УДК 004.522

РОЛЬ БОЛЬШИХ ДАННЫХ В АНАЛИЗЕ ГЕНОМНЫХ ДАННЫХ

О.М. Аббасова
Преподаватель кафедры «Программное обеспечение информационных технологий»
Институт Телекоммуникаций и информатики Туркменистана
ogultaganabbasowa@gmail.com

Г.Г. Аннасапаров
Преподаватель кафедры «Программное обеспечение информационных технологий»
Институт Телекоммуникаций и информатики Туркменистана
guwancannasaparow187@gmail.com

О.М. Аббасова

Окончил институт Телекоммуникаций и информатики Туркменистана. Область научных интересов связана с програмированием и использованием искусственного интеллекта.

Г.Г. Аннасапаров

Окончил Туркменский государственный университет имени Махтумкули. Область научных интересов связана с математикой и программированием.

Аннотация. Системы здравоохранения, генетика и геномика, население и общественное здравоохранение; все области биомедицины могут выиграть от Больших данных и связанных с ними технологий. При анализе геномных данных часто требуется интеграция различных типов данных (например, геномные, клинические, экологические). Поэтому важно владеть языками и инструментами, которые позволяют эффективно обрабатывать и интегрировать разнородные данные. Машинное обучение играет все более важную роль в анализе геномных данных. Python с его библиотеками машинного обучения является отличным выбором для этой задачи.

Ключевые слова: Медицинская информатика, интеллектуальный анализ данных, интеллектуальный анализ текстов, информационные системы, хранение и поиск информации, bioPython, Python.

Введение. Big Data — это горячая тема в здравоохранении и биомедицинских исследованиях. Более частое использование термина «Big Data» в биомедицинской литературе свидетельствует о растущей важности крупномасштабных наборов данных в здравоохранении и биомедицине, а также растет понимание роли, которую большие данные могут играть в научных и клинических исследованиях. Также было отмечено, что термин «Big Data» может означать разные вещи для разных групп людей. Однако было общее признание того, что здравоохранение, биомедицинские исследования и здоровье населения генерируют огромные, сложные, распределенные и часто динамические наборы данных, и что размер и сложность этих данных будут создавать как проблемы, так и возможности для организаций здравоохранения.

Считается, что термин «большие данные» возник в поисковых компаниях, которым приходилось запрашивать очень большие распределенные агрегации слабо структурированных данных.

С тех пор этот термин используется для обозначения огромных объемов данных, собранных с течением времени, которые трудно анализировать и обрабатывать с помощью обычных инструментов управления базами данных. Хотя может показаться, что этот

термин относится к объему данных, это не всегда так. В исследовательском отчете 2001 года аналитик Gartner (ранее META Group) Дуг Лэйни определил проблемы и возможности роста данных как трехмерные, т. е. увеличивающийся объем (количество данных), скорость (скорость входящих и исходящих данных) и разнообразие (диапазон типов данных и источников). Большая часть отрасли продолжает использовать эту модель «3 Vs» для описания больших данных. В 2012 году Gartner обновил свое определение следующим образом: «Большие данные — это информационные активы большого объема, высокой скорости и/или большого разнообразия, которые требуют новых форм обработки для обеспечения улучшенного принятия решений, обнаружения информации и оптимизации процессов». Кроме того, некоторые организации добавили новую букву V для «Veracity» для описания. Термин «Большие данные» может также относиться к технологии (например, хранилищам, инструментам и процессам), которые требуются организации для обработки больших объемов данных. Более прагматичное определение определяет Большие данные с точки зрения требования к аналитическим приложениям для обработки новых типов данных. С момента публикации первой последовательности генома человека в 2003 году область геномики стала основной движущей силой генерации больших данных в биомедицине. Прогресс в лабораторных аналитических методах (например, анализ последовательности ДНК) и мобильных технологиях (например, данные с мониторов физической активности и приложений) в настоящее время в значительной степени ответственны за постоянно растущее производство данных в реальном времени в больших объемах.

Однако использование больших данных теперь достигло всех областей здравоохранения, биомедицинских исследований и здоровья населения. Исследователи служб здравоохранения могут объединять административные и клинические базы данных для разработки прогностических моделей с целью улучшения политики здравоохранения. Фармацевтическая промышленность управляет огромными хранилищами клинических и молекулярных данных для рационального дизайна лекарств и подходов фармакогеномики. В области здоровья населения регистры заболеваний и данные из клинических записей используются для измерения воздействия вмешательств в здравоохранение. Биомедицинские исследователи имеют доступ к новым источникам геномных данных (например, микробиом, эпигеномика) и могут изучать новые гипотезы для понимания молекулярных причин заболеваний. Наконец, экологические данные начинают интегрироваться с генетическими и клиническими данными, которые (организация) ранее не отслеживала.

Хотя формального определения нет, обычно термин «структурированные данные» относится к данным с определенной схемой или моделью данных (т. е. явной семантикой). Данные, хранящиеся в базе данных, обычно структурированы. Измерения и сигналы являются примерами структурированных данных. Напротив, неструктурированные данные относятся к данным, содержащим информацию, которая нелегко доступна для систем управления вычислительными данными — содержащаяся в них информация не представлена в форме с четкой схемой данных, которая позволяет проводить прямую вычислительную интерпретацию и анализ. Этот тип данных обычно требует специализированных аналитических методов для извлечения содержащейся в них информации и преобразования ее в вычислимую форму. Текст на естественном языке, изображения и аудиопотоки являются примерами неструктурированных данных.

