ТЕХНИЧЕСКИЕ НАУКИ

УДК 004.8 Научная статья
DOI: 10.35330/1991-6639-2021-6-104-58-65

ПРИМЕНЕНИЕ МЕТОДА МАШИННОГО ОБУЧЕНИЯ ДЛЯ РЕШЕНИЯ ЗАДАЧИ МЕДИЦИНСКОЙ ДИАГНОСТИКИ

Л.А. ЛЮТИКОВА, Е.В. ШМАТОВА

Институт прикладной математики и автоматизации –
филиал Кабардино-Балкарского научного центра Российской академии наук
360000, Россия, Нальчик, ул. Шортанова, 89 А

Аннотация. В работе решается задача создания программного комплекса для компьютерной диагностики гастрита. В качестве входных данных используются показатели обследования пациентов и их диагнозы. Для успешного решения поставленной задачи разрабатывается логический подход анализа данных, который позволяет найти закономерности, необходимые для качественной диагностики. Эти закономерности выявляются, основываясь на данных, предоставленных специалистами, и включают в себя результаты обследований пациентов и существующий в медицинской практике опыт по постановке диагноза. Для выразительного представления данных используются системы многозначной логики предикатов. Предлагается алгоритм, реализующий и упрощающий рассматриваемые подходы. В результате разработанный программный комплекс по данным диагностики пациентов выбирает наиболее подходящие им типы заболевания с заранее заданной точностью. Если с заданной точностью по результатам обследования поставить диагноз не представляется возможным, то либо изменяется точность решения, либо предлагается пройти дополнительное обследование.

Ключевые слова: диагностика, база знаний, алгоритм, дизъюнкты, аксиомы

Статья поступила в редакцию 02.11.2021 *Принята к публикации 29.11.2021*

Для цитирования. Лютикова Л.А., Шматова Е.В. Применение метода машинного обучения для решения задачи медицинской диагностики // Известия Кабардино-Балкарского научного центра РАН. 2021. № 6 (104). С. 58–65. DOI: 10.35330/1991-6639-2021-6-104-58-65

Введение

Медицинская диагностика является достаточно известной задачей. Существуют различные методы ее решения, которые зависят от типа системы и ее назначения.

Это могут быть системы, основанные на статистических и других математических моделях, – их основой служат математические алгоритмы, которые осуществляют поиск частичного соответствия между симптомами наблюдаемого пациента и симптомами ранее наблюдавшихся пациентов, диагнозы которых известны [1–3].

Могут быть системы, основанные на знаниях экспертов. В них алгоритмы оперируют знаниями о заболеваниях, представленных в форме, приближенной к представлениям врачей и описанных экспертами-врачами.

Это могут быть системы на основе машинного обучения, которые нуждаются в достаточно большом количестве данных. Именно в этом случае алгоритм способен обучиться для самостоятельной работы.

Цель данной работы – разработка метода анализа данных и создание на его основе адекватного программного комплекса для диагностики гастрита.

Предлагаемый метод основан на логическом анализе данных и построении сложной дискретной функции, переменными которой являются соответствующим образом представленные симптомы и диагнозы. Это дает возможность даже при небольшом объеме

данных находить закономерности, строить классы по выявляемой общности признаков и отбирать наиболее важные свойства для принятия решения.

Постановка задачи

У 132 пациентов проводилась диагностика гастрита по данным гастроэнтерологических обследований. Данные были предоставлены республиканской клинической больницей МЗ КБР**.** Всего 28 симптомов, каждый из которых имеет от 2 до 4 вариантов ответов. Все симптомы приведены на рисунке 1. Количество диагностируемых типов гастрита – 17, таких, например, как хронический гастрит, хеликобактерный, аутоиммунный, рефлюкс-гастрит, радиационный, гранулематозный (болезнь Крона), эозинофильный, болезнь Менетрие и т. д. По этим данным нужно построить алгоритм для адекватной диагностики остальных пациентов.

