УДК 004.9:61+ 519.68

Реализация эффективных моделей классификации медицинских данных методами интеллектуального анализа текстовой информации

О.С. Кротова[1], И.В. Москалев[1], Л.А. Хворова[1], О.М. Назаркина[2]

[1]Алтайский государственный университет (Барнаул, Россия)
[2]Алтайский краевой клинический центр охраны материнства и детства (Барнаул, Россия)

Implementation of Effective Models for Classifying Medical Data Using Text Mining

O.S. Krotova[1], I.V. Moskalev[1], L.A. Khvorova[1], O.M. Nazarkina[2]

[1]Altai State University (Barnaul, Russia)
[2]Altai Regional Clinical Center for Maternal and Child Health (Barnaul, Russia)

Статья посвящена разработке и реализации эффективных моделей классификации медицинских данных методами интеллектуального анализа текстовой информации для поддержки принятия решений при диагностике пульмонологических заболеваний у детей и подростков Алтайского края. Медицинские данные содержат важную информацию о пациентах. В структурированном виде, как правило, хранятся результаты анализов. Такие данные, как анамнезы, результаты осмотров, описания результатов обследований, имеют неструктурированную форму (в виде текстов на естественном языке). В работе дана оценка качества разработанных методов и моделей извлечения информации из клинических текстов на русском языке. Проведена оценка метода автоматической диагностики пульмонологических заболеваний на тестовой выборке. Определены наиболее информативные признаки, а также подходящие методы машинного обучения для классификации пациентов по группам заболеваний. Применение методов интеллектуального анализа и обработки данных позволит автоматизировать решение многих медицинских задач, возникающих в клинической практике, повысив тем самым качество первичной медицинской помощи. Результаты исследования свидетельствуют о перспективности использования разработанных моделей для поддержки принятия решений при диагностике пульмонологических заболеваний у детей.

Ключевые слова: пульмонологические заболевания у детей, методы интеллектуальной диагностики, методы машинного обучения, лингвистический анализ текстов, интеллектуальная обработка медицинских данных.

The paper is devoted to the development and implementation of effective models of medical data classification by text mining for decision support in the diagnosis of pulmonological diseases in children and adolescents of the Altai Territory. Medical data contains important information about patients. Test results are usually retained as structured data, but some data are retained in the form of natural language texts (medical history, the results of physical examination, and the results of other examinations). The paper assesses the quality of the developed methods for extracting information from clinical texts. An assessment of the method for the automatic diagnosis of pulmonological diseases in a test sample is conducted. The most informative features, as well as suitable machine learning methods for classifying patients by disease groups, are identified. Many tasks arising in clinical practice can be automated by applying methods for intelligent analysis of structured and unstructured data that will lead to improvement of the healthcare quality. The results of the research indicate the prospect of using models to support decision-making in the primary diagnosis of pulmonological diseases in children and adolescents of the Altai Territory.

Key words: pulmonological diseases in children, methods of intellectual diagnostics, methods of machine learning, linguistic analysis of texts, intellectual processing of medical data.

DOI 10.14258/izvasu(2020)1-16

1. Введение

Анализ медицинских данных и создание систем поддержки принятия врачебных решений в диагностике — одно из актуальных и развиваемых в настоящее время направлений применения искусственного интеллекта.

В последние годы актуальными научными направлениями в медицине являются: переход к передовым цифровым, интеллектуальным технологиям, создание систем обработки больших объемов данных, развитие машинного обучения и искусственного интеллекта [1]; переход к персонализированной медицине, высокотехнологичному здравоохранению и технологиям здоровьесбережения, что соответствует Приоритетным направлениям Стратегии научно-технологического развития России [2] и Приоритетным направлениям социально-экономического развития Алтайского края [3].

Обзор и анализ современных публикаций по данной тематике показал, что значительное число работ посвящено описанию и применению методов интеллектуальной диагностики для прогнозирования рака груди [4], стадий хронической почечной недостаточности [5], астмы [6], диабета [7], диагностики хронических заболеваний для пациентов детской возрастной категории с болезнями органов дыхания, с аллергическими, нефрологическими и ревматическими болезнями [8]. В качестве методов интеллектуальной диагностики выступают методы машинного обучения: деревья решений, логистическая регрессия, градиентный бустинг.

