Аналитическая обработка медицинских данных в системе диагностики бронхолёгочных заболеваний

Г.Р. Шахмаметова
Уфимский государственный авиационный технический университет
Уфа, Россия
e-mail: shakhgouzel@mail.ru

А.А. Евграфов
Уфимский государственный авиационный технический университет
Уфа, Россия
e-mail: evgrafov.alexander92@yandex.ru

Р.Х. Зулкарнеев
Башкирский государственный медицинский университет
Уфа, Россия
e-mail: zrustem@ufanet.ru

Аннотация[1]

В данной статье рассмотрены информационные технологии, применяемые в здравоохранении для анализа и обработки медицинских данных, их функциональные возможности, а также предложено решение по разработке системы диагностики бронхолёгочных заболеваний, позволяющей устанавливать первичный диагноз пациента и предлагать эффективные методы его лечения на основе методов интеллектуальной обработки информации и принятия решений.

1. Введение

На сегодняшний день информационные технологии получили активное распространение и развитие практически во всех сферах человеческой жизнедеятельности. Отдельным вектором развития информационных технологий является интеллектуальный анализ данных, предполагающий извлечение полезной информации из большого потока данных, отдельные фрагменты которого могут быть неточными (в том числе противоречивыми), неполными и разнородными.

Как научное направление интеллектуальный анализ данных стал активно развиваться в 90-х годах XX века, что было вызвано широким распространением технологий автоматизированной обработки информации и накоплением в компьютерных системах больших объемов данных [1,2]. И хотя существующие технологии позволяли, например, быстро найти в базе данных нужную информацию, этого во многих случаях было уже недостаточно. Возникла потребность поиска взаимосвязей между отдельными событиями среди больших объёмов данных, для чего понадобились методы математической статистики, теории баз данных, теории искусственного интеллекта и ряда других областей [3].

Частным случаем использования возможностей интеллектуального анализа данных является медицина, в которой эффективность обработки информации, получаемой от пациентов, а также своевременно обновляемой информации о методах лечения различных заболеваний напрямую влияет на качество лечения больных в целом.

Современные подходы к интеллектуальному анализу данных предполагают выявление и извлечение закономерностей из баз фактов, в которых они содержатся в неявном виде, в частности о состоянии здоровья человека. Интеллектуальный анализ данных наиболее эффективен, когда он осуществляется посредством систем, не только имитирующих, но и усиливающих аналитические возможности экспертов [4].

2. Обзор и постановка проблемы

Программные комплексы, используемые в области медицины, условно можно разделить на 4 класса (рисунок 1).

Труды Шестой всероссийской научной конференции "Информационные технологии интеллектуальной поддержки принятия решений", 28-31 мая, Уфа-Ставрополь, Россия, 2018

Рисунок 1 – Виды систем обработки медицинских данных

1. ***Информационные системы обработки медицинских данных*** – обеспечивают базовый функционал по хранению и поиску биомедицинской информации о пациентах, не содержит средств анализа данных.

2. ***Аналитическое ПО общего назначения*** – включает в себя элементы математической статистики и анализа данных, не адаптированных для нужд медицины.

3. ***Аналитическое ПО специального назначения*** – программное обеспечение, содержащее в себе ряд аналитических и статистических аппаратов, адаптированных для использования с биомедицинскими данными.

4. ***Системы поддержки принятия решений*** – содержат алгоритмы информационного поиска, интеллектуального анализа данных, рассуждения на основе прецедентов, позволяющие помогать специалистам принимать решение в сложных условиях для полного и объективного анализа предметной деятельности [5].

К наиболее яркому примеру использования интеллектуальной обработки данных в медицине, относящуюся к классу систем поддержки принятия решений, сегодня с большой долей уверенности можно отнести когнитивную систему «IBM Watson». Данная система применяется в сфере здравоохранения для помощи врачам в работе по различным лечебным специализациям, при этом особенное внимание уделяется лечению онкологических заболеваний. «Watson for Oncology» [6] помогает врачам получить релевантные данные, объединив информацию из разных источников, и определить варианты лечения, которые будут персонализированы для каждого отдельного пациента. Система была применена уже более чем к 14000 онкобольным по всему миру и показала 96% соответствие результатов «Watson» с рекомендациями высококвалифицированных специалистов (по данным исследования, проведённого в онкологическом центре в Банголоре) [7]. При этом основной сложностью использования данной системы является низкое качество входных данных: электронные медицинские данные не имеют однозначной структуры, долгие годы их оцифровка преследовала лишь цели хранения данных, в значительной мере затрудняя поиск полезной информации [8].

