

Содержание

[Подзаголовок документа]

Natalia

Объединенный институт проблем информатики НАН Беларуси

При поддержке Министерства Белорусского республиканского

фонда фундаментальных исследований (БРФФИ)

проект №Ф19УКРГ-005

Разработка методов, алгоритмов и интеллектуальной аналитической системы для обработки и анализа разнородных клинических и биомедицинских данных с целью совершенствования диагностики сложных заболеваний

ПРОМЕЖУТОЧНЫЙ ОТЧЕТ

Разработка алгоритмов и программных модулей, реализующих предобработку и выделение информативных признаков, кластеризацию биомедицинских данных высокой размерности. Разработка гибридной модели классификации биомедицинских данных на основе ансамбля классификаторов.

31.03.2020

СПИСОК ИСПОЛНИТЕЛЕЙ

|  |  |  |  |
| --- | --- | --- | --- |
| Руководитель проекта,  ведущий научный сотрудник, к.т.н. |  | .12.2019 | Н.А. Новоселова (научное  руководство, разделы 1-2,  введение, заключение) |
| Ведущий научный сотрудник,  к.т.н. |  | .12.2019 | В.Ю. Скобцов  (раздел 3, заключение) |
| Научный сотрудник |  | .12.2019 | Т.С. Кухта  (раздел 1) |
| Ведущий инженер-про-граммист |  | .12.2019 | А.Н. Петрашко (оформление) |

Реферат

Отчет 49 с., 15 рис., 11 табл.

ПРЕДОБРАБОТКА ДАННЫХ, ФУНКЦИОНАЛЬНАЯ СЕТЬ, РАНЖИРОВАНИЕ ПРИЗНАКОВ, БИОМАРКЕР, КЛАСТЕРИЗАЦИЯ, КЛАССИФИКАЦИОННАЯ МОДЕЛЬ

Объектом исследования являются биомедицинские и клинические данные сложных заболеваний.

Цель работы – разработка методов, алгоритмов анализа разнородных клинических и биомедицинских данных и экспериментального прототипа интеллектуальной аналитической системы, который представляет собой программную реализацию всех этапов обработки и анализа данных в виде отдельных программных модулей и предназначен для проведения сравнительного анализа различных подходов, имеющих целью совершенствование диагностики и лечения сложных заболеваний.

В процессе выполнения проекта разработаны ряд алгоритмов для предобработки и выделения информативных признаков из биомедицинских данных, а также алгоритм кластеризации биомедицинских данных высокой размерности на основе методов спектрального анализа. Алгоритмы предобработки признаков позволяют как ранжировать признаки по их информативности относительно классов (фенотипов заболевания или групп «случай-контроль»), так и отбирать не избыточные комбинации биомаркеров с учетом их коррелированности. На основе интеграции двух источников данных, а именно данных генной экспрессии и данных функциональных и физических взаимодействий генов и их продуктов, представленных в виде сетей, разработан алгоритм выделения функциональных модулей, связанных с заболеванием. Алгоритм спектральной кластеризации позволяет представлять исходное множество данных в виде низко размерной структуры данных и выявлять устойчивые кластеры, соответствующие подтипам заболеваний, в данных, характеризующихся зашумленностью и наличием выбросов. Все вышеперечисленные алгоритмы реализованы в виде двух программных модулей: модуль предобработки и выделения комбинаций биомаркеров и модуль кластеризации биологических данных. Все разработанные алгоритмы и их соответствующие программные реализации были протестированы на искусственно сгенерированных и реальных биомедицинских данных.

Разработана общая схема гибридной модели классификации на основе ансамбля классификаторов, которая позволяет объединить несколько источников разнородных данных с целью повышения точности диагностики сложных заболеваний, надежности прогноза развития заболеваний и отклика на терапию.

**Содержание**

[Обозначения и сокращения 5](#_Toc34757560)

[Введение 6](#_Toc34757561)

[1 Реализация алгоритмов предобработки клинических и биомедицинских данных 7](#_Toc34757562)

[1.1 Краткое описание разработанных алгоритмов 7](#_Toc34757563)

[1.1.1 Алгоритм ранжирования признаков по информативности 7](#_Toc34757564)

[1.1.2 Алгоритмы выделения комбинаций биомаркеров на одном источнике данных 8](#_Toc34757565)

[1.2 Программная реализация алгоритмов предобработки клинических и биомедицинских данных 9](#_Toc34757566)

[1.2.2 Тестирование программного модуля 12](#_Toc34757567)

[**2 Алгоритм выделения комбинаций биомаркеров на интегрированных источниках данных** 18](#_Toc34757568)

[2.1 Краткое описание разработанного алгоритма 19](#_Toc34757569)

[2.2 Программная реализация алгоритма выделения комбинаций биомаркеров на интегрированных источниках данных 21](#_Toc34757570)

[2.2.1 Описание программной реализации 21](#_Toc34757571)

[2.2.2 Тестирование программной реализации алгоритма выделения комбинаций биомаркеров на интегрированных источниках данных 24](#_Toc34757572)

[**3 Алгоритм спектральной кластеризации биомедицинских данных на основе диффузионных карт и сетей** 27](#_Toc34757573)

[3.1 Диффузионные карты 28](#_Toc34757574)

[3.1.1 Построение графа и весовой матрицы исходного пространства данных 28](#_Toc34757575)

[3.1.2 Нормализация весовой матрицы графа исходного пространства данных 29](#_Toc34757576)

[3.1.3 Построение низко размерного многообразия , в которое вкладывается исходное многомерное пространство данных 30](#_Toc34757577)

[3.1.4 Кластеризация данных в итоговом евклидовом пространстве 30](#_Toc34757578)

[3.2 Обучение диффузионных отображений на основе модели диффузионных сетей 31](#_Toc34757579)

[3.2.1 Искусственные нейронные сети 31](#_Toc34757580)

[3.2.2 Диффузионные сети 33](#_Toc34757581)

[3.3 Программная реализация алгоритма спектральной кластеризации биомедицинских данных 33](#_Toc34757582)

[3.3.2 Тестирование программного модуля 34](#_Toc34757583)

[**4 Разработка гибридной модели классификации на основе ансамбля классификаторов** 39](#_Toc34757584)

[4.1 Разнородные источники биологических данных 41](#_Toc34757585)

[4.2. Различные схемы интеграции разнородных данных 42](#_Toc34757586)

[Список использованных источников 47](#_Toc34757587)

