СРАВНИТЕЛЬНЫЙ АНАЛИЗ АЛГОРИТМОВ СНИЖЕНИЯ РАЗМЕРНОСТИ ДАННЫХ ДЛЯ ИССЛЕДОВАНИЯ ЭКСПРЕССИИ ГЕНОВ

А. А. Горбунова

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

Работа посвящена сравнительному анализу алгоритмов снижения размерности данных, таких как методы главных и независимых компонент, стохастического вложения соседей с t-распределением, равномерного приближения и проекции, многомерного шкалирования, неотрицательной матрицы факторизации, с целью классификации групп пациентов на основе данных об экспрессии генов. Сравнительный анализ реализованных методов выполнен на смоделированных наборах данных, представляющих кластеры различной сложности. Наилучшими алгоритмами являются методы стохастического вложения соседей с t-распределением и равномерного приближения и проекции, позволяющие наиболее эффективно осуществить классификацию пациентов. В методе стохастического вложения соседей с t-распределением среднее отношение кластерных расстояний составляет 17,71, в методе равномерного приближения и проекции – 12,50. Среднее время работы метода стохастического вложения соседей с t-распределением составляет 12,7 с, метода равномерного приближения и проекции – 1,3 с.

Ключевые слова: методы снижения размерности; экспрессия генов; классификация; алгоритмы моделирования; критерии качества анализа.

ВВЕДЕНИЕ

Развитие биотехнологий напрямую связано с разработкой эффективных методов и алгоритмов обработки большого объема информации, получаемой в результате секвенирования последовательностей ДНК и РНК. Используя секвенирование РНК, можно количественно измерить генную экспрессию, обнаружить новые транскрипты и однонуклеотидные полиморфизмы. Однако реализация данной задачи требует оптимального использования существующих алгоритмов снижения размерности данных. Среди существующих алгоритмов 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. *multidimen-*

sional scaling), метод неотрицательной матрицы факторизации (NMF от англ. *non-negative matrix factorization*) [1-2].

Цель работы – изучить и исследовать алгоритмы снижения размерности данных на примере классификации групп пациентов на основе анализа данных об экспрессии генов.

МОДЕЛИРОВАНИЕ НАБОРОВ БИОДАННЫХ

В работе реализована имитационная модель кластеров многомерных данных [3], учитывающая количество кластеров, индекс разделения между кластерами, количество шумовых, нешумовых и прочих признаков. Для сравнения алгоритмов снижения размерности данных смоделированы три набора данных разной сложности, далее именуются система 1 (чистые данные), система 2 (средние данные) и система 3 (зашумленные данные). При моделировании данных важно реализовать наборы различными по степени разделения кластеров, наличию шумовых и нешумовых признаков и количеству выбросов. Наборы данных представляют кластеры 10-мерных данных. Количество кластеров в каждом наборе – 3.

Для моделирования чистых данных заданы следующие параметры: индекс разделения между кластерами – 0,6, количество нешумовых переменных – 8, количество шумовых переменных – 2, количество выбросов – 3. В результате получен набор данных, содержащий 462 объекта.

Для моделирования средних данных заданы следующие параметры: индекс разделения между кластерами – 0,4, количество нешумовых переменных – 4, количество шумовых переменных – 6, количество выбросов – 3. В результате получен набор данных, содержащий 485 объектов.

Для моделирования зашумленных данных заданы следующие параметры: индекс разделения между кластерами – 0,2, количество нешумовых переменных – 3, количество шумовых переменных – 7, количество выбросов – 10. В результате получен набор данных, содержащий 342 объекта.

РЕЗУЛЬТАТЫ

Качество анализа данных оценено по трем критериям: 1) время работы метода *t*, 2) средние внутрикластерные $\langle d_{\text{intra}} \rangle$ и межкластерные $\langle d_{\text{inter}} \rangle$ расстояния 3) отношения средних внутрикластерных и межкластерных расстояний $\frac{\langle d_{\text{intra}} \rangle}{\langle d_{\text{inter}} \rangle}$. Эффективность работы алгоритма определена следующим образом: чем меньше время, затраченное на работу алго-

ритма, тем он лучше; чем меньше внутрикластерное расстояние, тем больше точки сконцентрированы внутри кластера, тем эффективнее метод; чем больше межкластерное расстояние, тем дальше кластеры удалены друг от друга, тем эффективнее алгоритм.

Результаты классификации пациентов на основе анализа данных об экспрессии генов представлены для шести методов снижения размерности данных на рисунке. Наилучшие результаты показал алгоритм UMAP, отличительной особенностью которого является построение взвешенного неориентированного графа более низкой размерности. Алгоритм показал наименьшее значение $\langle d_{\mathbf{intra}} \rangle$ = 50,41, наибольшее значение $\langle d_{\mathbf{inter}} \rangle$ = 2,28, и наименьшее отношение $\frac{\langle d_{\mathbf{intra}} \rangle}{\langle d_{\mathbf{inter}} \rangle}$ = 22,15.