Образец анкеты с пунктами, необходимыми для постановки диагноза, выглядит следующим образом:

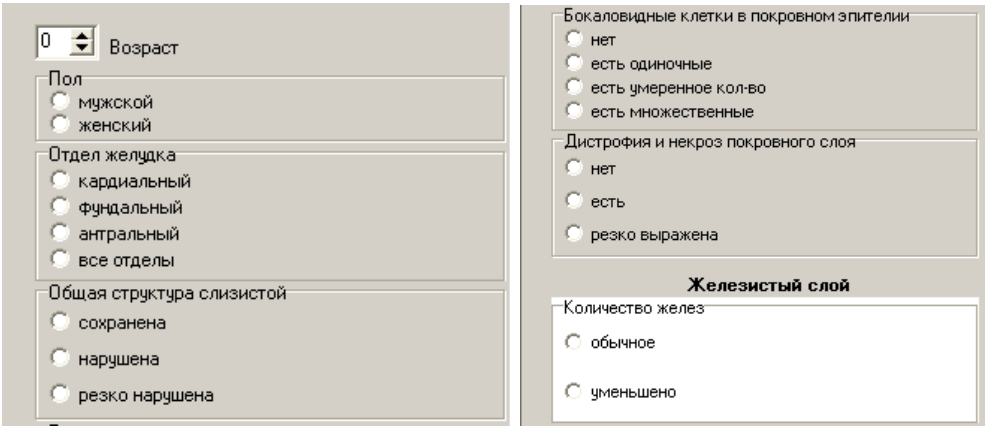

Рис. 1.

Можно говорить, что задана функция от 28 переменных, которая определена в 132 точках, область определения каждой переменой имеет разброс от 2 до 4. Нужно по этим данным восстановить значение функции в других запрашиваемых точках.

Постановка данной задачи сводится к постановке задачи по прецедентам, где симптомы – это вектор значений $X = \{x_1, x_2, ..., x_n\}$

$x_i \in \{0,1,...,k_i - 1\}$. В нашей системе входными данными будут являться n=28, а выходными *m*=17:

$$\begin{pmatrix} x_1(y_1) & x_2(y_1) & ... & x_n(y_1) \\ x_1(y_2) & x_2(y_2) & ... & x_n(y_2) \\ ... & ... & ... & ... \\ x_1(y_m) & x_2(y_m) & ... & x_n(y_m) \end{pmatrix} \to \begin{pmatrix} y_1 \\ y_2 \\ ... \\ y_m \end{pmatrix}. \quad (1)$$

Необходимо найти общие правила, порождающие заданную закономерность (1), исключить неинформативные переменные, разбить совокупность диагнозов на классы.

Может быть, в такой небольшой области знаний, как определение разновидности гастрита, хороший специалист, основываясь на своем опыте, даст более объективное и полное представление о возможностях постановки диагноза, чем те, которые будут получены в результате работы предлагаемого метода. Но важен общий подход, основанный на логическом анализе данных, который позволяет формально находить наиболее важные правила, совокупность которых способна полностью восстановить исходную информацию [4, 5].

Методы решения

Каждая строка (1) является зависимостью и может быть представлена следующим правилом: $\&_{j=1}^{m} x_j\left(y_i,\right) \to y_i$.

Эти правила, описывающие зависимость конкретного диагноза от проведенных исследований, представим в следующей дизъюнктивной форме: $\vee_{j=1}^{m} \overline{x_j\,(y_i)} \vee y_i$. А зависимость всех диагнозов от всех симптомов опишем следующей функцией: $f(x,y) = \&_{i=1}^{n} \vee_{j=1}^{m} \overline{x_j\,(y_i)} \vee y_i$.

Один и тот же диагноз может характеризоваться разной симптоматикой, данная функция поможет исключить несущественные симптомы, разобьет данные на классы. Вообще объединение каждого отдельного правила в общую функцию операцией конъюнкции предлагает широкой диапазон трактовки данных. В итоге мы получаем булеву функцию от $m+n$ переменных (симптомы и диагнозы), которая на каждом наборе будет равна единице, кроме тех наборов, где присутствуют все симптомы, но отрицается диагноз, соответствующий этим симптомам. Можно говорить, что данная функция допускает любые правила, кроме отрицания тех, которые существуют.

Пример. Пусть заданы следующие соотношения:

Таблица 1

x_1	x_2	W
0	*0*	*A*
0	*1*	*B*

$$X_1 = \{x_1(a), x_2(a)\}, \quad X_2 = \{x_1(b), x_2(b)\}, \quad W = \{a,b\} \quad w_1 = a, \quad w_2 = b.$$

Построим таблицу, задающую функцию для переменных: x_1, x_2, $y(a)$,$y(b)$.