Обозначения и сокращения

|  |  |  |
| --- | --- | --- |
|  |  |  |
| ROC | Receiver operating characteristic | Функциональные характеристики приемника |
| AUC | Area under curve | Площадь под кривой |
| MDL | Minimal description length | Минимальная длина описания |
| SD | Standard deviation | Стандартное отклонение |
| CFS | Correlation feature selection | Отбор признаков на основе корреляции |
| AML | Acute myeloid leukemia | Острый миелобластный лейкоз |
| ALL | Acute lymphoblastic leukemia | Острый лимфобластный лейкоз |
| ДП |  | Дисрегулированный биологичсекий путь |
| ДНК |  | Дезоксирибонуклеи́новая кислота́ |
| BFS | Breadth-first search | Алгоритм поиска в ширину |
| IFM | Integrated Feature Module | Интегрированный модуль признаков |
| HPRD | [Human Protein Reference Database](https://www.re3data.org/repository/r3d100010978) | Репозиторий белковых взаимодействий |
| HD | [Huntington's disease](https://en.wikipedia.org/wiki/Huntington%27s_disease) | Болезнью Гентингтона |
| HTT | Huntingtin | Белок Гентингтин |
| GEO | Gene Expression Omnibus | База данных генной экспрессии |
| PCA | Principal component analysis | Анализ главных компонент |
| MDS | [Multidimensional scaling](https://en.wikipedia.org/wiki/Multidimensional_scaling) | Многомерное шкалирование |
| РНК |  | Рибонуклеи́новая кислота́ |
| SNP | Single nucleotide polymorphism | Однонуклеотидный полиморфизм |
| GMM | Gaussian mixture modelling | Моделирование гауссовой смеси |
| *t*-SNE | *t*-distributed Stochastic Neighbor Embedding | Стохастическое вложение соседей с [*t*-распределением](https://ru.wikipedia.org/wiki/%D0%A0%D0%B0%D1%81%D0%BF%D1%80%D0%B5%D0%B4%D0%B5%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5_%D0%A1%D1%82%D1%8C%D1%8E%D0%B4%D0%B5%D0%BD%D1%82%D0%B0) |

Введение

В настоящее время с целью совершенствования диагностики и лечения сложных заболеваний большое внимание уделяется комплексному анализу различного рода биомедицинских и клинических данных для понимания процессов, происходящих в организме на клеточном уровне и изменений, вызываемых развитием заболевания. Как известно причиной сложных заболеваний наряду со внешними факторами является комбинация генетических нарушений, что не позволяет выделить в качестве биомаркера какую-либо одну генетическую мутацию. Сложность также заключается в том, что индивидуальные генетические факторы могут различаться и отдельные случаи одного и того же заболевания (фенотипа) могут быть вызваны различными генетическими изменениями. Кроме того, в случае комбинированного влияния многих мутаций, индивидуальный эффект каждой из них может быть достаточно незначительным и, следовательно, трудно выявляемым. Необходимо также учитывать высокую неоднородность сложного заболевания, т.е. неоднородность его проявлений (фенотипов). Таким образом, разработка подходов к анализу разнородных биомедицинских данных, исследуемых в рамках выполнения совместного белорусского-украинского проекта и позволяющих выявлять взаимосвязи между генотипом и фенотипом, включая определение информативных признаков (биомаркеров) и их комбинаций, выявление групп заболеваний с похожими признаками, построение классификационных моделей на основе интегрированных данных является важным шагом в понимании гетерогенности сложных заболеваний и будет представлять собой научно-практическую базу для проведения сравнительных модельных экспериментов, направленных на решение задачи повышения качества диагностики.

Основные промежуточные результаты, полученные в рамках выполнения проекта связаны с разработкой методов и алгоритмов предварительной обработки и выделения признаков из разнородных биомедицинских данных, разработкой подхода к кластеризации биомедицинских данных с целью выявления типов заболевания, связанных с конкретными клиническими показателями, а также общей схемы гибридной модели классификации на основе ансамбля классификаторов, которая позволяет объединить несколько источников разнородных данных с целью повышения точности диагностики сложных заболеваний. На первом этапе выполнения проекта авторами были разработаны алгоритмы предобработки биомедицинских данных, которые позволяют как ранжировать признаки по их информативности относительно классов (фенотипов заболевания или групп «случай-контроль»), так и отбирать не избыточные комбинации биомаркеров с учетом их коррелированности. На основе интеграции двух источников данных, а именно данных генной экспрессии и данных функциональных и физических взаимодействий генов и их продуктов, представленных в виде сетей, разработан алгоритм выделения функциональных модулей, связанных с заболеванием. На втором этапе, разработан программный модуль, реализующий предложенные алгоритмы выделения признаков, выполнена программная реализация алгоритма выделения функциональных модулей на основе интегрированных данных. С использованием ряда искусственно сгенерированных и реальных наборов данных проведены тестовые эксперименты разработанных алгоритмов. Кроме того, разработан алгоритм кластеризации данных с целью выявления групп пациентов со схожими значениями биомедицинских признаков и характеризующих некоторый определенный клинический показатель.

На третьем этапе разработан программный модуль, реализующий алгоритм спектральной кластеризации биологических данных, который протестирован на реальном наборе данных генной экспрессии. Разработана общая схема гибридной модели классификации на основе ансамбля классификаторов, которая позволяет объединить несколько источников разнородных биологических данных с целью повышения точности диагностики, надежности прогноза развития заболеваний и отклика на терапию.

1 Реализация алгоритмов предобработки клинических и биомедицинских данных

1.1 Краткое описание разработанных алгоритмов

1.1.1 Алгоритм ранжирования признаков по информативности

Разработанный алгоритм ранжирования признаков основан на построении ROC-кривой [1] для значений каждого отдельно взятого признака с последующим определением порогового значения и его значимости для классификации групп заболеваний. Построение ROC-кривой осуществляется для оценки признака для бинарной классификации (т.е. классификации на две группы).

ROC-кривая представляет собой множество возможных пар значений правильно и неправильно отнесенных к положительному классу элементов данных, определяемых путем разбиения диапазона значений  по пороговому значению на две подобласти:

, (1.1)

где  ‑ доля истинно положительных результатов и  ‑ доля ложно положительных результатов.

Признак является наиболее информативным, если полностью разделяет элементы двух классов. В этом случае для некоторого значения  значение  и . Большинство признаков имеют ROC-кривую, которая лежит между соответствующими кривыми для неинформативного и высоко информативного теста.