Цель данной работы — разработка и реализация эффективных моделей классификации медицинских данных методами интеллектуального анализа текста для поддержки принятия решений при первичной диагностике пульмонологических заболеваний у детей и подростков Алтайского края.

2. Задача диагностики пульмонологических заболеваний у детей и подростков

Постановка задачи диагностики пульмонологических заболеваний у детей и подростков сводится к задаче классификации: пусть X — пространство объектов — выписки из историй болезни пациентов. Атрибутом объектов является признаковое описание пациента, которое представляет собой формализованную историю болезни. Признаковое пространство каждого объекта $X = \{x_1, x_2, ..., x_m\}$ включает: признаки, характеризующие результаты обследований; симптомы заболевания и применявшиеся методы лечения; бинарные признаки — пол, наличие головной боли, слабости и т.д.; порядковые признаки — тяжесть состояния (удовлетворительное, средней тяжести, тяжелое, крайне тяжелое); количественные признаки — возраст, пульс, артериальное давление, содержание гемоглобина в крови и т.д. Пусть Y представляет собой конечное множество классов — диагнозы. Требуется построить такой алгоритм $\alpha : X \to Y$, который любому объекту $x \in X$ ставит в соответствие метку класса $y \in Y$ [9].

Информационная база исследования представлена обезличенными выписками из историй болезни пациентов в текстовой форме на естественном языке. Выписки содержат разнородную информацию (структурированные и неструктурированные медицинские данные) о пациентах и проведенных обследованиях. К структурированным данным относятся результаты анализов; к неструктурированным — анамнез и состояние пациента при поступлении в медицинское учреждение, результаты обследований у узких специалистов, результаты ультразвукового исследования, эхокардиограммы, рентгенограммы грудной клетки и другие. Целевым показателем является одна из категорий заболеваний бронхолегочной системы: бронхит, пневмония, астма, муковисцидоз, плеврит, другие заболевания.

3. Методы интеллектуальной обработки текстов

Для построения моделей классификации медицинских данных методами машинного обучения для поддержки принятия решений при диагностике пульмонологических заболеваний у детей и подростков осуществлялось извлечение информации из клинических текстов на русском языке (неструктурированные данные) на основе их полного лингвистического анализа. Трудности обработки медицинских текстов из карты пациента обусловлены отсутствием структуры в тексте, отсутствием эталонных фрагментов, а также наличием синтаксического шума, синонимии и неоднозначностей. Как показал проведенный анализ литературы по данной тематике, на сегодняшний момент не существует готового инструмента для обработки тестовых медицинских данных на русском языке [10, 11]. Всем исследователям приходится решать эту задачу за счет совместного использования нескольких методов и алгоритмов, приводящих к построению эффективных моделей.

В нашем случае из текстов извлекались: предварительный диагноз при поступлении, симптомы, тяжесть заболевания, течение заболевания. Для извлечения информации применялись медицинские тезаурусы, набор шаблонов — устойчивых, повторяющихся языковых конструкций, имеющих определенный семантический смысл и синтаксическую структуру, а также токенизация — процесс разделения письменного языка на отдельные компоненты (предложения, слова). Отметим, что предварительная обработка текстовых данных осуществлялась с использованием методов фильтрации и нормализации. Общая блок-схема процесса извлечения знаний из медицинских текстовых данных приведена на рисунке 1.

