Исследование больших данных в сфере биоинформатики

Суворов С.В., профессор, кандидат экономических наук,
ФГБОУ ВО «Московский политехнический университет», Москва, Россия
Царькова Н.И., кандидат педагогических наук, доцент,
ФГБОУ ВО «Московский политехнический университет», Москва, Россия
Жиляева И.А., доцент, кандидат экономических наук,
ФГБОУ ВО «Московский политехнический университет», Москва, Россия
Кальгина К.М., магистрант, ФГБОУ ВО «Московский политехнический университет», Москва, Россия

Аннотация. В данной статье рассмотрена одна из сфер биоинформатики — онкологические заболевания и злокачественные новообразования, а также рассмотрены несколько методик bigdata анализа, применимых к имеющимся статистическим данным, собранным в России за период с 2008 по 2018 год.

Ключевые слова: онкология, большие данные, интеллектуальный анализ больших данных, искусственный интеллект, Deductor, Python 3.

Research of big data in the field of bioniformatics

Suvorov S.V., professor, candidate of Economic Sciences Moscow Polytechnic University, Moscow, Russia
Tsarkova N.I., candidate of Pedagogical Sciences, associate Professor Moscow Polytechnic University, Moscow, Russia
Zhilyaeva I.A., candidate of economic Sciences, associate Professor Moscow Polytechnic University, Moscow, Russia
Kalgina K.M., undergraduate Moscow Polytechnic University, Moscow, Russia

Annotation. This article discusses one of the bioinformatics areas — oncological diseases and malignant neoplasms, as well as several methods for

analyzing big data that are applicable to the available statistical data collected in Russia from 2008 to 2018.

Keywords: oncology, big data, big data mining, artificial intelligence, Deductor, Python 3.

Введение

За последние десять лет объемы данных и скорость их появления экспоненциально росли. Согласно опубликованным отчетам аналитической фирмы IDC, каждый день появляется более 3 квинтиллионов байтов данных. Общее количество информации ежегодно возрастает на 35%. Это является следствием развития информационных технологий, которые на сегодняшний день используются во всех сферах жизни общества. Медицинские организации собирают данные обследований пациентов, частные лаборатории собирают результаты проведения диагностик и исследований, приложения собирают статистику действий пользователей. Растущие объёмы информации ставят перед обществом новые сложные задачи по организации её хранения и обработки.

Биоинформатика — это наука, которая занимается анализом молекулярно-биологических данных. Это могут быть последовательности геномов, структуры белков, данные о том, как гены работают и с чем взаимодействуют. В процессе развития технологий в биоинформатике выделились различные направления. К этим направлениям относятся геномика, транскриптомика, протеомика и другие. Каждое из направлений имеет свои технологии получения данных и свои объекты для изучения, но все направления порождают огромные объемы данных, которые необходимо хранить, систематизировать, и извлекать из них полезную информацию. Одним из направлений биоинформатики является изучение онкологии. В рамках этого направления изучаются генетические основы развития раковых болезней, молекулярная динамика. Рак является второй из основных причин смерти в мире.[1]

[1] Oncology — официальный сайт компании онкологических заболеваний [Электронный ресурс]. Режим доступа: http://www.oncology.ru/service/statistics/malignant_tumors/2018.pdf, свободный. – (дата обращения: 24.12.2019).

Основная часть

В настоящий момент лидером медицинских исследований, базирующихся на анализе больших данных, является испанская компания «BigMedilytics». Вместе со своими партнерами из 12 стран они активно работают в трех направлениях: поддержание здоровья населения, диагностика хронических заболеваний и борьба с онкологическими заболеваниями. Согласно их исследованию, произведенному вместе с российскими коллегами и национальный медицинский исследовательский центром, за 2018 наиболее распространёнными пораженными органами у мужчин выступают: бронхи, легкие, предстательная железа и кожа (рисунок 1), а у женщин: молочные железы и кожа (рисунок 2).[2]