Не визуальным методом сравнения ROC-кривых является оценка площади под кривыми. Численный показатель площади под кривой называется AUC [2]. Таким образом, ранжирование признаков, характеризующих группы «случай-контроль», может выполняться на основе результатов вычислений AUC для каждого из признаков с последующей сортировкой результатов по убыванию значений AUC.

1.1.2 Алгоритмы выделения комбинаций биомаркеров на одном источнике данных

Разработанный алгоритм отбора комбинаций информативных признаков использует метод прямого отбора признаков на основе значений интегрированного критерия, состоящего из комбинации двух критериев: критерия релевантности и критерия избыточности признака. Для расчета критериев используется выражение для нормализованной взаимной информации случайных величин и :

, (1.2)

где – представляет собой энтропию случайной величины .

Критерий релевантности рассчитывается для каждого признака, характеризующего объекты данных и целевого признака, соответствующего классу, следующим образом

, (1.3)

где является множеством всех признаков, – множество отобранных признаков, – вектор меток классов.

Критерий избыточности рассчитывается для каждого признака относительно уже отобранных признаков следующим образом

. (1.4)

Согласно алгоритму, задача отбора информативных признаков формулируется как однокритериальная оптимизационная задача вида:

, (1.5)

где функция  может быть представлена в следующих двух видах:

 (1.6)

 (1.7)

Точное решении данной оптимизационной задачи требует выполнения количества вычислений равное  и является сложным в вычислительном плане. Поэтому в нашем исследовании мы используем эвристический подход к решению оптимизационной задачи (1.5).

Эвристический подход:

1. Из входного множества маркеров  отбираем признак , такой что  и включаем его в подмножество , исключая из множества .
2. Для отбора каждого последующего признака в подмножество *S* выполняем следующие шаги:
3. Для каждого маркера  рассчитываем значение релевантности по следующим формулам:

1)  в случае оптимизации функции (1.6);

2)  в случае оптимизации функции (1.7).

1. Для включения в подмножество *S* выбираем маркер  такой, что  и включаем в подмножество с исключением из множества.
2. Шаги 2-4 повторяются до выполнения условия остановки: отбор заданного количества *s* признаков или в случае когда .

1.2 Программная реализация алгоритмов предобработки клинических и биомедицинских данных

1.2.1 Описание программного модуля

Программный модуль, реализующий алгоритмы предобработки и выделения признаков реализован как набор функций в среде RStudio на языке программирования R [3]. Программный модуль включает функции для ранжирования признаков, выделения подмножеств информативных признаков, а также ряд функций для выполнения классификации и оценки отобранных комбинаций признаков с использованием моделей классификации. В дополнение к разработанным алгоритмам реализованы ряд существующих подходов к ранжированию и отбору признаков, а также ряд известных моделей классификации. В связи с тем, расчет значений взаимной информации выполняется только для номинальных признаков, то для возможности ее применения к непрерывным признакам используется дискретизация их значений. Основными функциями программного модуля являются «select.process» и «classifier.loop». Функция «select.process» является программной оболочкой для вызовы конкретных алгоритмов ранжирования и отбора признаков. Функция «classifier.loop» предназначена для построения классификаторов со встроенной процедурой отбора признаков и оценкой результатов. Все реализованные функции программного модуля перечислены в таблице 1.

Таблица 1 – Описание основных функций программного модуля

|  |  |
| --- | --- |
| Функций | Описание |
| ProcessData | Выполняет дискретизацию значений непрерывных признаков. Предоставляется выбор нескольких способов дискретизации: на основе принципа минимальной длины описания (MDL), с использованием среднего и стандартного отклонения значений признака (SD), равными интервалами (equal interval) и интервалами с равной частотой (equal frequency). |
| Продолжение таблицы 1 | |
| compute.aucs | Реализует разработанный алгоритм ранжирования признаков. Функция позволяет рассчитать веса признаков с использованием AUC значений. Данные значения рассчитываются только для числовых признаков по отношению к классу. Результатом является структура «data.frame», состоящая из трех полей: имя признака, AUC значение и уровень положительного класса. Данная функция является внутренней для функции построения классификатора «classifier.loop» со встроенным отбором признаков с входным параметром «auc». |
| select.inf.chi2 | Выполняет ранжирование признаков путем расчета весов с использованием критерия хи-квадрат [4]. Применима как для числовых, так и для номинальных признаков. Для непрерывных признаков изначально выполняется дискретизация с использованием функции «[ProcessData](http://127.0.0.1:19146/help/library/Biocomb/help/ProcessData)». Результатом является структура «data.frame», состоящая из трех полей: имя признака, значение хи-квадрат и порядковый номер признака в наборе данных. Данная функция является внутренней для функции построения классификатора «classifier.loop» со встроенным отбором признаков с входным параметром «Chi-square» . |
| select.inf.gain | Выполняет ранжирование признаков путем расчета весов с использованием критерия информационного выигрыша (Information Gain) [4]. Применима как для числовых, так и для номинальных признаков. Для непрерывных признаков изначально выполняется дискретизация с использованием функции «[ProcessData](http://127.0.0.1:19146/help/library/Biocomb/help/ProcessData)». Результатом является структура «data.frame», состоящая из трех полей: имя признака, значение информационного выигрыша и порядковый номер признака в наборе данных. Данная функция является внутренней для функции построения классификатора «classifier.loop» со встроенным отбором признаков с входным параметром «InformationGain». |
| select.inf.symm | Выполняет ранжирование признаков путем расчета весов с использованием симметричного критерия неопределенности [4]. Применима как для числовых, так и для номинальных признаков. Для непрерывных признаков изначально выполняется дискретизация с использованием функции «[ProcessData](http://127.0.0.1:19146/help/library/Biocomb/help/ProcessData)». Результатом является структура «data.frame», состоящая из трех полей: имя признака, значение критерия неопределенотси и порядковый номер признака в наборе данных. Данная функция является внутренней для функции построения классификатора «classifier.loop» со встроенным отбором признаков с входным параметром «symmetrical.uncertainty». |
| select.mrmr.filter | Реализует разработанный алгоритм выделения комбинаций признаков (биомаркеров). Функция использует эвристический метод прямого отбора признаков в комбинацию на основе интегрированного критерия, оценивающего релевантность признака по отношению к классу и его уровень избыточности по отношению к уже отобранному набору признаков. Применима как для числовых, так и для номинальных признаков. Для непрерывных признаков изначально выполняется дискретизация с использованием функции «[ProcessData](http://127.0.0.1:19146/help/library/Biocomb/help/ProcessData)». Результатом является структура «data.frame», состоящая из двух полей: имя признака и порядковый номер признака в наборе данных. Данная функция является внутренней для функции построения классификатора «classifier.loop» со встроенным отбором признаков с входным параметром «MRMR». |
| Продолжение таблицы 1 | |
| select.cfs | Выполняет отбор комбинации признаков с использованием стратегии поиска «лучший — первый» на основе корреляционной меры (CFS). CFS оценивает подмножество признаков по эффективности каждого признака предсказывать класс и обеспечивать добавочную информацию. Применима как для числовых, так и для номинальных признаков. Для непрерывных признаков изначально выполняется дискретизация с использованием функции «[ProcessData](http://127.0.0.1:19146/help/library/Biocomb/help/ProcessData)». Результатом является структура «data.frame», состоящая из двух полей: имя признака и порядковый номер признака в наборе данных. Данная функция является внутренней для функции построения классификатора «classifier.loop» со встроенным отбором признаков с входным параметром «CFS». |
| select.forward.wrapper | Выполняет отбор комбинации признаков с использованием встроенной стратегии с помощью классификационной модели деревьев решений и метода прямого отбора признаков. Классификатор используется для оценки качества отобранной комбинации. Функция использует встроенную процедуру перекрестной проверки для оценки точности классификации на подмножестве признаков. Результатом является вектор с именами отобранных признаков. Данная функция является внутренней для функции построения классификатора «classifier.loop» со встроенным отбором признаков с входным параметром «CorrSF». |
| select.process | Основная функция для вызова всех вышеперечисленных функций. Представляет собой программную оболочку для выполнения ранжирования и отбора признаков для двух и более классов. Результатом является числовой вектор с номерами колонок отобранных признаков в случае алгоритмов отбора комбинаций признаков и отсортированными по соответствующему критерию номерами колонок в случае ранжирования признаков. Максимальное возвращаемое количество признаков может быть ограничено с использованием входного параметра «max.no.features». Выходные результаты используются функцией «classifier.loop» в процессе классификации. |
| classifier.loop | Основная функция построения и оценки классификаторов со встроенной процедурой отбора признаков. Для оценки качества классификации используется процедура перекрестной проверки ("fold-crossval"). Реализовано несколько известных моделей классификации, включая наивный байесовский классификатор ("nn"), мультиномиальную логистическую регрессию ("mlr"), машины опорных векторов ("svm"), линейный дискриминантный анализ ("lda"). Функция вызывает «[select.process](http://127.0.0.1:19146/help/library/Biocomb/help/select.process)» для выполнения отбора признаков, что позволяет улучшить точность классификации. in order to perform feature selection for classification, which helps to improve the quality of the classifier. Выходом функции является структура «list», включая «data.frame» с результатами классификации для каждого классификатора, матрицу частоты отбора каждого признака, вектор или матрицу со значениями точности классификации. |