Среди более доступных для применения программных комплексов, созданных с целью извлечения полезной информации из скопления необработанных данных, следует отметить «STADIA»[9], «SPSS»[10], «STATA»[11], «STATISTICA»[12], «JMP»[13] и другие. Такие комплексы следует относить к классу аналитического ПО общего назначения. Данные программы позволяют анализировать массивы информации, строить различные графические и статистические модели и выводить результаты в виде сформированных отчётов в доступной форме. Тем не менее, отсутствие адаптации данных продуктов для нужд медицины существенно снижает возможности их применения в клиниках.

К классу аналитического ПО специального назначения можно отнести, к примеру, платформу «HealthSuite Insights», которая предоставляет ученым, клиницистам, разработчикам программного обеспечения и другим представителям сферы здравоохранения доступ к расширенным возможностям хранения и анализа медицинских данных [14]. Однако на сегодняшний день ПО такого типа находится на стадии внедрения в здравоохранение и массово не используется, что затрудняет оценку показателей его эффективности.

В настоящее время рынок программного обеспечения в области информационных систем обработки медицинских данных в России предлагает множество решений по организации, хранению и обработке биомедицинской информации. К решениям такого типа относятся системы «ПроМед» [15], «MEDESK» [16], «MEDODS» [17], «МедОфис» [18], «MedWork» [19] и др. Практически все вышеуказанные продукты устроены по единому принципу и предлагают ряд общих возможностей:

- функционал регистратуры (запись пациентов на приём, отслеживание статуса приёма);

- ведение электронных карт пациентов, учёт готовности и сравнение результатов анализов;

- интеграция с онлайн-кассами.

Таким образом, можно заметить, что применение данных программных комплексов предполагает улучшения в области организационной структуры лечебных учреждений, при этом непосредственному анализу данных, получаемых в ходе диагностирования и лечения пациентов, отводится несущественная роль (возможности ограничиваются исключительно просмотром, ускоренным поиском необходимой информации и её наглядным графическим представлением).

Исходя из вышеперечисленного, можно сделать следующие выводы: используемые в рассмотренных программных средах методы обработки медицинской информации на сегодняшний день являются недостаточно эффективными по причине отсутствия в эксплуатации действующих систем профессионального уровня с функциональными возможностями в области интеллектуальной обработки и принятия решений, ввиду чего задача обработки данных при работе с данными больного, полученными в ходе проведения анализов, и ответственность за результат лечения целиком возлагаются на медицинский персонал. Это приводит к невозможности использования большинства вспомогательных методик обработки информации, требующих адаптации в области медицины и значительного количества вычислительных ресурсов, а также увеличивает риск погрешности человеческого фактора, вследствие чего разработка распределённой системы обработки медицинских данных является актуальной и необходимой задачей.

3. Предлагаемое решение

В связи с невозможностью одновременного охвата широкого спектра диагностируемых заболеваний ввиду существенного увеличения трудоёмкости, а также по причине наибольшего распространения в России таких заболеваний, как пневмония и бронхит [20], в качестве диагностируемых выбран ряд бронхолёгочных заболеваний. Впоследствии система предполагает возможность своего эффективного использования на широком спектре заболеваний (в том числе и за пределами бронхолёгочных) при условии интеграции в систему соответствующих баз знаний.

К бронхолёгочным заболеваниям относятся такие виды заболеваний, как острый бронхит, асбестоз, пневмония, бронхиальная астма, ателектаз и другие [21].

Перечисленные заболевания можно регистрировать на основании полученных от пациентов жалоб и результатов анализов, в том числе и средствами машинной обработки данных.

Входным объёмом данных для системы будет являться информация, полученная в ходе проведения следующих методов диагностики:

- физикальные (визуальный осмотр, выслушивание и т.д.);

- инструментальные (рентген,УЗИ, МРТ и другие);

- лабораторные (анализы крови и т.д.).