Таблица 2

x_1	x_2	$y(a)$	$y(b)$	$f(X,Y)$
0	*0*	*0*	*0*	*0*
0	*0*	*0*	*1*	*0*
0	*0*	*1*	*0*	*1*
0	*0*	*1*	*1*	*1*
0	*1*	*0*	*0*	*0*
0	*1*	*0*	*1*	*1*
0	*1*	*1*	*0*	*0*
0	*1*	*1*	*1*	*1*
1	*0*	*0*	*0*	*1*
1	*0*	*0*	*1*	*1*
1	*0*	*1*	*0*	*1*
1	*0*	*1*	*1*	*1*
1	*1*	*0*	*0*	*1*
1	*1*	*0*	*1*	*1*
1	*1*	*1*	*0*	*1*
1	*1*	*1*	*1*	*1*

Из таблицы видно, что *f(X,Y)=0,* когда заданы признаки объекта (*x₁=0, x₂=0),* при этом сам объект отрицается *y(a)=0,* а также признаки объекта *b* (*x₁=1, x₂=0),* при этом *y(b)=0.*

СКНФ, построенная по таблице 2, будет выглядеть следующим образом:

$$f(X,Y) = (x_1 \vee x_2 \vee y(a) \vee y(b)) \,\&\, (x_1 \vee x_2 \vee y(a) \vee \overline{y(b)}) \,\&$$
$$(\overline{x_1} \vee x_2 \vee y(a) \vee y(b)) \,\&\, (\overline{x_1} \vee x_2 \vee \overline{y(a)} \vee y(b)) = x_2 \vee y(a) \,\&\, \overline{x_1} \vee y(b) \,\&\, x_1$$

Эта функция может легко модифицироваться. Каждое новое правило путем операции конъюнкции входит в систему уже существующих, с возможной некоторой их модификацией [6–7].

Также она может быть представлена в следующем рекурсивном виде:

$$W(X) = Z_k(q_k w_k X);$$
$$Z_k(q_k w_k X_k) = Z_{k-1} \,\&\, (\overset{n}{\underset{i=1}{\vee}} \overline{x_k(w_i)} \vee w_k) \vee q_{k-1} \,\&\, (\overset{n}{\underset{i=1}{\vee}} \overline{x_k(w_i)} \vee w_k);$$
$$q_k = q_{k-1} \,\&\, (\overset{n}{\underset{i=1}{\vee}} \overline{x_k}(w_i); q_1 = \overset{n}{\underset{i=1}{\vee}} \overline{x_1(w_i)}); \; j = 2...m; Z_1 = w_1.$$

где *W(X)* – моделируемая функция, Z_j – характеристика объектов на текущий момент, Q_j – состояние системы на текущий момент. Состояние системы – это элементы настройки.

Если функцию представить в СДНФ и удачно сократить, она может выражать компактное представление данных. Причем структурированных, в которых будут наши диагнозы, будут классы, в которые диагнозы объединяются по сходным симптомам, и будут сочетания симптомов, не характерные для рассматриваемых диагнозов.

Вообще в случае больших данных такой подход может выглядеть несколько громоздким, поэтому дальше предлагается алгоритм для реализации этого метода.

Алгоритмы моделирования системы знаний

Алгоритм отбора правил, из которых можно получить весь объем рассматриваемых данных, может быть следующим: количество столбцов в таблице $\sum_{i=1}^{n} k_i$ – это количество пунктов диагностики с учетом числа вопросов в каждом пункте. Количество строк будет соответствовать количеству диагнозов, в нашем случае это 17 плюс количество классов, которые будут найдены.

Соблюдая порядок следования, записываем данные по каждому пункту всех пациентов в таблицу следующим образом.

Берем каждого пациента, его диагноз разносим по соответствующим столбцам, диагноз w_1 будет размещен в столбце каждого пункта в соответствии с результатами обследования этого пациента. Например, пункт «пол» будет иметь два столбца со значениями 0 и 1, и диагноз будет помещен в столбец в зависимости от пола пациента. Общий вид таблицы приведен ниже.