1.2.2 Тестирование программного модуля

Для тестирования работы функций и сравнительного анализа различных подходов к отбору признаков был проведен ряд экспериментов как на искусственно сгенерированных, так и реальных данных генной экспрессии из базы данных TCGA [5].

Искусственно сгенерированные данные data\_test, используемые для тестирования состоят из 300 объектов, признаков и делятся на три класса (рисунок 1.1). Признаки x1-x5 являются информативными и определяют кластерную структуру набора данных. Кластеры сгенерированы в двухмерном пространстве x1-x2. Значения признаков x3-x5 имеют такое же распределение как признак x2. Значения признаков x1-x5 имеют нормальное распределение для каждого класса с различными значениями средних и равными стандартными отклонениями. Остальные признаки набора данных являются случайными величинами с равномерным распределением в интервале [0, 1].

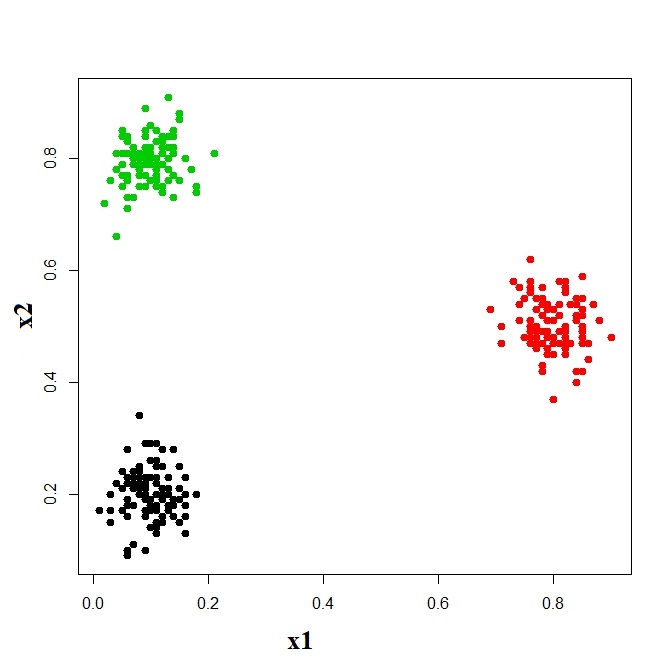


Рисунок 1.1 – Проекция искусственно сгенерированного набора данных на двухмерное пространство x1-x2

Два набора данных генной экспрессии являются реальными данными из базы данных TCGA и представляют собой данные по лейкемии [6] и раку кишечника [7] (таблица 2). Данные по раку кишечника содержат образцы нормальных и злокачественных образцов. Данные по лейкемии состоят из трех подтипов заболевания: острый миелобластный лейкоз (AML), T-линейный и B- линейный острый лимфобластный лейкоз (ALL). Данный набор данных может использоваться как для бинарной классификации: 47 ALL и 25 AML, так и для классификации на 3 класса: 38 B-ALL, 9 T-ALL и 25 AML.

Таблица 2 – Описание реальных наборов данных

|  |  |  |  |
| --- | --- | --- | --- |
| Набор данных | Количество объектов | Количество признаков | Количество и состав классов |
| Лейкемия (leukemia) | 72 | 7070 | 2 (ALL, AML)  3 (T-ALL, B-ALL, AML) |
| Рак кишечника  (colon) | 62 | 2000 | 2 (Normal, Tumor) |

Общая схема эксперимента состоит из выполнения следующих шагов:

* выполнение дискретизации непрерывных признаков наборов данных с использованием подхода на основе принципа минимальной длины описания (MDL) для использования алгоритмов отбора, основанных на расчете информационных критериев;
* выполнение ранжирования признаков для искусственно сгенерированных данных с использованием разработанного алгоритма и функции «compute.aucs» и select.mrmr.filter. Сравнительный анализ полученных результатов;
* выполнение ранжирования признаков реального набора данных leukemia с использованием разработанного алгоритма отбора признаков и функции «select.mrmr.filter», и с использованием алгоритма ранжирования признаков на основе информационного выигрыша и функции select.inf.gain. Сравнение полученных результатов с использованием классификационной модели на подмножестве признаков фиксированного размера;
* выполнение ранжирования признаков реального набора данных leukemia с использованием разработанного алгоритма отбора признаков и функции «select.mrmr.filter», и с использованием алгоритма ранжирования признаков на основе информационного выигрыша и функции select.inf.gain. Сравнение полученных результатов с использованием классификационной модели при увеличении количества отобранных признаков от 1 до 30 и оценка ее эффективности. Для оценки используется процедура перекрестной проверки.

