РАЗРАБОТКА И ПРОГРАММНАЯ РЕАЛИЗАЦИЯ АЛГОРИТМОВ СНИЖЕНИЯ РАЗМЕРНОСТИ ДАННЫХ БИОФИЗИЧЕСКИХ ЭКСПЕРИМЕНТОВ

А. А. Горбунова

Белорусский государственный университет, г. Минск;
anastasia.gorbunova.so@yandex.ru;
науч. рук. – Н. Н. Яцков, канд. физ.-мат. наук, доц.

В работе представлены результаты сравнительного анализа алгоритмов методов главных и независимых компонент, стохастического вложения соседей с t-распределением, равномерного приближения и проекции, многомерного шкалирования, неотрицательного матричного разложения для классификации данных об экспрессии генов в заболевании рака груди. Сравнительный анализ реализованных методов выполнен на данных о метилировании ДНК, экспрессии микроРНК и информационной РНК молекулах, представляющих кластеры различной сложности. Наилучшим алгоритмом является метод равномерного приближения и проекции, точность классификации которого на данных о метилировании ДНК – 73%, экспрессии микроРНК – 69%, экспрессии информационной РНК – 79%.

Ключевые слова: алгоритмы снижения размерности данных; экспрессия генов; классификация; метилирование ДНК; микроРНК; информационная РНК; критерии качества анализа.

ВВЕДЕНИЕ

Высокопроизводительные геномные технологии становятся стандартом для молекулярной диагностики рака [1]. Геномные секвенаторы нового поколения позволяют регистрировать большие наборы экспериментальных данных о нуклеотидном составе молекул ДНК/РНК, что затрудняет визуализацию, интерпретацию и понимание результатов. Алгоритмы снижения размерности данных частично или полностью устраняют это ограничение, проецируя данные в пространство нескольких измерений с сохранением важных свойств объектов [2].

Среди существующих алгоритмов снижения размерности данных следует выделить наиболее перспективные методы, такие как метод главных компонент (далее используется аббревиатура PCA от англ. principal component analysis), метод независимых компонент (ICA от англ. independent component analysis), метод стохастического вложения соседей с t-распределением (tSNE от англ. t-distributed stochastic neighbor embedding), метод равномерного приближения и проекции (UMAP от англ. uniform approximation and projection), многомерное шкалирование (MDS от англ. multidimensional scaling), метод неотрицательного матричного разложения (NMF от англ. non-negative matrix factorization) [1, 3].

Цель работы – сравнить эффективность наиболее популярных алгоритмов снижения размерности данных в применении к решению задачи группировки данных эпигеномного метилирования ДНК (meDNA от англ. DNA methylation), микроРНК (miRNA от англ. Micro RNA) и информационной РНК (mRNA от англ. messenger RNA), получаемых в экспериментах геномного секвенирования по исследованию рака груди (BRCA от англ. Breast Cancer).

АЛГОРИТМЫ СНИЖЕНИЯ РАЗМЕРНОСТИ ДАННЫХ

Рассмотрим кратко методы снижения размерности данных. Основная идея PCA – уменьшение размерности набора данных, сохраняя переменные с набольшей вариаций. ICA преобразует исходные признаки наблюдений объекта исследования в независимые подкомпоненты. tSNE, нелинейный метод уменьшения размерности, моделирует новый набор данных таким образом, что подобные объекты отображаются ближайшими точками в сниженном пространстве. UMAP – метод визуализации и нелинейного снижения размерности, особенностью которого является построение взвешенного неориентированного графа. MDS создает карту, отображающую относительные положения ряда объектов. На карте преобразуются расстояния между объектами. NMF – линейный алгоритм, суть которого состоит в разложении исходной матрицы данных на две простые, упрощающие дальнейшую обработку, при условии, что матрицы не будут иметь отрицательных элементов [4].

ЭКСПЕРИМЕНТАЛЬНЫЕ ДАННЫЕ

Исследованы данные опухолей рака груди, представленные наборами эпигеномного метилирования ДНК, микроРНК и информационной РНК для различных групп пациентов. Взято три набора данных, поскольку каждый из них представляет практическую значимость при диагностике рака груди. Наборы данных различаются по числу генов и пациентов, степени зашумленности, что позволяет всесторонне исследовать эффективность алгоритмов. Наборы данных содержат пять подтипов рака груди BRCA: LumA, LumB, Basal, Her2 и Normal (эталонная группа здоровых людей) (таблица 1). Подтип заболевания используется при оценке эффективности алгоритмов снижения размерности данных при разделении пациентов на кластеры.

Таблица 1

Количество пациентов и генов в наборах данных BRCA

Вид молекулы	Число пациентов					Число генов
	LumA	LumB	Basal	Her2	Normal	
meDNA	430	148	136	46	110	21776
mRNA	579	219	191	82	141	20126
miRNA	407	139	133	58	110	625

ВЫЧИСЛИТЕЛЬНЫЙ ЭКСПЕРИМЕНТ

Выполнен анализ данных о метилированной ДНК, микроРНК и информационной РНК молекулах с использованием методов снижения размерности данных. Эффективность алгоритмов оценена по четырём критериям качества анализа: 1) время работы метода – t, 2) отношение средних внутрикластерных и межкластерных расстояний Q_1, 3) отношение суммы квадратов внутрикластерных и межкластерных расстояний Q_2, 4) точность классификации наборов данных в сниженном пространстве наиболее информативных компонент с использованием алгоритма случайного леса (с англ. – «Random forest») acc_{RF}.

