм-множества входных и выходной переменных

Наименование переменной	Терм-множество	
	Множество	Символический вид
Входные переменные		
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}
Выходная переменная		
Вероятность наличия заболевания	T5={«очень низкая», «низкая», «средняя», «высокая», «очень высокая»}	T5={NB, NS, Z, PS, PB}

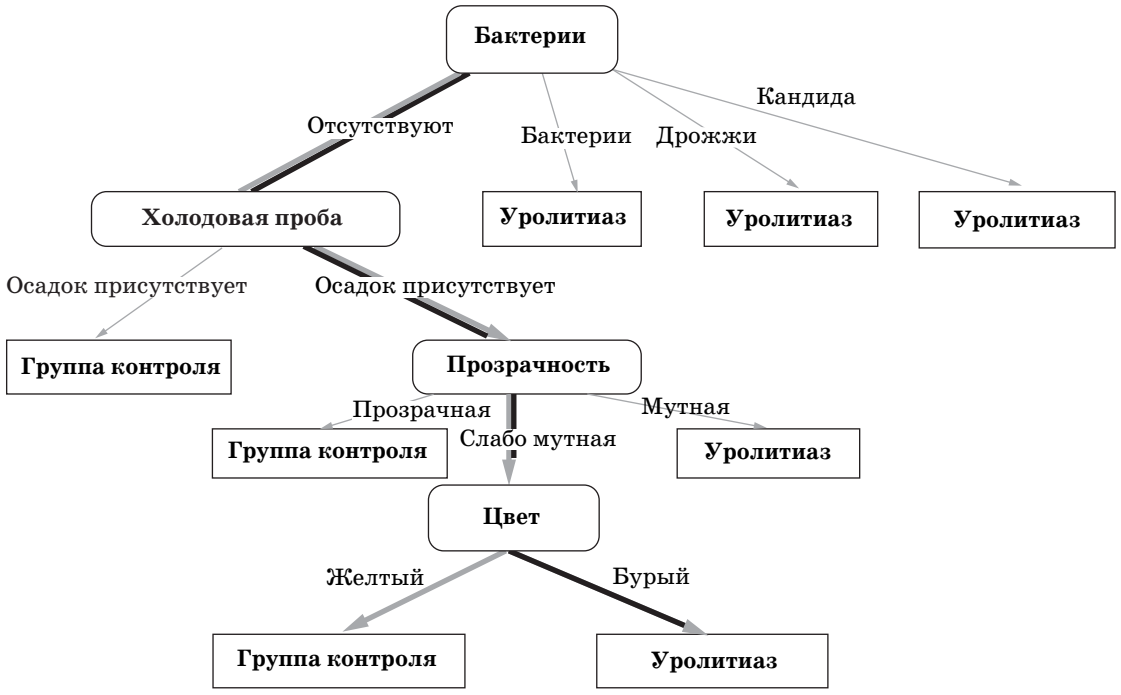
входных переменных используются только качественные параметры многомерных медицинских данных. В качестве выходной переменной используется вероятность обнаружения у пациента наличия мочекаменной болезни.

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

Следующим этапом построения модели является построение базы нечетких правил. Для этой цели использовались четкие правила, сгенерированные на основе деревьев решений. Стандартные деревья решений построены на всех атрибутах и только на порядковых имеют небольшую точность согласно кросс-валидации. Поэтому эксперту было предоставлено дерево решений, построенное на порядковых атрибутах (рис. 2), и набор деревьев решений, полученных согласно процедуре классификации «случайный лес».

Эксперт выбирал правила из деревьев решений и на их основе формулировал собственные нечеткие правила. Всего экспертом была сформулирована 71 нечеткая продукция (табл. 2). На рис. 2 жирными стрелками выделены два правила, которые эксперт использовал явным образом в нечеткой системе (правила 16, 26).