Для набора данных data\_test было выполнено ранжирование признаков с использованием разработанных алгоритмов на основе расчета AUC и интегрированного критерия, а также с использованием подхода на основе информационного выигрыша. Результаты приведены в таблице 3.

Из таблицы 3 видно, что признаки x1-x5 имеют наибольшие значения для соответствующих критериев, что согласуется со схемой генерации данных. Согласно этой схемы признаки x2-x5 определяют три кластера и признак x1 определяет два кластера в наборе данных. Остальные признаки являются неинформативными для формирования кластерной структуры. В отличие от других методов ранжирования предложенный алгоритм MRMR позволяет зафиксировать избыточность информации, которая присутствует в связи с наличием коррелированных значений признаков x2-x5. Согласно этому алгоритму в качестве наиболее информативного признака выбирается признак x5, тогда как следующим в процесс отбора определяется неинформативный признак x7 со значением интегрированного критерия равным 0. В случае выбора порогового значения критерия в качестве критерия останова, алгоритм позволяет отобрать комбинацию не избыточных информативных признаков, позволяющих объяснить кластерную структуру данных. В случае набора данных data\_set в качестве такой комбинации выбирается один признак x5.

Таблица 3 – Результаты ранжирования признаков набора данных data\_test с использованием нескольких критериев

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Information Gain | | | MRMR | | | AUC | | |
| Признак | Значение | Порядковый номер | Признак | Значение | Порядковый номер | Признак | Значение | Порядковый номер |
| x2 | 1,0986 | 2 | x5 | 1,5850 | 5 | x1 | 0,8415 | 1 |
| x3 | 1,0986 | 3 | x7 | 0,0000 | 7 | x2 | 1,0000 | 2 |
| x4 | 1,0986 | 4 | x3 | 0,7843 | 3 | x3 | 1,0000 | 3 |
| x5 | 1,0986 | 5 | x2 | 0,5229 | 2 | x4 | 1,0000 | 4 |
| x1 | 0,6365 | 1 | x4 | 0,3922 | 4 | x5 | 1,0000 | 5 |
| x10 | 0,0125 | 10 | x1 | 0,1820 | 1 | x6 | 0,5266 | 6 |
| x7 | 0,0113 | 7 | x10 | 0,0017 | 10 | x7 | 0,5390 | 7 |
| x6 | 0,0066 | 6 | x6 | 0,0011 | 6 | x8 | 0,5115 | 8 |
| x9 | 0,0018 | 9 | x8 | -0,0040 | 8 | x9 | 0,4986 | 9 |
| x8 | 0,0013 | 8 | x9 | 0,0000 | 9 | x10 | 0,4973 | 10 |

Результаты применения различных алгоритмов ранжирования и отбора признаков оцениваются с использованием классификационной модели. При этом используются различные схемы разбиения набора данных на обучающее и тестовое множество. В наших экспериментах для оценки эффективности классификатора использовалась процедура перекрестной проверки. При ее использовании на различных подмножествах признаков отбор признаков и построение модели классификации осуществляются на обучающем множестве, эффективность же классификации оценивается на тестовом множестве. Данный подход позволяет получить несмещенную оценку точности классификации. Для реальных наборов данных точность классификации с использованием разработанного алгоритма отбора комбинаций признаков MRMR сравнивалась с методом на основе информационного выигрыша (InformationGain), который не принимает во внимание коррелированность признаков в наборе данных. В качестве классификаторов использовались следующие модели:

1. наивный байесовский классификатор (NBС);
2. классификатор ближайших соседей (NN);
3. мультиномиальная логистическая регрессия (MLR);
4. машины опорных векторов (SVM).

На первом этапе было выполнено сравнение эффективности классификации для наборов признаков фиксированного размера , полученных с использованием двух подходов. Результаты представлены в таблице 4.

Таблица 4 – Результаты оценки количества ошибок классификаторов на наборах реальных данных с фиксированным значением признаков .

|  |  |  |  |
| --- | --- | --- | --- |
| Классификатор |  | Набор данных | |
|  | Метод отбора | leukemia | colon |
| NBС | MRMR | 2 | 7 |
| InformationGain | 4 | 11 |
| NN | MRMR | 3 | 6 |
| InformationGain | 4 | 8 |
| MLR | MRMR | 2 | 6 |
| InformationGain | 5 | 10 |
| SVM | MRMR | 4 | 7 |
| InformationGain | 5 | 8 |

Из таблицы 4 видно, что применение алгоритма отбора MRMR позволяет повысить точность классификации наборов данных по сравнению со стандартным ранжированием признаков на основе критерия информационного выигрыша.

В таблице 5 приведены результаты эффективности классификации для набора данных leukemia c использованием возрастающего количества признаков от 1 до 30.

Таблица 5 – Результаты оценки количества ошибок классификаторов на наборе данных leukemia

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Классификатор |  | Количество признаков | | | | | | |
| Метод отбора | 1 | 3 | 5 | 7 | 10 | 20 | 30 |
| NBС | MRMR | 10 | 5 | 4 | 2 | 4 | 3 | 3 |
| InformationGain | 9 | 6 | 7 | 6 | 4 | 3 | 3 |
| NN | MRMR | 11 | 6 | 6 | 5 | 5 | 4 | 4 |
| InformationGain | 10 | 6 | 5 | 5 | 6 | 3 | 3 |
| MLR | MRMR | 10 | 4 | 4 | 3 | 4 | 3 | 3 |
| InformationGain | 9 | 8 | 7 | 6 | 5 | 3 | 4 |
| SVM | MRMR | 11 | 5 | 4 | 3 | 2 | 1 | 3 |
| InformationGain | 9 | 7 | 5 | 4 | 3 | 3 | 4 |