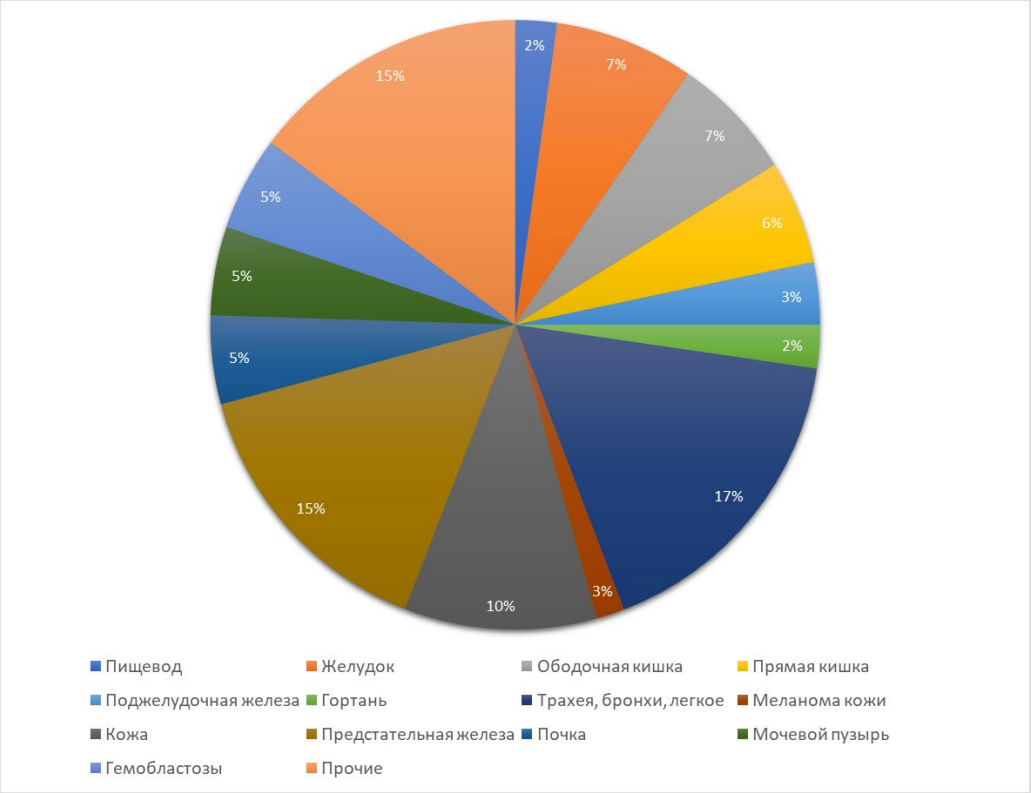

Рис. 1 – Структура заболеваемости злокачественными новообразованиями мужского населения России в 2018 год

[2] UECS — электронный научный журнал управления экономическими системами [Электронный ресурс]. Режим доступа: http://uecs.ru/index.php?option=com_flexicontent&view=items&id=5595e, свободный. – (дата обращения: 24.12.2019).