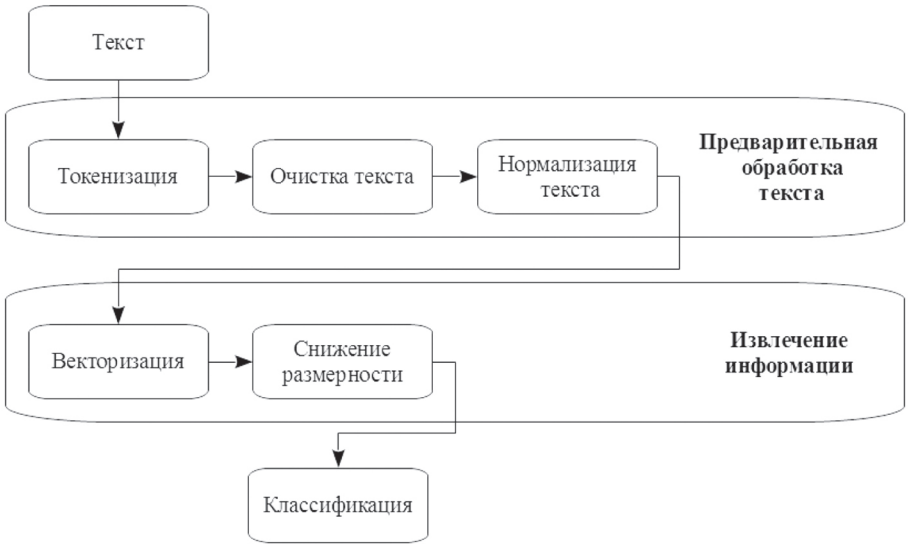

Рис. 1. Блок-схема процесса извлечения знаний из медицинских текстовых данных

4. Построение интеллектуальных моделей классификации медицинских данных и диагностики пульмонологических заболеваний

Исследование проводилось с использованием языка программирования Python; набора инструментов для работы с естественными языками Natural language toolkit, библиотек Pandas, NumPy, Scikit-Learn, Matplotlib, Yellowbrick.

4.1. Предварительная обработка текстовых данных

Исходные данные представляют собой базу из 1100 документов в формате .docx, которые были разделены на 6 непересекающихся классов, 5 из которых — диагнозы заболеваний бронхолегочной системы: астма (250 выписок), бронхит (396), муковисцидоз (60), пневмония (204), плеврит (22). В последний класс (168 выписок) попали выписки, диагнозы в которых не соответствуют ни одной из 5 основных категорий. Классы являются несбалансированными. Это следует учитывать при выборе итоговой метрики классификации.

Основная задача подготовки текста — получение максимального количества информации для дальнейшего использования в классификации. Текстовые данные содержали большое количество слов и символов, имеющих малый информационный вес. Были выделены 30 наиболее часто встречающихся токенов: от, анализ, общий, крови и т.д. При этом словарь полученного корпуса насчитывал 9717 слов, суммарное же количество слов — 261 481. Значительная часть из них представлена символами пунктуации, предлогами и союзами, которые не содержат какой-либо информации. Из исходных текстов были удалены слова с малой информативностью или встречающиеся во всех документах. После чего была проведена нормализация оставшихся токенов, которая заключалась в приведении всех слов к нижнему регистру и лемматизации. Таким образом было исключено влияние грамматической составляющей на дальнейший анализ.

По частотному распределению лексем был сформирован итоговый словарь корпуса — 6211 слов и определено общее количество слов во всех документах — 162 503. Это позволило значительно повысить информативность данных и сократить время для проведения исследования.

4.2. Векторизация данных

Библиотека Scikit-Learn предлагает несколько способов представления текстовых данных в числовом виде. В процессе исследования путем сравнения качества классификации данных был определен оптимальный метод векторизации — TF-IDF. Метод представляет текстовые данные в виде числовых векторов, отражающих важность использования каждого слова из некоторого набора слов в отдельном документе. Для оценки важности используется одноименная с методом статистическая мера, которая определяется как произведение двух сомножителей:

$$tf\_idf(t,d,D) = tf(t,d) \times idf(t,D),$$

где $tf(t,d)$ определяется как отношение числа вхождений некоторого слова к общему числу слов документа:

$$tf(t,d) = \frac{n_t}{\sum_k n_k},$$

n_t — число вхождения слова t во всех документах, а в знаменателе — общее число слов в данном документе.

$idf(t,D)$ — обратная частота появления слова во всех документах корпуса:

$$idf(t, D) = \log_a \frac{|D|}{|\{d_i \in D \,|\, t \in d_i\}|},$$

$|D|$ — число документов в коллекции, $|\{d_i \in D \,|\, t \in d_i\}|$ — число документов из коллекции D, в которых встречается t.

Выбор основания логарифма a в формуле не имеет значения, поскольку изменение основания приводит к изменению веса каждого слова на постоянный множитель, что не влияет на соотношение весов. Чем ближе к единице показатель TF-IDF, тем больше информации в себе несет данное слово. Данный векторизатор естественным образом решает проблему стоп-слов, которые, как правило, присутствуют во всех документах, но получают крайне малый вес.

В дальнейшем при построении моделей классификации был использован алгоритм TF-IDF, реализованный в классе TfidfVectorizer библиотеки Scikit-Learn. В результате векторизации методом TF-IDF были получены оценки значимости токенов для каждого класса.