Согласно таблице 5 при увеличении количества признаков количество ошибок классификации уменьшается (точность классификации увеличивается), особенно это заметно для наивного байесовского классификатора (nbc) и классификатора машины опорных векторов (svm). Эта тенденция сохраняется вне зависимости от выбранного метода отбора признаков, однако количество ошибок классификации (см. таблица 5) с использованием метода на основе информационного выигрыша превосходит число ошибок, полученных с использованием разработанного авторами алгоритма отбора комбинаций признаков MRMR. Из рисунка 1.2 видно, что в случае использования разработанного алгоритма с увеличением количества признаков точность классификации растет быстрее, чем в случае использования метода информационного выигрыша.

|  |  |
| --- | --- |
|  |  |

Рисунок 1.2 – Графики зависимости точности от количества признаков для набора данных leukemia

На рисунке 1.3 представлены диаграммы зависимости точности классификации от количества признаков для наиболее эффективных классификаторов.

|  |  |
| --- | --- |
|  |  |

Рисунок 1.3 – Графики вида «ящики с усами», отражающие зависимость точности классификаторов от количества используемых признаков для набора данных leukemia. Зависимость оценивалась с использованием процедуры перекрестной проверки (кросс-валидации)

В таблице 6 приведены результаты эффективности классификации для набора данных colon с использованием возрастающего количества признаков от 1 до 30.

Согласно таблице 6 при увеличении количества признаков количество ошибок классификации уменьшается только при отборе признаков с использованием предложенного алгоритма и учета их релевантности, в случае отбора с использованием информационного выигрыша точность не изменяется и даже падает при использовании логистической регрессии. Из рисунка 1.4 видно, что тенденции изменения точности классификации отличаются.

Таблица 6 – Результаты оценки количества ошибок классификаторов на наборе данных colon

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Классификатор |  | Количество признаков | | | | | | |
| Метод отбора | 1 | 3 | 5 | 7 | 10 | 20 | 30 |
| NBС | MRMR | 11 | 11 | 10 | 10 | 10 | 9 | 7 |
| InformationGain | 16 | 15 | 14 | 16 | 15 | 14 | 14 |
| NN | MRMR | 12 | 11 | 12 | 12 | 11 | 8 | 11 |
| InformationGain | 14 | 15 | 16 | 13 | 14 | 15 | 14 |
| MLR | MRMR | 13 | 12 | 13 | 13 | 11 | 11 | 12 |
| InformationGain | 16 | 15 | 15 | 18 | 21 | 22 | 21 |
| SVM | MRMR | 11 | 10 | 11 | 10 | 9 | 9 | 9 |
| InformationGain | 15 | 14 | 15 | 14 | 14 | 15 | 13 |

|  |  |
| --- | --- |
|  |  |

Рисунок 1.4 – Графики зависимости точности от количества признаков для набора данных leukemia

На рисунке 1.5 представлены диаграммы зависимости точности классификации от количества признаков для наиболее эффективных классификаторов с использованием предложенного алгоритма отбора признаков.

|  |  |
| --- | --- |
|  |  |

Рисунок 1.5 – Графики вида «ящики с усами», отражающие зависимость точности классификаторов от количества используемых признаков для набора данных colon. Зависимость оценивалась с использованием процедуры перекрестной проверки (кросс-валидации)

**2 Алгоритм выделения комбинаций биомаркеров на интегрированных источниках данных**

Предложенный алгоритм основан на использовании двух источников данных: данные генной экспрессии и данные молекулярных и функциональных взаимодействий. Интеграция двух и более источников данных для выделения подмножеств признаков, характеризующих сложное заболевание, является одним из актуальных направлений в области компьютерной биологии [8-9]. Поиск модулей с использованием биологической сети позволяет определить группу ген, которые не только дифференциально экспрессированы в случае сложного заболевания, но и расположены близко друг к другу на сети, т.е. связаны между собой функционально, так как биологические сети определяют некоторую функциональную зависимость между генами, белками и другими биологическими единицами. Известно, что сложное заболевание обычно характеризуется нарушениями функционирования ряда биологических путей, а не нарушением работы какого-либо отдельного определенного гена. Поэтому интерес представляет учет функциональной зависимости между генами для определения модулей, которые нарушаются в процессе течения заболевания, и которые могут объединять как гены, активно изменяющие свою экспрессию для большинства случаев заболевания, так и гены, реагирующие только в отдельных случаях заболевания и не выявляемые с использованием стандартных статистических тестов. В ряде имеющихся работ [9] предполагается что одни и те же гены в биологическом пути дифференциально экспрессированы во всех случаях заболевания, что не всегда имеет место в человеческом геноме, где генетические нарушения в отдельных случаях могут различаться в связи с разнородностью генетических типов, тканей, внешних факторов и т.д. В нашем исследовании предложен алгоритм, который позволяет выделить подсети (дисрегулированный путь -ДП) взаимодействующих генов, для которых нарушена регуляция в множестве случаев заболевания, однако в каждом отдельном случае могут быть дисрегулированы различные подмножества генов. Кроме того, для каждой найденной подсети оценивается уровень ее статистической значимости. Поиск оптимальных ДП является NP-полной задачей, поэтому алгоритм представляет собой эвристически подход к поиску субоптимальных решений, которые обладают достаточным уровнем эффективности.

2.1 Краткое описание разработанного алгоритма

Входными данными для алгоритма являются сеть молекулярных/функциональных взаимодействий, которая представляет собой ненаправленный и невзвешенный граф и набор профилей генной экспрессии, который разделены на две группы - случаев и контрольных измерений. Каждый генный профиль включает уровни экспрессии одного пациента для некоторых узлов сети. Некоторые гены не имеют данных экспрессии, так как отсутствовали в ДНК микрочипах. Дисрегулированный путь в этом случае определяется как наименьший связный подграф сети, в котором в каждом случае заболевания дисрегулировано определенное количество генов.

На первом шаге алгоритма для каждого генного профиля случая заболевания определяется подмножество генов дифференциально экспрессированных по отношению к контрольным случаям. Далее целью является определить наименьший подграф, покрывающий по крайней мере генов из каждого такого подмножества, за исключением не более выбросов, для которых количество генов может быть меньшим. Таким образом алгоритм имеет два параметра: – количество генов, нарушенных в ДП для каждого случая, - максимальное количество выбросов (исключений).