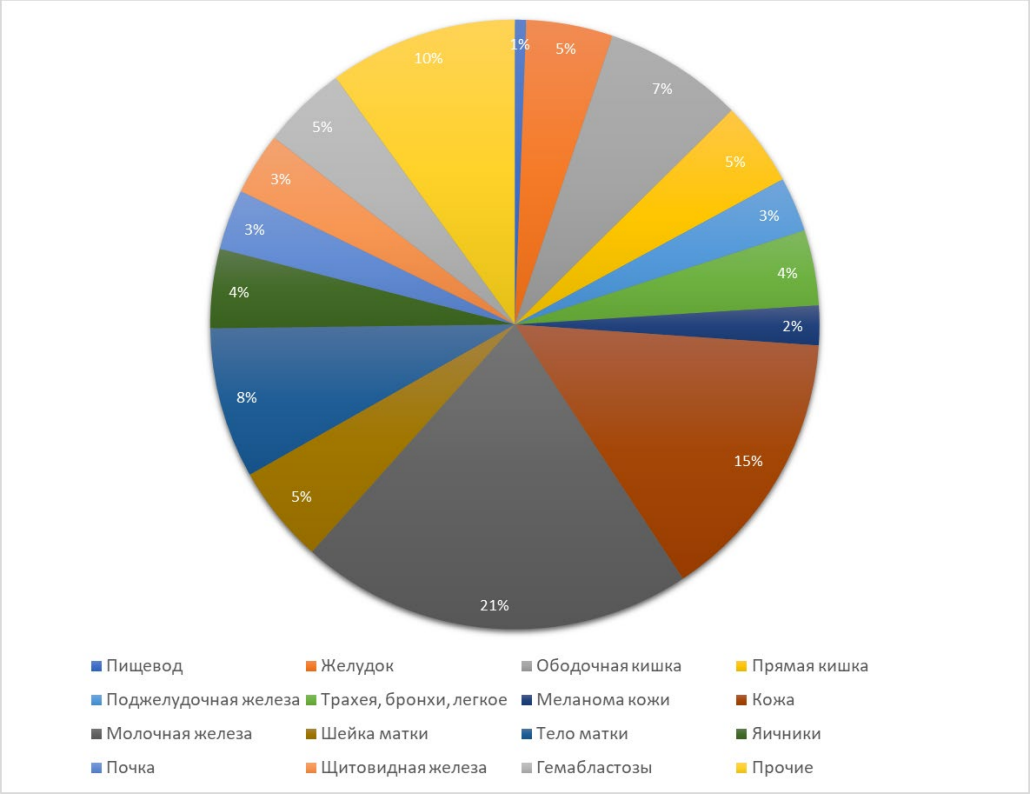

Рис. 2 – Структура заболеваемости злокачественными новообразованиями женского населения России в 2018 год

Согласно проведенному исследованию, средний возраст заболевших в 2018 году в России составил 64,5 года. При этом средний возраст для мужчин составлял 64,9 года, а для женщин 64,2 года. При сравнении возрастных показателей с данными за 2008 год, была выявлена тенденция увеличения среднего возраста заболевших. Так, в 2008 году совокупный возраст составлял 63,5 года, в детальном рассмотрении для мужчин он был равен 63,8 года, для женщин 63,3 года.[3]

Различия среднего возраста между заболевшими женщинами и мужчинами особенно велики при новообразованиях губы (8,1 года), пищевода (5,6 года), поджелудочной железы (5,1 года), головного мозга (4,0 года), лимфатической и кроветворной ткани (4,0 года), желудка (3,1 года), почки (3,1 года), мочевого пузыря (2,8 года), трахеи, бронхов, легкого (2,0 года), кожи (без меланомы) (2,2 года).

[3] Википедия — свободная энциклопедия [Электронный ресурс]. Режим доступа: https://ru.wikipedia.org/wiki/онкология, свободный. – (дата обращения: 24.12.2019).

Максимальный уровень совокупной онкологической заболеваемости в популяции России отмечается в возрастной группе 75-79 лет (1647,0 на 100 000 населения соответствующего возраста). Показатель детской (0-14 лет) заболеваемости злокачественными новообразованиям составил в 2018 г. 12,7 на 100 000 детского населения. При этом мальчики заболевают в 1,2 раза чаще девочек. Соотношения показателей заболеваемости мужского и женского населения различаются в разных возрастных группах: 15-29 лет – 0,6; 30-39 лет – 0,4; 40-49 лет – 0,6; 50-59 лет – 1,0; 60-69 лет – 1,6; 70-79 лет – 1,9; 80 лет и старше – 1,7.

Всего за 2018 год было выявлено 624 699 случаев злокачественных новообразований у ранее не болевших пациентов. На мужчин приходится 285 949 случаев, а на женщин 338 750, что в процентном соотношении составляет 45,77% для мужчин и 54,23% для женщин. Разница между количеством мужчин и женщин составляет 52 801 человек (8,46%). Прирост количества заболевших по сравнению с 2017 годом составил 1,2%. Совокупный показатель распространенности онкологических заболеваний составил 2 562 человека на 100 000 человек.