Метод Мамдани использован в качестве схемы нечеткого вывода, методом активации является 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. Для аккумуляции заключений



■ ***Рис. 2***. Дерево решений

■ ***Таблица 2.*** Правила нечетких продукций для рассматриваемой системы нечеткого вывода

№	Цвет	Прозрачность	Бактерии	Холодовая проба	Вероятность наличия заболевания	№	Цвет	Прозрачность	Бактерии	Холодовая проба	Вероятность наличия заболевания
1	–	–	–	N	ON	37	Y	M	K	Y	OV
2	–	–	N	N	ON	38	SY	M	K	Y	OV
3	–	P	N	N	ON	39	S	M	K	Y	OV
4	S	P	N	N	ON	40	B	SM	K	Y	OV
5	S	SM	N	N	ON	41	Y	SM	K	Y	OV
6	SY	P	N	N	ON	42	SY	SM	K	Y	OV
7	SY	SM	N	N	N	43	S	SM	K	Y	OV
8	Y	SM	N	N	N	44	B	P	K	Y	OV
9	B	SM	N	N	N	45	Y	P	K	Y	OV
10	SY	M	N	N	N	46	SY	P	K	Y	OV
11	Y	M	N	N	N	47	S	P	K	Y	OV
12	B	M	N	N	S	48	B	M	D	Y	OV
13	S	P	N	Y	S	49	Y	M	D	Y	OV
14	SY	P	N	Y	S	50	SY	M	D	Y	OV
15	Y	P	N	Y	S	51	S	M	D	Y	OV
16	S	SM	N	Y	S	52	B	SM	D	Y	OV
17	B	P	B	N	S	53	Y	SM	D	Y	OV
18	Y	P	B	N	S	54	SY	SM	D	Y	OV
19	SY	P	B	N	S	55	S	SM	D	Y	OV
20	S	P	B	N	S	56	B	P	D	Y	OV
21	SY	SM	B	N	S	57	Y	P	D	Y	OV
22	S	SM	B	N	S	58	SY	P	D	Y	OV
23	B	M	N	Y	V	59	S	P	D	Y	OV
24	Y	M	N	Y	V	60	B	M	B	Y	OV
25	SY	M	N	Y	V	61	Y	M	B	Y	OV
26	B	SM	N	Y	V	62	SY	M	B	Y	OV
27	Y	SM	N	Y	V	63	S	M	B	Y	OV
28	SY	SM	N	Y	V	64	B	SM	B	Y	OV
29	B	P	N	Y	V	65	Y	SM	B	Y	OV
30	Y	P	N	Y	V	66	SY	SM	B	Y	OV
31	B	M	B	N	V	67	S	SM	B	Y	OV
32	Y	M	B	N	V	68	B	P	B	Y	OV
33	SY	M	B	N	V	69	Y	P	B	Y	OV
34	B	SM	B	N	V	70	SY	P	B	Y	OV
35	Y	SM	B	N	V	71	S	P	B	Y	OV
36	B	M	K	Y	OV						

Без предварительного снижения разрядности и без использования нечетких правил, точность (60 ± 18,37) %			
	Фактический класс 0	Фактический класс 1	Точность распознавания
Предполагаемый класс 0	25	19	56,82 %
Предполагаемый класс 1	10	26	72,22 %
Точность предсказания	71,43 %	57,78 %	
С использованием нечетких правил, точность (63,75 ± 19,72) %			
	Фактический класс 0	Фактический класс 1	Точность распознавания
Предполагаемый класс 0	25	19	56,82 %
Предполагаемый класс 1	10	26	72,22 %
Точность предсказания	71,43 %	57,78 %	
С предварительным снижением размерности и без использования нечетких правил, точность (60,00 ± 9,35) %			
	Фактический класс 0	Фактический класс 1	Точность распознавания
Предполагаемый класс 0	11	8	57,89 %
Предполагаемый класс 1	24	37	60,66 %
Точность предсказания	31,43 %	82,22 %	
С предварительным снижением размерности и использованием нечетких правил, точность (68,75 ± 16,06) %			
	Фактический класс 0	Фактический класс 1	Точность распознавания
Предполагаемый класс 0	27	17	61,36 %
Предполагаемый класс 1	8	28	77,78 %
Точность предсказания	77,14 %	62,22 %	