Таблица 3

0_1	1_1	k_1-1	0_2	1_2.....	k_2-1....	..0_n	1_n	k_n-1
	w_1		w_1					w_{n1}
	w_2				w_2		w_2	
w_m				w_m				w_m

По ходу заполнения таблицы проверяем столбец, в который попадает диагноз рассматриваемого пациента. Если в столбце уже есть другие диагнозы, то вычеркиваем их и заносим в класс с рассматриваемым диагнозом, заносим их в следующую строку в тот же столбец. Эти диагнозы объединяются в класс по данному пункту диагностики, что продемонстрировано в таблице 4.

Таблица 4

0_1	1_1	k_1-1	0_2	1_2.....	k_2-1....	..0_n	1_n	k_n-1
	~~w_1~~		w_1					~~w_1~~
	~~w_2~~				w_2		w_2	
	w_1 w_2							
w_m				w_m				~~w_m~~
								w_1 w_m

Далее последовательно рассматриваем строки, если в строке, соответствующей какому-либо диагнозу, остались в квадратах не вычеркнутые диагнозы, то выбираем соответствующий этому диагнозу столбец и считаем это уникальным признаком именно этого диагноза. Также рассматриваем классы, образованные в результате анализа данных [8].

Таким образом, алгоритм позволяет построить те дизъюнкты, которые содержат диагнозы, т. е., по которым и ведется распознавание.

Описание программы

Программа реализует представленный выше алгоритм, состоит из двух исполняемых модулей:

Модуль 1. Выполняет декодирование базы данных с использованием словаря, загружает симптомы и диагнозы в форме вопросов и ответов и анализирует результаты.

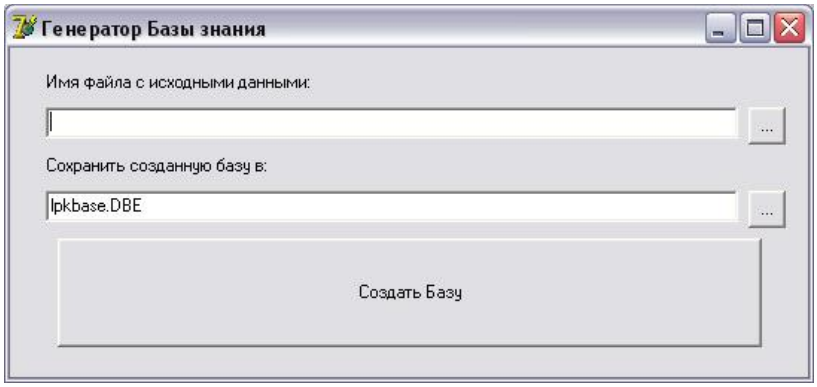

Рис. 2.

Модуль 2. Создает информацию на основе исходного файла с данными или для уточнения этой системы знаний. Уменьшает размер базы данных в соответствии с приблизи-

тельным значением, тогда стоит либо снизить точность алгоритма, либо добавить информацию для проверки правильности сохраненных данных.

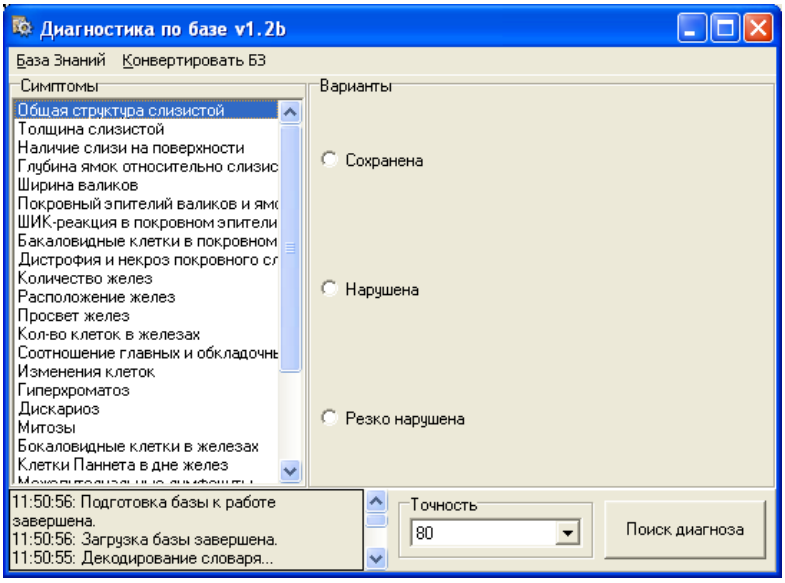

Рис. 3.