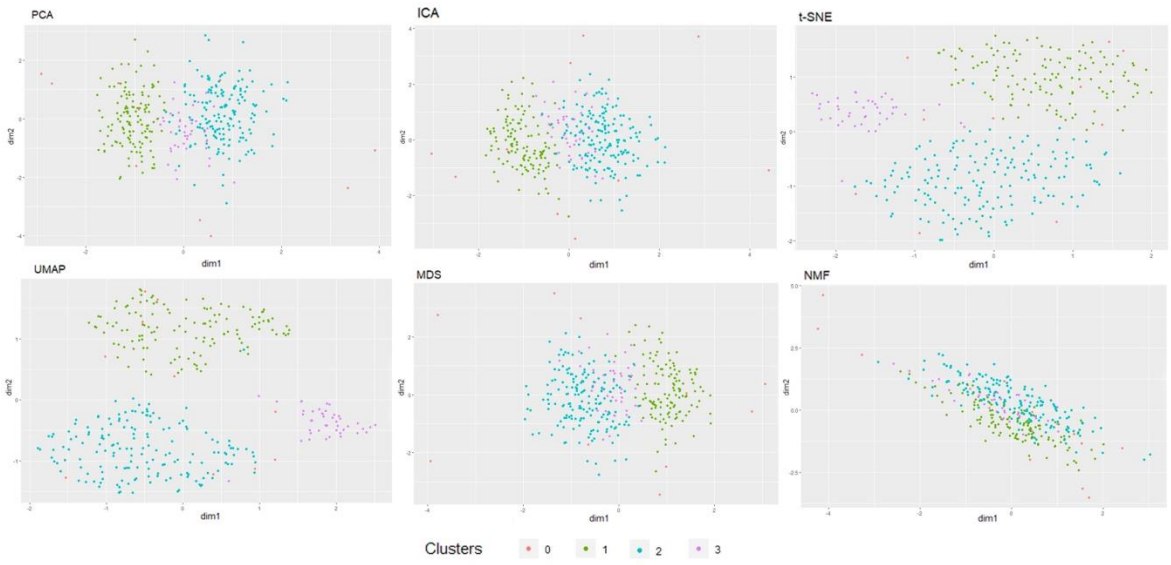

Рис. 1. Диаграммы разброса для кластеров данных системы 3 в пространствах наиболее информативных координат, вычисленных методами PCA, ICA, tSNE, UMAP, MDS, NMF

Методы tSNE и UMAP даже на сложных данных успешно выполняют задачу, однако UMAP работает качественнее, в tSNE наблюдается менее выраженное разделение, в то время как после выполнения остальных методов наблюдается частичное или полное перекрытие нескольких кластеров. По степени разделения данных наилучшие результаты демонстрирует метод UMAP. Значения $\langle d_{\mathbf{intra}} \rangle$ и $\langle d_{\mathbf{inter}} \rangle$ наихудшие у NMF. PCA имеет низкие показатели по рассматриваемым критериям, однако обладает более высокой скоростью работы. Самый быстрый алгоритм ICA плохо справляется с зашумленными данными, наблюдается значительное перекрытие кластеров. MDS показывает схожие результаты с ICA, однако проигрывает по времени работы. Оценки критериев качества работы алгоритмов представлены в таблице.

Таблица 1

Оценки критериев качества работы алгоритмов на смоделированных наборах данных

Система	Параметр	Алгоритм					
		PCA	ICA	tSNE	UMAP	MDS	NMF
Система 1	*t*, с	0,02	0,01	15,42	1,38	0,22	0,22
	$\frac{\langle d_{\text{intra}} \rangle}{\langle d_{\text{inter}} \rangle}$	28,99	13,27	11,74	5,19	13,27	10,70
Система 2	*t*, с	0,01	0,01	15,11	1,41	0,27	0,23
	$\frac{\langle d_{\text{intra}} \rangle}{\langle d_{\text{inter}} \rangle}$	51,18	28,10	15,18	10,05	28,10	19,58
Система 3	*t*, с	0,03	0,02	7,57	1,06	0,11	0,25
	$\frac{\langle d_{\text{intra}} \rangle}{\langle d_{\text{inter}} \rangle}$	50,19	54,15	26,20	22,15	54,15	112,95

ЗАКЛЮЧЕНИЕ

Выполнен сравнительный анализ шести методов снижения размерности данных: PCA, ICA, tSNE, UMAP, MDS, NMF. Лучше всего с задачей разделения кластеров данных об экспрессии генов справляется UMAP (среднее отношение кластерных расстояний 12,5). Методы PCA (43,5) и NMF (47,7) подходят преимущественно для чистых данных невысокой размерности, при усложнении данных наблюдается большое перекрытие кластеров. Самый быстрый из представленных метод ICA (31,8) и сходный в работе с ним метод MDS (31,8) следует использовать на чистых или средних данных. Метод tSNE (17,7) отлично подходит для разделения чистых и средних данных, на зашумленных данных метод работает приемлемо, но хуже, чем UMAP. При увеличении начальной или конечной размерности данных, tSNE значительно проигрывает UMAP.

Выполненная работа позволяет сделать вывод о том, что наиболее оптимальными алгоритмами снижения размерности данных для исследования экспрессии генов являются методы tSNE и UMAP.

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. *Maaten, L., Hinton, G.* Visualizing data using t-SNE. Journal of machine learning research. 2008. Vol. 9, P. 2579-2605.
3. *Qiu W., Joe H.* Generation of Random Clusters with Specified Degree of Separation // Journal of Classification. 2006. Vol. 23. DOI: 10.1007/s00357-006-0018-y.