МЕТОДЫ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА БОЛЬШИХ ДАННЫХ ФЛУОРЕСЦЕНТНОЙ СПЕКТРОСКОПИИ ДЛЯ ИССЛЕДОВАНИЯ СЛОЖНЫХ БИОФИЗИЧЕСКИХ СИСТЕМ

Н. Н. Яцков, А. А. Горбунова, В. В. Апанасович

Белорусский государственный университет, Минск, Беларусь
E-mail: yatskou@bsu.by

Предложена схема выбора и применения методов интеллектуального анализа больших данных флуоресцентной спектроскопии с целью исследования сложных биофизических систем. Выполнен сравнительный анализ алгоритмов на смоделированных и экспериментальных данных, представляющих собой примеры модельных систем сбора и обработки больших данных при исследовании биомолекулярных соединений. Произведена оценка эффективности алгоритмов.

Ключевые слова: *биофизические системы, анализ больших данных флуоресцентной спектроскопии, методы интеллектуального анализа данных*.

ВВЕДЕНИЕ

Появление методов анализа больших данных связано с развитием информационных технологий, увеличением производительности экспериментальных методов, удешевлением систем сбора и хранения данных [1, 2]. Традиционно алгоритмы интеллектуального анализа данных (ИАД) применяются для решения стандартных задач принятия решений на основе моделей «черного ящика», таких как снижение размерности данных, классификация, кластеризация, визуализация [3, 4]. Предложен системный или комплексный подход к анализу больших данных прикладной флуоресцентной спектроскопии, основанный на применении алгоритмов ИАД и имитационного моделирования, который позволяет выявить наиболее информативные данные и проанализировать их с физическими моделями «белого ящика» [5, 6]. Критически важной задачей является систематизация и развитие части подхода, связанной с отбором, выделением и обработкой наиболее информативных данных, оценкой параметров моделей «белого ящика» и качества анализа. Для решения данной задачи применяются алгоритмы ИАД, что требует их исследования, обоснованного выбора и усовершенствования при применении в ходе анализа больших данных флуоресцентной спектроскопии [7, 8]. Для использования методов ИАД в комплексном подходе при обработке данных флуоресцентной спектроскопии накладывается ряд требований или ограничений, которые могут выступать в виде критериев эффективности применения, связанных с особенностями больших данных, имитационных моделей или предметной областью, например – точность оценива-

ния параметров моделей, скорость обработки данных и минимальность использования вычислительных ресурсов, критически необходимых при проведении экспериментальных исследований в стандартных условиях. Вышеизложенное предполагает проведение всестороннего сравнительного анализа и исследования методов ИАД с учетом предъявляемых к ним требований, применяемых при обработке биофизических систем в ходе комплексного анализа.

Целью работы является разработка схемы выбора и применения методов ИАД для эффективного использования в комплексном анализе больших данных прикладной флуоресцентной спектроскопии.

СХЕМА ПРИМЕНЕИЯ МЕТОДОВ ИАД

Методы ИАД используются для повышения точности, эффективности и устойчивости обработки данных, а также для упрощения применения, обеспечения надежности и автоматизации комплексного подхода. Методы можно разделить на группы, связанные с конкретными задачами анализа данных, такими как снижения размерности, кластерного анализа (КА), классификации, регрессии, оптимизации и визуализации данных.

В работе предложена схема применения методов ИАД в комплексном подходе (рисунок 1) с учетом задач анализа, требований или ограничений, связанных с особенностями больших данных, имитационных моделей и предметной областью флуоресцентной спектроскопии. Сформулированы критерии оптимальности методов с целью выбора наиболее эффективных для решения задач комплексного анализа биофизической информации. Рассмотрены наиболее популярные алгоритмы ИАД.

РЕЗУЛЬТАТЫ

Выполнен сравнительный анализ алгоритмов ИАД на смоделированных и экспериментальных данных, представляющих собой примеры модельных систем сбора и обработки больших данных при исследовании сложных биомолекулярных соединений. Определены наиболее эффективные алгоритмы для решения задач анализа больших данных молекулярных и клеточных систем с учетом предъявляемых к ним требований (таблица 1).

ЗАКЛЮЧЕНИЕ

Предложена схема выбора и применения методов анализа систем области флуоресцентной спектроскопии для повышения точности, эффективности и устойчивости обработки данных в комплексном подходе на основе имитационного моделирования. Представлены результаты иссле-

дования наиболее оптимальных методов интеллектуального анализа данных при сжатии и разделении кластеров данных, определении физических факторов, выделении наиболее информативных признаков и кластеров данных, классификации наблюдений, оценке параметров имитационных моделей и их доверительных интервалов, охарактеризовании качества анализа, поиске ассоциативных правил.

Рис. 1. Схема применения методов ИАД в комплексном подходе

Таблица 1

Методы ИАД для использования в комплексном подходе при обработке больших данных флуоресцентной спектроскопии

Задача	Особенности	Метод
Снижение размерности	Визуализация	UMAP
	Сжатие	МГК
	Физические компоненты	МНК
	Информативные признаки	Алгоритм на основе критерия

		Фишера
Задача	**Особенности**	**Метод**
Кластерный анализ	Небольшие данные Большие данные	Иерархический DBSCAN
Классификация	Универсальность	Случайный лес, xgboost
Оптимизация	Многопараметрическая, в условиях высокого шума	Нелдера-Мида
Критерий качества анализа	Гистограммы экспериментальных данных	Хи-квадрат-подобный
Доверительные интервалы	Гистограммы экспериментальных данных	Асимптотических стандартных ошибок
Нейросетевая аппроксимация имитационных моделей	Применена в исследовании резонансного переноса энергии между окрашенными белками	Многослойный персептрон
Поиск ассоциативных правил	По результатам исследования транскриптов онкогена	FPG

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Яцков Н. Н., Апанасович В. В. Комплексный анализ данных при исследовании сложных биомолекулярных систем // Информатика. 2021. Т. 18, № 1. С. 105-122.
2. Yu S., Guo S. Big Data Concepts, Theories, and Applications / New York: Springer, 2016. 445 p.
3. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. Data Mining, Inference, and Prediction / 2nd ed. New York : Springer, 2009. 739 p.
4. Яцков Н. Н. Интеллектуальный анализ данных : пособие / Минск: БГУ, 2014. 151 с.
5. Яцков Н. Н., Скакун В. В., Апанасович В. В. Метод обработки кинетических кривых затухания флуоресценции с использованием алгоритмов интеллектуального анализа данных // Журнал прикладной спектроскопии. 2020. Т. 87, № 4. С. 311-322.
6. Яцков Н. Н., Скакун В. В., Недервин-Шипперс Л. И др. Комплексный анализ флуктуаций интенсивности флуоресценции молекулярных соединений // Журнал прикладной спектроскопии. 2020. Т. 87, № 4. – С. 628-636.
7. Datta R., Heaster T.M., Sharick J.T. et al. Fluorescence lifetime imaging microscopy : fundamentals and advances in instrumentation, analysis, and applications // J. Biomed. Opt. 2020. Vol. 25(7):071203. DOI: 10.1117/1.JBO.25.7.071203
8. Datta R., Gillette A., Stefely M., Skala M.C. Recent innovations in fluorescence lifetime imaging microscopy forbiology and medicine // J. Biomed. Opt. 2021. Vol. 26(7):070603. DOI: 10.1117/1.JBO.26.7.070603