Существует огромное количество информации, относящейся к пониманию здоровья человека, зафиксированной в неструктурированных ресурсах. Тексты, изображения, аудио- и видеопотоки обычно производятся в клиническом контексте. Эти ресурсы требуют разработки стратегий для извлечения и обобщения содержащейся в них информации, чтобы наложить структуру и значение, используя внутреннюю структуру или закономерности, присущие источнику. Технологические решения, позволяющие автоматически

интерпретировать такие ресурсы, в дополнение к помощи людям, которым поручено интерпретировать эти источники данных, позволяют масштабировать анализ и рассматривать гораздо более широкий набор данных — по больнице, населению или научно-исследовательскому сообществу. Это привело к некоторым весьма инновационным исследованиям, демонстрирующим силу крупномасштабного анализа данных в медицине.

Анализ геномных данных — это процесс изучения и интерпретации генетической информации, закодированной в ДНК организма. Он включает в себя использование различных методов и инструментов для выявления закономерностей, вариаций и мутаций в геноме, а также для определения их влияния на биологические процессы, здоровье и болезни.

Геномные данные относятся к полной генетической информации организма, хранящейся в его ДНК. Эта информация включает в себя последовательность оснований ДНК, гены, регуляторные элементы и другие функциональные элементы.

Проблемы анализа геномных данных. Анализ геномных данных представляет собой уникальные проблемы из-за огромного объема данных, сложности генома и необходимости интеграции данных из различных источников.

Большие данные — это термин, используемый для описания огромных объемов данных, которые настолько сложны, что традиционные методы обработки данных не могут их эффективно обрабатывать. Большие данные предлагают несколько преимуществ для анализа геномных данных:

Масштабируемость: системы больших данных могут обрабатывать и анализировать огромные объемы геномных данных, что позволяет исследователям изучать большие популяции и выявлять редкие генетические вариации.

Скорость: технологии больших данных могут анализировать геномные данные гораздо быстрее, чем традиционные методы, что позволяет исследователям быстрее получать результаты.

Интеграция данных: платформы больших данных могут интегрировать геномные данные с другими типами данных, такими как клинические данные и данные о стиле жизни, для получения более полного представления о здоровье человека.

Машинное обучение: алгоритмы машинного обучения могут анализировать геномные данные для выявления закономерностей и взаимосвязей, которые трудно обнаружить с помощью традиционных статистических методов.

Для анализа геномных данных с использованием технологий больших данных применяется широкий спектр языков программирования и инструментов. Выбор конкретного языка часто зависит от специфики задачи, предпочтений аналитика и доступных ресурсов. Вот наиболее популярные и востребованные языки:

Основные языки программирования:

Python:

– широко используется в биоинформатике и анализе данных благодаря богатому набору библиотек (например, NumPy, Pandas, SciPy, scikit-learn);

– идеален для статистического анализа, машинного обучения и визуализации данных;

– библиотеки, такие как Biopython, специально разработаны для работы с биологическими данными.

R:

– специализированный язык для статистического анализа и визуализации данных;

– имеет обширный набор пакетов для биоинформатики и геномного анализа (например, Bioconductor);

– часто используется для анализа экспрессии генов, генетической вариации и других геномных данных.

Java:

– используется для разработки высокопроизводительных приложений и систем обработки больших данных;

– применяется в разработке биоинформатических инструментов и платформ;

– используется для таких технологий как Apache Hadoop.

C/C++:

– обеспечивают высокую производительность и используются для разработки низкоуровневых биоинформатических инструментов и алгоритмов;

– используются, когда критична скорость выполнения.

Python — мощный инструмент для анализа геномных данных, благодаря своему широкому набору библиотек и простоте использования.

Заключение. Анализ геномных данных на Python открывает широкие возможности для исследований в области биоинформатики, медицины и других наук. Благодаря мощным библиотекам, таким как Biopython, Pandas и Scikit-learn, Python позволяет эффективно обрабатывать, анализировать и визуализировать сложные геномные данные. Мы рассмотрели основные шаги анализа, начиная от загрузки данных из распространенных форматов, таких как FASTA и VCF, до анализа последовательностей и генетических вариаций. Визуализация данных с помощью Matplotlib и Seaborn помогает лучше понять результаты анализа, а применение алгоритмов машинного обучения позволяет выявлять скрытые закономерности и делать прогнозы. Однако важно помнить, что анализ геномных данных требует не только технических навыков, но и глубокого понимания биологических процессов. Интерпретация результатов должна проводиться с учетом контекста и с использованием соответствующих биологических знаний.

Список литературы

[1] PC Magazine Encyclopedia. http://www.pcmag.com/encyclopedia/term/62849/big-data.

[2] O'Driscoll A, Daugelaite J, Sleator RD. 'Big data', Hadoop and cloud computing in genomics. J Biomed Inform 2013

THE ROLE OF BIG DATA IN GENOMIC DATA ANALYSIS

O. Abbasova
Lecturer of the Department of Software of information technologies,
The Institute of Telecommunications and Informatics of Turkmenistan

G. Annasaparov
Lecturer of the Department of Software of information technologies,
The Institute of Telecommunications and Informatics of Turkmenistan

Annotation. Health systems, genetics and genomics, population and public health; all areas of biomedicine can benefit from Big Data and related technologies. Genomic data analysis often requires the integration of different types of data (e.g. genomic, clinical, environmental). It is therefore important to have languages and tools that allow efficient processing and integration of heterogeneous data. Machine learning is playing an increasingly important role in genomic data analysis. Python with its machine learning libraries is an excellent choice for this task..

Keywords: Medical informatics, data mining, text mining, information systems, information storage and retrieval, bioPython, Python.