РЕЗУЛЬТАТЫ

Реализованы методы PCA, ICA, tSNE, UMAP, MDS и NMF. Результаты сравнительного анализа алгоритмов для трех наборов данных представлены в таблице 2, диаграммы наиболее информативных компонент для набора данных meDNA – на рисунке. Наилучшие результаты в ходе анализа данных meDNA получены для алгоритма UMAP, для которого значения критериев $Q_1 = 26$ и $Q_2 = 19$, $acc_{RF} = 73\%$, что ненамного ниже, чем в исходном пространстве признаков (79%), время работы – $t = 33$ с. Время работы алгоритма с увеличением объёма данных и сложности существенно не изменяется. Установлено более четкое разделение подтипов рака Basal, Normal и Her2, в сравнении с другими алгоритмами. Кластеры подтипов LumA и LumB плохо разделимы.

Таблица 2

Оценки критериев качества работы алгоритмов на экспериментальных наборах данных

Набор BRCA	t, c						Q_1					
	PCA	ICA	tSNE	UMAP	MDS	NMF	PCA	ICA	tSNE	UMAP	MDS	NMF
meDNA	2072	1941	2006	**33**	237	3225	51	51	38	**26**	51	30
mRNA	2424	1481	1788	**45**	386	2500	59	29	41	**25**	29	35
miRNA	8	**1**	194	3	3	7	52	52	43	38	52	**34**

Продолжение таблицы 2

Набор BRCA	Q_2						acc_{RF}, %					
	PCA	ICA	tSNE	UMAP	MDS	NMF	PCA	ICA	tSNE	UMAP	MDS	NMF
meDNA	68	67	43	**19**	67	25	61	50	71	**73**	61	62
mRNA	66	17	35	**15**	17	24	56	74	77	**79**	75	62
miRNA	66	61	51	37	61	**27**	50	52	67	**69**	53	56

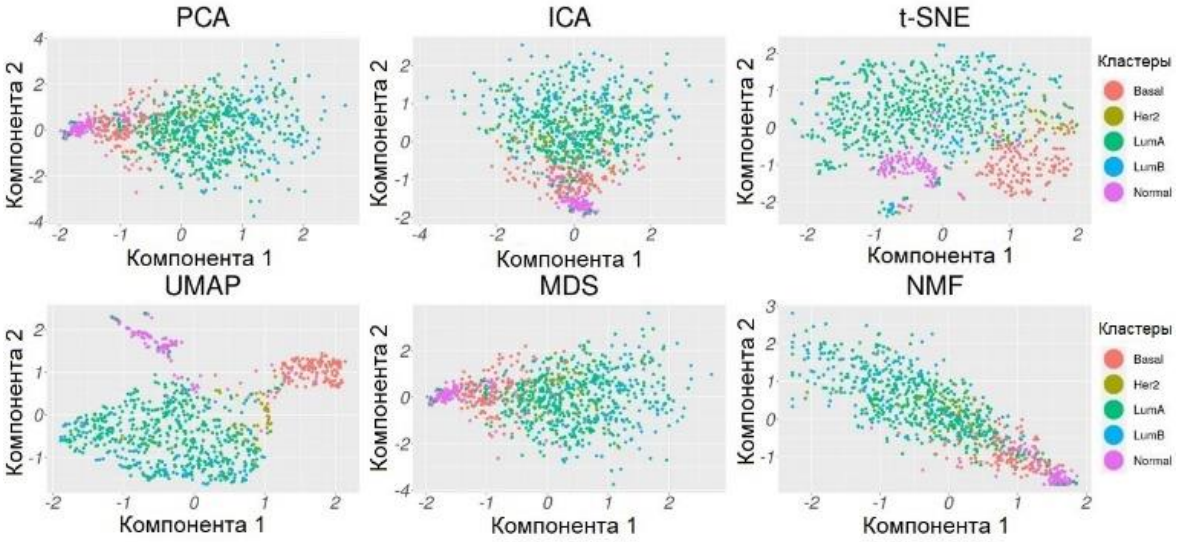

Рис. 1. Диаграммы разброса для кластеров данных meDNA в координатах двух наиболее информативных компонент, вычисленных методами PCA, ICA, tSNE, UMAP, MDS и NMF.

ЗАКЛЮЧЕНИЕ

Выполнен сравнительный анализ шести методов снижения размерности данных. Наиболее оптимальным алгоритмом снижения размерности данных для исследования экспрессии генов данных рака груди является метод UMAP, точность которого на данных о метилировании ДНК – 73%, экспрессии микроРНК – 69%, экспрессии информационной РНК – 79%.

Библиографические ссылки

1. *Li, X.* Genomic Analysis of Liver Cancer Unveils Novel Driver Genes and Distinct Prognostic Features / *X. Li, W. Xu, W. Kang* // Theranostics. 2018. № 8. P. 1740–1751.
2. *Яцков, Н. Н.* Комплексный анализ данных при исследовании сложных бимолекулярных систем / *Н. Н. Яцков, В. В. Апанасович* // Информатика. – 2021. − Т. 18, № 1. – С. 105–122.
3. *Maaten, L., Hinton, G.* Visualizing data using t-SNE. Journal of machine learning research. 2008. Vol. 9, P. 2579–2605.
4. *Горбунова, А.А.* Сравнительный анализ алгоритмов снижения размерности // 77-я научн. конф. студентов и аспирантов Белорусского государственного университета: материалы конф. В 3 ч. Ч. 1, Минск, 11–22 мая 2020 г. / Белорус. гос. ун-т; редкол.: *В. Г. Сафонов (пред.) [и др.]*. – Минск: БГУ, 2020. – С. 161–164.