Размерность данных, полученных при векторизации текстов, оказалась достаточно высокой, а содержащиеся в них некоторые признаки малоинформативны. Поэтому на следующем этапе исследования к исходным данным были применены методы понижения размерности: PCA — метод главных компонент, SVD — сингулярное разложение и t-SNE — метод t-распределенного стохастического вложения соседей.

4.3. Построение моделей классификации

Для построения моделей классификации были использованы следующие методы: наивный байесовский классификатор (NB), случайный лес (RF), градиентный бустинг (GB), логистическая регрессия (LR), метод опорных векторов (SVM) и многослойный персептрон (MLP). Для каждого классификатора было сформировано пространство параметров, по которому осуществлялась оптимизация. В качестве метрики использовался коэффициент корреляции Мэтьюса (mcc). В таблице 1 приведены результаты оценки качества моделей классификации.

Таблица 1

Качество перекрестной проверки на оптимальных параметрах

	NB	LR	SVM	RF	GB	MLP
mcc	0.432	**0.511**	**0.533**	0.495	0.495	0.504

Как следует из таблицы, методы показали различное качество классификации на этапе подбора параметров. Лучшие показатели по сравнению с остальными алгоритмами продемонстрировали алгоритмы логистической регрессии (LR) и метод опорных векторов (SVM).

4.4. Оценка качества моделей

На данном этапе была проведена проверка уже обученных моделей на тестовой выборке — на данных, которые не участвовали в настройке параметров и обучении моделей. Таким образом мы установили степень обобщения и оценили качество предсказаний моделей.

Для оценки качества пользовались следующими метриками: полнота (Recall), точность (Precision), F-мера (f1-score), mcc. В таблице 2 приведены значения метрик качества, полученных на тестовой выборке.

Таблица 2

Значения метрик качества на тестовой выборке

	Полнота (Recall)	Точность (Precision)	F-мера (f1-score)	mcc
NB	0.536	0.544	0.533	0.386
LR	0.639	0.643	0.635	0.521
SVM	0.645	0.644	0.641	0.535
RF	0.606	0.594	0.585	0.472
GB	0.624	0.619	0.608	0.497
MLP	0.636	0.657	0.625	0.510

Из таблицы 2 следует, что метод опорных векторов (SVM) обладает высокими значениями качества. Хорошие результаты имеют многослойный персептрон и логистическая регрессия.

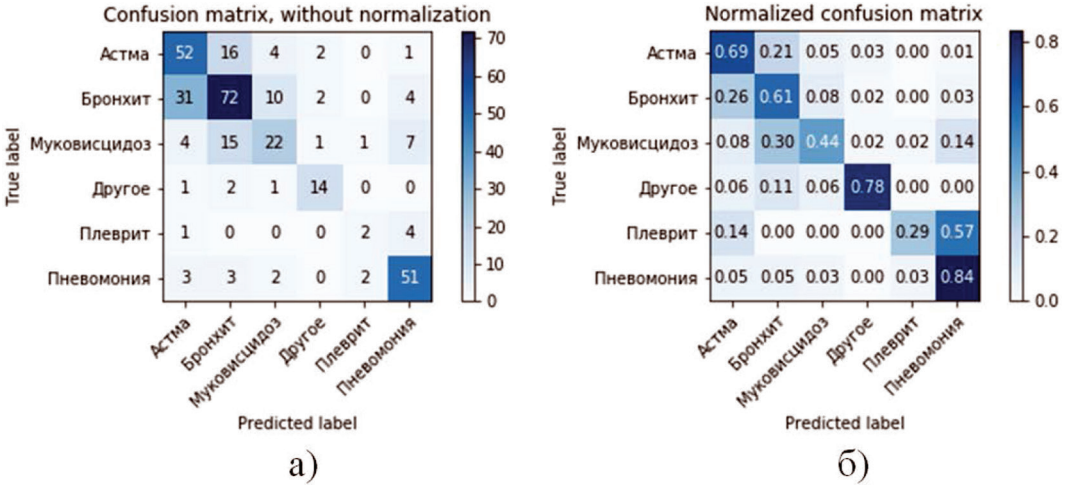

Рис. 2. Матрицы ошибок метода опорных векторов:
а) без нормализации; б) с нормализацией

Еще одним способом оценить качество построенных моделей классификации служит матрица ошибок. На рисунке 2 приведены ненормализованная и нормализованная матрицы. По элементам матрицы можно судить о том, как влияет дисбаланс классов на общую оценку качества. Несмотря на то что алгоритму удалось правильно классифицировать 72 пациента с диагнозом бронхит, в относительных значениях это только 61 % от общего числа пациентов. С другой стороны, диагноз пневмония был правильно классифицирован с точностью 84 %, что является высоким показателем.