На данном этапе разработки в области обработки такого рода данных можно выделить следующие сложности и особенности:

- формат представления анализируемых данных может быть количественным, качественным, а в отдельных случаях и смешанным, вследствие чего возникает необходимость обеспечить верное восприятие информации машинными алгоритмами при обработке;

- данные, представленные для обработки, могут оказаться неполными и не содержать достаточной информации для установления точного диагноза пациента, следовательно, требуется оценить работу системы и предусмотреть алгоритмы её поведения в подобных случаях;

- анализируемая информация является узкоспециализированной и содержит значительное количество медицинской терминологии. С одной стороны это упрощает задачу поиска и выделения необходимых данных, с другой – при отсутствии термина в используемом системой словаре получение части необходимых данных может оказаться невозможным.

Всё это в значительной мере усложняет выделение из общего объёма анализируемых данных полезной информации и требует использования методов, позволяющих сделать эффективной работу разрабатываемой системы с учётом указанных особенностей.

Перейдём к рассмотрению принципов работы разрабатываемой системы.

Разрабатываемая система должна предусматривать выполнение следующих основных функций (рисунок 2):

- анализ данных пациента в виде блока текстовой информации;

- постановка предполагаемого диагноза на основе проведённого анализа и выявленных закономерностей;

- формирование и выдача врачебных рекомендаций для повышения эффективности лечения больного.

При поступлении на вход системы информации в виде сканированных текстовых документов система при помощи методов распознавания текста преобразует информацию в ряд текстовых блоков, после чего они подвергаются интеллектуальной машинной обработке, где из блоков извлекается

информация, необходимая для постановки первичного диагноза больного. Постановка диагноза производится при помощи использования методов интеллектуального анализа данных с опорой на продукционную базу знаний. Далее на основе установленного диагноза вырабатываются решения по рекомендуемому лечению пациента, при этом используется база знаний прецедентов. Конечной стадией работы системы является формирование наглядного отчёта, включающего наглядные данные о поставленном диагнозе и рекомендуемую методику лечения. При этом следует отметить, что система не является полностью автоматизированной: на стадиях постановки первичного диагноза и выработки решений по лечению предполагается участие лица, принимающего окончательное решение. Это позволит снизить риск неверного диагностирования при пограничных или комплексных заболеваниях, а также провести дополнительное обучение нейросетей, используемых в системе, что постепенно увеличит её эффективность.