При распределении количества заболевших по территориальному признаку, максимальные показатели суммарной онкологическойзаболеваемости (при рассмотрении количества заболевших на 100 000 человек населения) отмечены в Архангельской (537,0), Пензенской (536,6) областях, Алтайском крае (532,4), Рязанской (526,5), Курской (525,0), Ярославской (523,9) областях. Минимальные показатели заболевших отмечены в республиках Чечня (155,3), Дагестан (164,2), Ингушетия (175,2), Ямало-Ненецком автономном округе(217,6), Республике Тыва (243,0). Более детально территориально распределение заболевших визуализировано на карте (рисунок 3). [4]

[4] Vesty — новостной портал [Электронный ресурс]. Режим доступа: https://nauka.vesti.ru/article/1041960, свободный. – (дата обращения: 24.12.2019).

Рис. 3 – Территориально распределение онкобольных за 2018 год

На основе проведённого анализа можно сделать вывод, что существует некий дисбаланс в количестве онкобольных как по территориальному и локализационному (тип раковой опухоли) признаку, так и по возрастному и половому признаку. Более детальное изучение и систематизация данных может помочь выявить другие закономерности и причин следственные связи появления злокачественных новообразований.

По своему количеству и разнообразию факторов, данные по онкобольным, собранные за период с 2008 по 2018 год, в совокупности представляют собой достаточно большой датасет, включающий в себя такие показатели, как:

- Год
- Пол
- Возраст
- Локализация опухоли
- Территориальная принадлежность
- Абсолютное число заболевших
- Показатели на 100 000 населения
  - «Грубый» показатель

- Стандартизированный показатель
  - Ошибка стандартизированного показателя
- Код МКБ 10
- Среднегодовой темп прироста заболевших
- Прирост заболевших

Чтобы корректно обработать такого рода данные и сделать достоверные выводы, следует применить к ним технологию интеллектуального анализа данных (DataMining). Учитывая, что данные собраны за период с 2008 по 2018 год (10 лет), появляется возможность составить прогноз на дальнейшие года. Уместно в данном случае применить машинное обучение и нейросети на базе платформы для создания законченных аналитических решений – Deductor.

Таким образом, следуя алгоритму интеллектуального анализа данных, первым шагом будет являться очистка и подготовка набора данных для дальнейшего анализа. Можно выделить важный под этап процесса – оценивание качества данных. Качество данных – это критерий, определяющий полноту, точность, своевременность и возможность интерпретации данных. Помимо этого, следует проверить данные на наличие шумов и выбросов. Различные методы DataMining имеют разную чувствительность к выбросам, этот факт нужно учитывать при выборе метода анализа данных. Также некоторые инструменты DataMining имеют встроенные процедуры очистки от шумов и выбросов. Процесс очистка данных по своей сути занимается выявлением и удалением ошибок и несоответствий в данных и направлен на улучшение качества данных. В целом, очистка данных включает следующие этапы:

1. Анализ данных.
2. Определение порядка и правил преобразования данных.
3. Подтверждение.
4. Преобразования.
5. Противоток очищенных данных.

После окончания этапа подготовки данных можно переходить к построению модели. Для построения моделей используются различные методы

и алгоритмы DataMining. Некоторые задачи могут быть решены при помощи моделей, построенных на основе различных методов. К сожалению, идеальной модели, которая бы позволила решать разнообразные задачи, не существует. Конкретно к имеющимся данным применим метод нейросетевого прогнозирования.

Нейронные сети — это система соединенных и взаимодействующих между собой простых процессоров (искусственных нейронов). К главным областям применения нейронных сетей можно отнести такие процессы, как прогнозирование, автоматизация процессов распознавания образов, адаптивное управление, создание экспертных систем, организация ассоциативной памяти, обработка аналоговых и цифровых сигналов, синтез и идентификация электронных цепей и систем. [5]

Перед использованием нейронной сети необходимо обучить ее. Процесс обучения нейронной сети заключается в подстройке ее внутренних параметров под конкретную задачу. Алгоритм работы нейронной сети итеративный, его шаги называют эпохами или циклами.