5. Заключение

В работе проведено исследование методов извлечения информации из клинических текстов, методов интеллектуальной диагностики пульмонологических заболеваний у детей и подростков Алтайского края, а также методов определения значимости признаков заболеваний. При выполнении исследования были решены основные научно-технические проблемы: организация первичной обработки разнородных данных о пациентах, неравномерное распределение заболеваний в обучающей выборке и интеграция различных методов интеллектуального анализа данных и текстов в рамках единого исследования.

В процессе исследования были выявлены наиболее значимые для диагностики признаки заболеваний и скрытые зависимости в клинических данных, что свидетельствует о перспективах использования программного модуля в целях автоматизации процесса диагностики. Более того, разработанный программный модуль способен анализировать разнородные данные как структурированного, так и неструктурированного характера и позволяет автоматизировать не только диагностический этап медицинской помощи, но и широкий спектр мероприятий в рамках лечебного процесса. Для пациента с пульмонологическим заболеванием несвоевременность установки диагноза и промедление в назначении лечения являются основной составляющей прогноза течения и исхода болезни. Дальнейшее совершенствование программного модуля, внедрение его в систему поддержки принятия врачебных решений позволит обеспечить процесс качественного оказания медицинской помощи детям. Необходимо отметить, что совместное использование методов анализа структурированных и текстовых данных в рамках единой процедуры позволит значительно повысить качество диагностики пульмонологических заболеваний у детей и подростков.

Библиографический список

1. О развитии искусственного интеллекта в Российской Федерации : Указ Президента РФ от 10.10.2019 № 490. URL: http://www.consultant.ru/document/cons_doc_LAW_335184/.

2. О Стратегии научно-технологического развития Российской Федерации : Указ Президента РФ от 01.12.2016 № 642. URL: http://static.kremlin.ru/media/events/files/ru/uZiATIOJiq5tZsJgqcZLY9YyL8PWTXQb.pdf.

3. Об утверждении стратегии социально-экономического развития Алтайского края до 2025 г. : Закон Алтайского края от 21.11.2012 № 86-ЗС. URL: http://docs.cntd.ru/document/453123097.

4. Isa NAM. Towards intelligent diagnostic system employing integration of mathematical and engineering model // Proceedings of International Conference on Mathematics, Engineering and Industrial Applications. AIP Publishing. 2015. DOI: 10.1063/1.4915633.

5. Abeer Y.A., Ahmad M.A., Majid A.A. Clinical decision support system for diagnosis and management of chronic renal failure // Proceedings of Applied Electrical Engineering and Computing Technologies. IEEE; 2013.

6. Zarandi MHF, Zolnoori M., Moin M., Heidarnejad H. A fuzzy rule-based expert system for diagnosing asthma. Transaction E: Industrial Engineering. 2010. № 17(2).

7. Кротова О.С., Пиянзин А.И., Хворова Л.А. Оценка качества моделей прогнозирования стадий компенсации и декомпенсации сахарного диабета // Ломоносовские чтения на Алтае: фундаментальные проблемы науки и техники : сб. научных ст. междун. конф. 2018.

8. Баранов А.А., Намазова-Баранова Л.С., Смирнов И.В. и др. Технологии комплексного интеллектуального анализа клинических данных // Вестник РАМН. 2016. № 71(2).

9. Воронцов К.В. Математические методы обучения по прецедентам. 2007. URL: http://www.machinelearning.ru/wiki/ images/6/6d/Voron-ML-1.pdf/

10. Лапаев М.В., Водяхо А.И., Смирнов А.Б., Жукова Н.А. Система обработки текстовых медицинских данных // Известия СПбГЭТУ «ЛЭТИ». 2016. № 9.

11. Bird S., Klein E., Loper E. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit // O'Reilly Media, Inc. 2009.