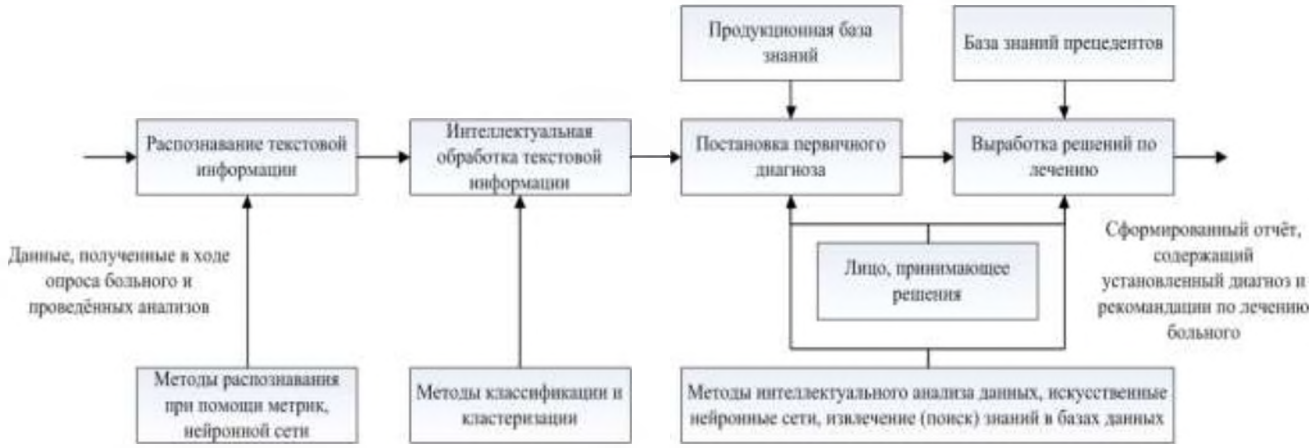

Рисунок 2 – Функционал разрабатываемой системы

В качестве методов распознавания текста предполагается использовать распознавание при помощи метрик, либо нейронных сетей. Метрика – некоторое условное значение функции, определяющее положение объекта в пространстве. Таким образом, если два объекта расположены близко друг от друга, то есть похожи (например, две буквы А написанные разным шрифтом), то метрики для таких объектов будут совпадать или быть предельно похожими [22]. Среди нейросетевых механизмов для распознавания символов можно использовать, к примеру, сеть Кохонена. Слой Кохонена состоит из адаптивных линейных сумматоров («линейных формальных нейронов»). Как правило, выходные сигналы слоя Кохонена обрабатываются по правилу «победитель забирает всё»: наибольший сигнал превращается в единичный, остальные обращаются в ноль [23]. Следует учитывать, что нейросеть данного типа необходимо обеспечить корректной обучающей выборкой, поскольку от неё зависит эффективность процедуры распознавания элементов.

В качестве методов интеллектуального анализа данных предлагается использовать алгоритмы классификации и кластеризации.

Под классификацией понимается один из разделов машинного обучения, посвящённый решению следующей задачи. Имеется множество объектов (ситуаций), разделённых некоторым образом на классы. Задано конечное множество объектов, для которых известно, к каким классам они относятся. Это множество называется обучающей выборкой. Классовая принадлежность остальных объектов не известна. Требуется построить алгоритм, способный классифицировать произвольный объект из исходного множества [24].

К алгоритмам классификации относятся:

- N-граммы;
- Tf*ldf (взвешивание);
- Признаки из матричных разложений;
- Тематическое моделирование;
- Дистрибутивная семантика.

Под кластеризацией понимают задачу разбиения заданной выборки объектов (ситуаций) на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались. Задача кластеризации относится к широкому классу задач обучения без учителя [25].

Среди алгоритмов кластеризации можно выделить следующие:

- Графовые алгоритмы кластеризации;

- K-Means;

- EM-алгоритм;

- DBScan;

- Иерархическая кластеризация;

- Тематическое моделирование;

- Ансамбль кластеризаторов;

- Нейронная сеть Кохонена.

Рекомендации по лечению пациентов вырабатываются за счёт использования базы знаний прецедентов и механизмов принятия решений. Выработка решений происходит в результате циклического процесса, в котором участвуют: система поддержки принятия решений в роли вычислительного звена и объекта управления; человек как управляющее звено, задающее входные данные и оценивающее полученный результат вычислений на компьютере. Её отличительные характеристики: ориентация на решение плохо структурированных задач; сочетание традиционных методов доступа и обработки компьютерных данных с возможностями математического моделирования; направленность на непрофессионального пользователя; высокая адаптивность – приспосабливаемость к особенностям используемого технического и программного обеспечения, требованиям пользователя.

К основным принципам формирования и использования механизмов принятия решений можно отнести: обеспечение логики принятия решений необходимой информацией в максимально возможном объеме; возможность оперативного поиска информации; генерирование альтернативных вариантов решений; предоставление прогнозных оценок результатов реализации возможных альтернатив; постоянная эволюция системы за счет наращивания ее возможностей.

Для анализа и выработок предложений в алгоритмах принятия решений используются разные методы. Это могут быть: информационный поиск, интеллектуальный анализ данных, поиск знаний в базах данных, рассуждение на основе прецедентов, имитационное моделирование, эволюционные вычисления и генетические алгоритмы, нейронные сети, ситуационный анализ, когнитивное моделирование и другие методы [26].

С учётом специфики разрабатываемой системы в ходе подготовительного этапа разработки был проведён анализ, в результате которого были выявлены наиболее приоритетные задачи, качественное решение которых окажет непосредственное влияние на результат разработки. К данным задачам следует отнести:

- Необходимость критического анализа типовой информации, предполагаемой для использования в разрабатываемой системе, с целью повышения эффективности выбора методов её обработки;

- Необходимость подбора и создания методов обработки информации, выраженной в виде неструктурированных русскоязычных текстов биомедицинской направленности;

- Необходимость разработки и использования методов тестирования, подтверждающих эффективность разрабатываемой системы.

3. Заключение

Подводя итоги, можно сказать, что реализация системы, обладающей перечнем указанных возможностей, позволит в значительной мере повысить качество медицинского обслуживания, вследствие снижения рисков влияния человеческого фактора за счёт использования машинной обработки информации, что в конечном итоге даст возможность провести цифровизацию медицинских организаций и повысить их экономическую эффективность.