После заполнения всех полей, приведенных на рисунке 3, можем получить результат диагностики с заданной точностью (рис. 4). Там же все рассматриваемые симптомы для постановки диагноза.

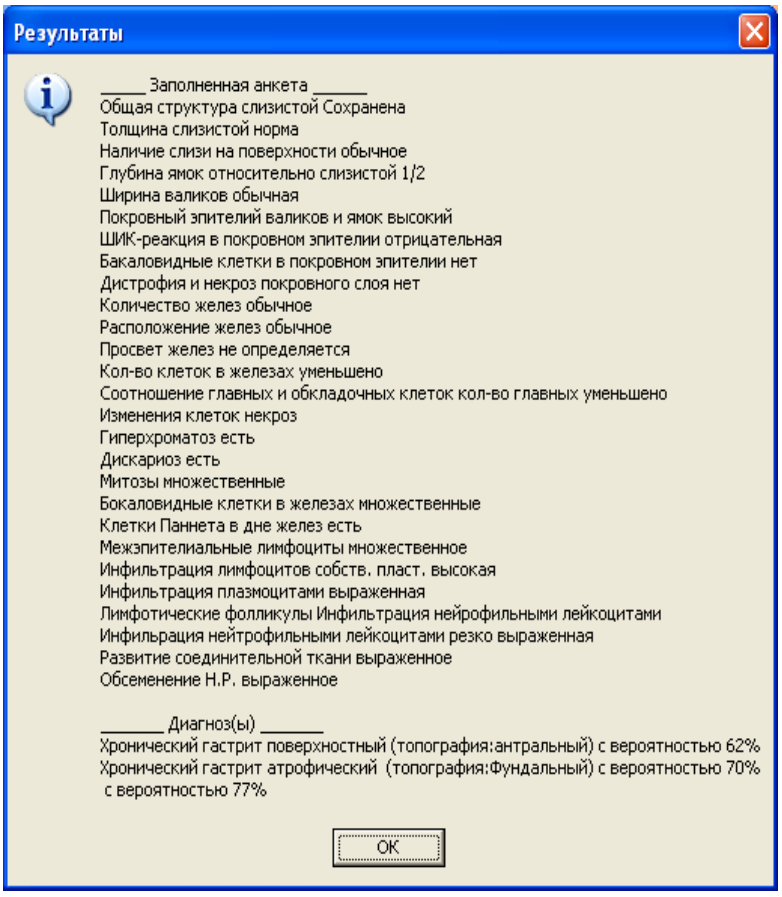

Рис. 4

Заключение

Логические алгоритмы могут быть использованы для анализа данных, они дают возможность рассмотреть исходные данные как некий набор общих правил, среди которых можно выявить минимальных набор тех правил, которых достаточно, чтобы получить все рассматриваемые. Эти правила будут порождающими для рассматриваемой области, они помогут лучше понять природу исследуемых объектов и минимизировать поиск правильных ответов.

Информация об авторах

Лютикова Лариса Адольфовна, канд. физ.-мат. наук, зав. отделом «Нейроинформатика и машинное обучение», Институт прикладной математики и автоматизации – филиал Кабардино-Балкарского научного центра РАН;

360000, Россия, Нальчик, ул. Шортанова, 89 А;

lylarisa@yandex.ru, ORCID: https://orcid.org/0000-0003-4941-7854

Шматова Елена Витальевна, стажер-исследователь отдела «Нейроинформатика и машинное обучение», Институт прикладной математики и автоматизации – филиал Кабардино-Балкарского научного центра РАН;

360000, Россия, Нальчик, ул. Шортанова, 89 А;

lenavsh@yandex.ru, ORCID: https://orcid.org/0000-0003-1344-1924

СПИСОК ЛИТЕРАТУРЫ / REFERENCES

1. *Zhuravlev Yu.I.* On an algebraic approach to solving problems of recognition or classification. *Problemy kibernetiki* [Problems of Cybernetics]. 1978. Vol. 33. Pp. 5–68. (In Russian)

2. *Shibzukhov Z.M.* Correct Aggregation Operations with Algorithms. Pattern Recognition and Image Analysis. 2014, Vol. 24. No. 3. Pp. 377–382.