Эпоха – одна итерация в процессе обучения, включающая предъявление всех примеров из обучающего множества и, возможно, проверку качества обучения на контрольном множестве. Процесс обучения осуществляется на обучающей выборке. Обучающая выборка включает входные значения и соответствующие им выходные значения набора данных. В ходе обучения нейронная сеть находит некие зависимости выходных полей от входных. Процесс обучения нейросети можно производить как через программу Deductor, так и применив к имеющимся данным такие библиотеки Python 3, как: SciKit-Learn, Theano, Keras[6].

Завершающим этапом, после обучения нейросети является ее проверка и, при наличии нескольких вариантов, выбор наиболее оптимальной модели.

[5] Царькова Н.И., Ерисов В.Д., Пекова Е.А. Технология BigData как инструмент управления в межкультурной коммуникации // Управление экономическими системами: электронный научный журнал, 2019. № 7.
[6] Суворов С.В., Царькова Н.И., Спиридонова А.К. Анализ больших данных компании UberTechnologiesInc с помощью технологии DataMining // Управление экономическими системами: электронный научный журнал, 2019. № 7.

Следует отметить, что со временем данные будет необходимо обновлять и дополнять, чтоб сохранять актуальность выводов.

Заключение

Рассмотренная проблема является актуальной и злободневной как для России, так и для всего мира в целом. Проведенное в дальнейшем исследование поможет совершить качественное преобразование системы оказания первичной и специализированной, в том числе высокотехнологичной, медицинской помощи; усиление роли первичного медико-санитарного звена для раннего выявления злокачественных новообразований и повышение профессионализма медицинских работников.

В качестве выходных данных будут выступать рассчитанные и спрогнозированные значения количественного изменения онкобольных мужчин и женщин, с территориальным и возрастным распределением и указанием очага заболевания. Полученные результаты будут проанализированы и визуализированы.

Библиографический список

1. Oncology — официальный сайт компании онкологических заболеваний [Электронный ресурс]. Режим доступа: http://www.oncology.ru/service/statistics/malignant_tumors/2018.pdf, свободный. – (дата обращения: 24.12.2019).

2. UECS — электронный научный журнал управления экономическими системами [Электронный ресурс]. Режим доступа: http://uecs.ru/index.php?option=com_flexicontent&view=items&id=5595e, свободный. – (дата обращения: 24.12.2019).

3. Википедия — свободная энциклопедия [Электронный ресурс]. Режим доступа: https://ru.wikipedia.org/wiki/онкология, свободный. – (дата обращения: 24.12.2019).

4. Vesty — новостной портал [Электронный ресурс]. Режим доступа: https://nauka.vesti.ru/article/1041960, свободный. – (дата обращения: 24.12.2019).

5. Царькова Н.И., Ерисов В.Д., Пекова Е.А. Технология BigData как инструмент управления в межкультурной коммуникации // Управление экономическими системами: электронный научный журнал. – 2019. – № 7.

6. Суворов С.В., Царькова Н.И., Спиридонова А.К. Анализ больших данных компании UberTechnologiesInc с помощью технологии DataMining// Управление экономическими системами: электронный научный журнал. – 2019. – № 7.

Bibliographic list

1. Oncology — the official website of the cancer company [Electronic resource]. Access mode: http://www.oncology.ru/service/statistics/malignant_tumors/2018.pdf, free - (accessed date: 12/24/2019).

2. UECS — electronic scientific journal of economic systems management [Electronic resource]. Access mode: http://uecs.ru/index.php?option=com_flexicontent&view=items&id=5595e, free. – (accessed: 12/24/2019).

3. Wikipedia — the free encyclopedia [Electronic resource]. Access mode: https://ru.wikipedia.org/wiki/онкология, free. – (accessed date: 12/24/2019).

4. Vesty— news portal [Electronic resource]. Access mode:https://nauka.vesti.ru/article/1041960, free. – (accessed: 12/24/2019).

5. Tsarkova N.I., Erisov V.D., Pekova E.A. Big Data technology as a management tool in inter-cultural communication // Management of economic systems: electronic scientific journal. – 2019. – № 7.

6. Suvorov S.V., Tsarkova N.I., Spiridonova A.K. Big Data Analysis by Uber Technologies Inc Using Data Mining Technologies // Management of Economic Systems: Electronic Scientific Journal. – 2019. – № 7.