Схема работы алгоритма состоит из выполнения следующих шагов:

1. Для каждого гена на основе распределения значений контрольных случаев определяются случаи заболевания, для которых его регуляция нарушена. Для этого, для каждого гена определяется среднее значение и стандартное отклонение его экспрессии в контрольных случаях для подгонки нормального распределения к данным. На основе полученного нормального распределения рассчитываются p-значения экспрессии гена в каждом случае заболевания. Ген считается дифференциально экспрессирующим, если p-значение меньше 0.05 и отношение значения экспрессии в случае заболевания и среднего значения экспрессии контрольных случаев больше или равно 1.4.

2. Осуществляется идентификация подмножества узлов сети, которые являются стартовыми для поиска ДП. Для каждого отдельного узла сети исследуются его r-окрестности. Узлы с минимальной окрестностью, которая содержит удовлетворяющий задаче ДП, составляют стартовое подмножество поиска оптимальных ДП.

3. Для каждого стартового узла пошагово выполняется процедура определения ДП. Поиск начинается с ДП, содержащего один корневой узел. В качестве непокрытых определяются случаи, для которых менее генов содержатся в текущем ДП. На каждой последующей итерации просматриваются узлы, смежные с узлами из ДП и добавляется узел, покрывающий наибольшее число непокрытых случаев.

4. Условием останова алгоритма является количество непокрытых случаев меньшее .

Задачу поиска оптимального ДП можно сформулировать также как задачу теории графов. Для этого введем ряд определений.

Граф представляет сеть молекулярных взаимодействий. – множество смежных вершин (соседей) вершины . Расстояние между вершинами и является длиной кратчайшего пути между и на графе. -окрестностью вершины является множество вершин графа на расстоянии r от . Радиус графа – наименьшее такое что существует вершина, чья -окрестность содержит весь граф или . Поиск в ширину (BFS) на графе представляет собой метод последовательного просмотра отдельных уровней графа, начиная с узла источника .

В обозначениях теории графов задача поиска ДП формулируется следующим образом. Имеется неориентированный граф и набор подмножеств на множестве элементов , где – множество случаев заболевания, и – подмножество случаев, для которых ген дисрегулирован. Дополнительно рассматривается двудольный граф , где для тогда и только тогда, когда . Множество является связным -покрытием ( если образует связный компонент в графе и существует подмножество такое что и для всех , т.е. в выделенном подграфе графа минимальная степень вершины в равна . Интерес представляет поиск наименьшей мощности. Обозначим данную задачу минимизации как .

В случае, если является кликой (полный граф), то каждое множество является связным подграфом и следовательно эквивалентно классической задаче о покрытии множества. Данная задача является NP-полной и для ее решения используется простой жадный приближенный алгоритм с коэффициентом приближения равным [10]. В нашем алгоритме решается задача множественного покрытия, в случае когда граф может не являться полносвязным.

2.2 Программная реализация алгоритма выделения комбинаций биомаркеров на интегрированных источниках данных

2.2.1 Описание программной реализации

Программная реализация алгоритма представляет собой консольное приложение IFM, написанное на языке Java в программной среде разработки IntelliJ IDEA 2019.2. Данное приложение позволяет извлекать информативные функциональные модули или подмножества ген из биологической сети и набора данных генной экспрессии, характеризующего подтипы заболеваний или данные типа «случай-контроль». Данная реализация позволяет создать библиотеку классов и таким образом интегрировать функционал приложения в систему, использующую другой программный язык для разработки как интерфейса, так и входящих в ее состав программных модулей. Основные функции программного приложения следующие

1. функция-парсер для чтения файла, описывающего входные параметры выполнения алгоритма;
2. функции-парсеры для чтения исходных набора данных генной экспрессии и файла, содержащего описание биологической сети (в форматах GRAPHML, SIF и др.);
3. функция предобработки и преобразования непрерывных значений входной матрицы экспрессии генов в бинарное представление. Предобработка состоит в стандартизации значений каждого гена, преобразование состоит в расчете p- значений для значений экспрессий образцов больных тканей на основе оценки распределения значений в здоровых тканях. После этого значения экспрессии приравнивается к 1, если ;
4. функция интегрирования двух наборов данных и формирование встроенного описания сети с атрибутами узлов, содержащими соответствующие бинарные генные профили;
5. функция предварительного отбора стартовых узлов для алгоритма поиска функционального модуля на основе поиска узлов, которые совместно с узлами, находящимися в некоторой окрестности раз покрывают по крайней мере случаев заболевания, причем окрестности этих узлов являются минимальными;
6. функция поиска функционального модуля на основе эвристического алгоритма последовательного расширения модуля, начиная с некоторого корневого узла;
7. функция поиска функционального модуля на основе процедуры поиска кратчайшего пути на графе;
8. функция формирования и сохранения результатов.

Все вышеперечисленные функции реализованы как методы соответствующих программных классов, описанных в таблице 7.

Основным классом программной реализации алгоритма является класс IFMRunner, который используется для вызова функциональности алгоритма. Причем предоставляется возможность вызова алгоритма как для фиксированных значений параметров и , так и для диапазонов значений этих параметров.

Таблица 7 – Основные классы программной реализации алгоритма

|  |  |  |  |
| --- | --- | --- | --- |
| Программный класс | Методы | Описание метода | Описание класса |
| IFMSettings | IFMSettings() | Конструктор класса | Класс для объявления и хранения параметров алгоритма |
| IFMGraph | readGraph | Чтение структуры графа (сети) из файла | Класс для хранения и обработки структуры биологической сети |
| createGraph | Создание графа |
| removeGeneNode | Удаление узла графа |
| getEdgesConnecting | Получение списка ребер, соединяющих узлы графа |
| getNodeList | Получение списка узлов графа |
|  | saveToFile | Сохранение структуры графа в файл |
| IFMData | readData | Чтение данных генной экспрессии | Класс для хранения и обработки данных генной экспрессии |
|  | preProcessData | Предобработка данных. |
|  | saveToFile | Сохранение данных в файл |
| GeneNode | GeneNode | Конструктор класса | Класс для выполнения манипуляций с узлами сети |
| GeneEdge | GeneEdge | Конструктор класса | Класс для выполнения манипуляций с ребрами сети |
| IFMRunner | runSingle | Выполнение запуска алгоритма для фиксированных значений параметров и | Основной класс для вызова функции поиска функциональных модулей (подмножеств ген) с использованием параметров, определенных в классе IFMSettings |
| runMultiple | Выполнение запуска алгоритма для диапазона значений параметров и |
| getResults | Формирует класс IFMResult с результатами работы алгоритма |
| AlgoComputations | Run | Позволяет вызывать несколько вариантов выполнения алгоритма, включая эвристический и с использованием поиска кратчайших путей на графе | Класс для запуска различных вариантов выполнения поиска подмножеств ген на сети |
| IFMResult | getResultsInfoTable | Получение информации о каждой подсети в табличном виде с заголовком | Класс для хранения результатов работы алгоритма |
| Продолжение таблицы 7 | | | |
|  | getAllComputedNodeSets | Получение множества узлов, составляющих подсеть, полученную с использованием алгоритма для фиксированных значений входных параметров |  |
| RunStats | getNumPathways | Возвращает количество найденных модулей | Класс для расчете статистических показателей на основе полученных результатов, включая количество экспрессированных ген в каждом узле сети, и в среднем по функциональному модулю и других |
| getNumNodes | Возвращает количество узлов в модуле |
| getMinNodeHits | Возвращает минимальное количество активных узлов в модуле |
| getMaxNodeHits | Возвращает минимальное количество активных узлов в модуле |
| computeCounts | Рассчитывает статистические показатели |