Список используемых источников

1.Богдан Криват, Джеми Макленнен, Чжаохуэй Танг, Microsoft SQL Server 2008 :Datamining - интеллектуальный анализ данных, СПб. : БХВ-Петербург, 2009

2.А.А. Барсегян, И.И. Холод, М.Д.Тесс, М.С. Куприянов, С.И. Елизаров, Анализ данных и процессов, СПб.: БХВ-Петербург, 2009

3.«Лекция 1: Интеллектуальный анализ данных: базовые понятия», Источник: ИНТУИТ, URL: https://www.intuit.ru/studies/courses/2312/612/lecture/13260 (дата обращения: 18.03.2018)

4.«Основы интеллектуального анализа», URL: http://www.studmedlib.ru/ru/doc/ISBN9785970436899-0008.html (дата обращения: 18.03.2018)

5.Ларичев О. И., Петровский А. Б. Системы поддержки принятия решений. Современное состояние и перспективы их развития. // Итоги науки и техники. Сер. Техническая кибернетика. — Т.21. М.: ВИНИТИ, 1987, с. 131—164

6.«IBM Watson Health», URL: www.ibm.com/watson/health/oncology-and-genomics/oncology/ (дата обращения: 20.03.2018)

7.Статья «IBM Watson для онкологии: только факты», Источник: IBM. URL: evercare.ru/watson-facts (дата обращения: 20.03.2018)

8.Статья « Недостатки IBM Watson связаны с проблемами, от которых страдает все здравоохранение», URL: evercare.ru/stat-about-watson (дата обращения: 20.03.2018)

9.Официальный сайт программного продукта «STADIA», URL: protein.bio.msu.ru/~akula/Podr2~1.htm (дата обращения: 24.03.2018)

10.Официальный сайт программного продукта «SPSS», URL: www.ibm.com/analytics/ru/ru/technology/spss/ (дата обращения: 24.03.2018)

11.Официальный сайт программного продукта «STATA», URL: www.stata.com/ (дата обращения: 24.03.2018)

12.Официальный сайт программного продукта «STATISTICA», URL: statsoft.ru/ (дата обращения: 24.03.2018)

13.Официальный сайт программного продукта «JMP», URL: www.jmp.com/en_us/home.html (дата обращения: 24.03.2018)

14.Philips запускает платформу с искусственным интеллектом для здравоохранения, URL: http://www.iksmedia.ru/news/5486423-Philips-zapuskaet-platformu-s-iskus.html (дата обращения: 29.03.2018)

15.Официальный сайт программного продукта «ПроМед», URL: swanit.ru/elektronnoe_zdravoohranenie/riams_promed/ (дата обращения: 24.03.2018)

16. Официальный сайт программного продукта «MEDESK», URL: www.medesk.ru/ (дата обращения: 25.03.2018)

17. Официальный сайт программного продукта «MEDODS», URL: medods.ru/ (дата обращения: 25.03.2018)

18. Официальный сайт программного продукта «МедОфис», URL: medoffice.ru/ (дата обращения: 25.03.2018)

19. Официальный сайт программного продукта «MedWork», URL: www.medwork.ru/ (дата обращения: 25.03.2018)

20.«Минздрав перечислил самые распространённые болезни РФ», URL: https://regnum.ru/news/2312591.html (дата обращения: 20.03.2018)

21.«Виды бронхолёгочных заболеваний», URL: http://studbooks.net/2471639/meditsina/vidy_bronholegochnyh_zabolevaniy (дата обращения: 22.03.2018)

22.«Методы распознавания текста», URL: https://habr.com/post/220077 (дата обращения: 25.03.2018)

23.«Нейронная сеть Кохонена», URL: http://www.machinelearning.ru/wiki/index.php?title=Нейронная_сеть_Кохонена (дата обращения: 23.03.2018)

24.«Классификация», URL: http://www.machinelearning.ru/wiki/index.php?title=Классификация (дата обращения: 24.03.2018)

25.«Кластеризация», URL: http://www.machinelearning.ru/wiki/index.php?title=Кластеризация (дата обращения: 24.03.2018)

26.«Функции, основные характеристики и классификация СППР», URL: http://decision-make.ru/index.php?action=full_article&id=175 (дата обращения: 24.03.2018)