3. *Ashley I. Naimi, Laura B.* Balzer Stacked generalization: an introduction to super learning. European Journal of Epidemiology. 2018. No. 33. Pp. 459–464.

4. *Mesiar R., Komornikova M., Kolesarova A., Calvo T.* Fuzzy Aggregation Functions: A revision. Sets and Their Extensions: Representation, Aggregation and Models. Springer-Verlag, Berlin, 2008.

5. *Fan Yang Zhilin Yang William W.* Cohen Differentiable Learning of Logical Rules for Knowledge Base Reasoning. Advances in Neural Information Processing Systems. Vol. 2017. 2017. Pages 2320–2329.

6. *Flach P.* Machine Learning: The Art and Science of Algorithms that Make Sense of Data. Cambridge University Press, 2012. 396 p. ISBN: 978-1107096394.

7. *Rahman Akhlaqur, Tasnim Sumaira.* Ensemble Classifiers and Their Applications: A Review. International Journal of Computer Trends and Technology. 2014. Vol. 10. No 1. Pp. 31–35.

8. *Lyutikova L.A., Shmatova E.V.* Application of Variable-Valued Logic to Correct Pattern Recognition Algorithms. Advances in Intelligent Systems and Computing. 2020. Vol. 948. Pp. 308–314.

9. *Christopher J.C. Burges.* A Tutorial on Support Vector Machines for Pattern Recognition, Appeared in: Data Mining and Knowledge Discovery 2. 1998. Pp. 121–167.

10. *Lyutikova L.A.* Use of logic with a variable valency under knowledge bases modeling. CSR-2006.

MSC: 68T30 Original article

APPLICATION OF THE MACHINE LEARNING METHOD TO SOLVE THE PROBLEM OF MEDICAL DIAGNOSTICS

L.A. LYUTIKOVA, E.V. SHMATOVA

Institute of Applied Mathematics and Automation –
branch of Kabardino-Balkarian Scientific Center of the Russian Academy of Sciences
360000, Russia, Nalchik, 89 A Shortanov street

Annotation. The paper solves the problem of creating a software package for computer diagnostics of gastritis. Patient examination indicators and their diagnoses are used as input data. To successfully solve the task, a logical approach to data analysis is being developed, which allows us to find the patterns necessary for qualitative diagnostics. These patterns are identified based on the data provided by specialists and include the results of patient examinations and existing medical practice experience in diagnosis. Systems of multivalued predicate logic are used for expressive representation of data. An algorithm is proposed that implements and simplifies the approaches under consideration. As a result, the developed software package selects the most suitable types of the disease with a predetermined accuracy according to the data of the diagnosis of patients. If it is not possible to make a diagnosis with a desired accuracy based on the results of the examination, then either the accuracy of the solution should be changed, or the patient is proposed to undergo an additional examination.

Keywords: diagnostics, knowledge base, algorithm, clauses, axioms

The article was submitted 02.11.2021 *Accepted for publication 29.11.2021*

For citation. Lyutikova L.A., Shmatova E.V. Application of the machine learning method to solve the problem of medical diagnostics. News of the Kabardino-Balkarian Scientific Center of RAS. 2021. No. 6 (104). Pp. 58–65. DOI: 10.35330/1991-6639-2021-6-104-58-65

Information about the authors

Lyutikova Larisa Adolfovna, Candidate of Physical and Mathematical Sciences; Head of the Department of Neuroinformatics and Machine Learning, Institute of Applied Mathematics and Automation – branch of Kabardino-Balkarian Scientific Center of the Russian Academy of Sciences;

360000, Russia, Nalchik, 89 A Shortanov street;

lylarisa@yandex.ru, ORCID: https://orcid.org/0000-0003-4941-7854

Shmatova Elena Vitalevna, Trainee Researcher of the Department of Neuroinformatics and machine learning, Institute of Applied Mathematics and Automation – branch of Kabardino-Balkarian Scientific Center of the Russian Academy of Sciences;

360000, Russia, Nalchik, 89 A Shortanov street;

lenavsh@yandex.ru, ORCID: https://orcid.org/0000-0003-1344-1924