■ ***Рис. 3.*** Точность классификатора на основе метода опорных векторов

■ ***Таблица 3.*** Точность классификаторов согласно кросс-валидации, %

Метод	Классификатор без предварительного снижения размерности	Классификатор с предварительным снижением размерности, без использования нечеткой логики	Классификатор без предварительного снижения размерности, с использованием нечеткой логики	Классификатор с предварительным снижением размерности и добавлением нового атрибута
Опорных векторов	60	60	63,75	68,75
k-ближайших соседей	52,5	52,5	52,5	53,75
Наивный байесовский классификатор	62,5	62,5	67,5	61,25
Нейронные сети	53,75	62,5	57,5	61,25

правил использовался метод max-дизъюнкции. В качестве метода дефаззификации планируется использовать метод центра тяжести.

Результаты и обсуждение

Для определения уровня точности использовались метод опорных векторов, метод *k*-ближайших соседей, наивный байесовский классификатор, нейронные сети. Оценка классификаторов производилась с помощью методов кросс-валидации. В качестве примера на рис. 3 показано различие между точностью классификации пациентов согласно кросс-валидации с использованием метода опорных векторов.

Результаты точности работы всех классификаторов без предобработки и с предобработкой сведены в табл. 3.

Классификаторы без предобработки показали довольно посредственный результат: точность определения наличия заболевания с помощью процедуры кросс-валидации составила в среднем 57,19 %.

Уровень точности после снижения размерности повысился в среднем на 2,19 %, данное значение остается посредственным (59,38 %). Такой результат можно объяснить загрубленной оценкой порядковых атрибутов. Фаззифицируя эти атрибуты с помощью лингвистических переменных, эксперты могут дать больше информации о градации соответствующих данных. В результате использования нечетких правил набор исходных атрибутов дополнился новым атрибутом, отражающим вероятность наличия заболевания по порядковым атрибутам. Уровень точности после добавления нового атрибута в среднем увеличился на 3,13 % и составил

60,31 %. Такой уровень уже можно признать удовлетворительным.

Заключение

В данной статье приведен пример использования нечеткой логики в совокупности с методами интеллектуального анализа данных (Data Mining) и привлечением эксперта. Нечеткие правила были сформулированы с применением деревьев решений и привлечением эксперта — заведующего кафедрой клинической лабораторной диагностики с курсом молекулярной СПбГМУ им. академика И. П. Павлова, доктора медицинских наук, профессора В. Л. Эмануэля.

Совместное применение нескольких математических методов обработки данных позволило повысить точность классификации, согласно результатам кросс-валидации, в среднем на 4,06 %. Максимальное увеличение точности было достигнуто с использованием классификатора на основе метода опорных векторов и составило 8,75 %.

Литература

1. **Дюк В. А., Самойленко А. П.** Data Mining: учебный курс. – СПб.: Питер, 2001. – 368 с.
2. **Дюк В., Эмануэль В.** Информационные технологии в медико-биологических исследованиях. – СПб.: Питер, 2003. – 528 с.
3. **Колесников А. В.** Гибридные интеллектуальные системы: Теория и технология разработки / под ред. А. М. Яшина/СПбГТУ. – СПб., 2001. – 711 с.
4. **Simerville J. A., Maxted W. C., Pahira J. J.** Urinalysis (review)// Amer. Fam. Physician. 2005. Vol. 71. N 6. P. 1153–1162.
5. **Эмануэль В. Л.** Пособие для семейного врача по лабораторным технологиям и интерпретации исследования мочи: учеб. пособие. – СПб.: Триада; Тверь: Триада, 2007. – 128 с.