Входные параметры алгоритма, используемые для запуска программного приложения представлены в таблице 8.

Таблица 8 – Входные параметры алгоритма

|  |  |  |
| --- | --- | --- |
| Название | Значение по умолчанию | Описание |
| graph\_file | sampleNetwork.sif | Путь к файлу, содержащему биологическую сети с связями между генами. Этот файл может быть в форматах SIF или GRAPHML |
| graph\_file\_separator | TAB | Разделить используемый для разделения колонок в файле |
| datasets\_file | datasets\_file.txt | Путь к файлу, содержащему данные генной экспресии |
| datasets\_file\_separator | TAB | Разделить используемый для разделения колонок в файле |
| positive\_file | positive\_list.txt | Путь к файлу со списком положительных ген, т.е. ген, которые имеют высокий приоритет при поиске функциональных модулей по сети |
| negative\_file | negative\_list.txt | Путь к файлу со списком отрицательных ген, т.е. ген, которые имеют низкий приоритет при поиске функциональных модулей по сети |
| results\_folder | results | Путь к директории для записи результатов |
| summary\_file | summary | Путь к файлу с общими результатами анализа |
| pathways\_file | pathways | Путь к файлу с описанием всех выделенных функциональных модулей |
| Продолжение таблицы 8 | | |
| pathways\_stats\_file | pathways\_stats | Путь к файлу с описанием значений статистических показателей выделенных функциональных модулей |
| generate\_summary\_file | TRUE | Генерация файла общих результатов |
| generate\_pathways\_file | TRUE | Генерация файла описания функциональных модулей |
| generate\_pathways\_stats\_file | TRUE | Генерация файла статистических показателей |
| gene\_exceptions | 0 | Количество исключений или ген, не удовлетворяющих условиям алгоритма, соответствует параметру в описании алгоритма. |
| algorithm | heuristic | Стратегия поиска функциональных модулей, (heuristic или shortPath) |
| max\_solutions | 20 | Максимальное количество функциональных модулей |
| treat\_nodes | negative | Определение ген в список отрицательных/положительных в случае отсутствия значений экспрессии |

Данные параметры записываются в файл параметров ifs\_parameters.txt, который используются для вызова исполняемого файла приложения.

2.2.2 Тестирование программной реализации алгоритма выделения комбинаций биомаркеров на интегрированных источниках данных

Для тестирования предложенного алгоритма использовалась биологическая сеть белок-белковых взаимодействий, полученная из базы данных HPRD [11]. HPRD является репозиторием белковых взаимодействий, состоящий из высококачественных курируемых данных, полученных, как с использованием высокопроизводительных технологий, так и с использованием компьютерного моделирования. Биологическая сеть состоит из 7384 генов (узлов сети), обозначенных соответствующими Entrez [12] идентификаторами и 23462 связей или ребер между генами. В качестве данных генной экспрессии использовались данные пациентов с неврологическими заболеваниями, а именно с болезнью Гентингтона (HD). Данное заболевание является дегенеративным неврологическим заболеванием, вызванным дефектом четвертой хромосомы, кодирующей мутированную версию белка гентингтина (HTT). Данный белок вступает во взаимодействие со многими транскрипционными факторами и, следовательно, генная экспрессия может быть нарушена во множестве тканей. Системные исследования данного заболевания представляются перспективными для понимания возможных биологических путей его развития. В результате экспериментов с использованием предложенного алгоритма было выполнено выделение нескольких подсетей или функциональных модулей, содержащих наибольшее количество дифференциально экспрессированных генов среди всех возможных подмодулей и проведен анализ их функциональности. Данные экспрессии генов были получены из архива данных генной экспрессии GEO [13] и включают экспрессию более 2000 ген для 70 пациентов, включая 32 образца контрольных тканей и 38 образца пораженных заболеванием.

В качестве начальных параметров разработанного алгоритма IFM были использованы значения по умолчанию, за исключением количества исключений для ген и количества исключений для случаев заболевания .

Для проведения сравнительного анализа полученных результатов использовался алгоритм jActiveModules [14] и статистический тест. В качестве статистического теста использовался [t-критерий Стьюдента, который был применен к данным генной экспресии без учета биологической сети для выделения наиболее статистически значимых генов. Алгоритм jActiveModules реализован в виде программного плагина системы Cytoscape [15] и также предназначен для поиска дифференциально экспрессированных подсетей в биологической сети.](https://ru.wikipedia.org/wiki/T-%D0%9A%D1%80%D0%B8%D1%82%D0%B5%D1%80%D0%B8%D0%B9_%D0%A1%D1%82%D1%8C%D1%8E%D0%B4%D0%B5%D0%BD%D1%82%D0%B0)

В таблице 9 представлены характеристики функциональных модулей, полученных в результате применения нескольких подходов.

Таблица 9 – Результаты сравнительного анализа нескольких подходов к выделению комбинаций информативных ген, связанных с заболеванием

|  |  |  |  |
| --- | --- | --- | --- |
|  | IFM | jActiveModules | t-test |
| Количество генов | 37 | 282 | 34 |
| Включение HTT | Да | Нет | Нет |
| Включает гены биологическиого пути болезни HD из KEGG [16] | 8 | 4 | 0 |
| Включает гены кальциевого сигнального пути из KEGG | 5 | 10 | 3 |

При увеличении значения до 8 предложенный алгоритм позволяет выделить функциональный модуль большего размера c включением дополнительно 3 генов биологического пути болезни HD и 3 гена сигнального пути. Необходимо отметить, что гены CTNNB1, EGFR, участвующие в биологическом пути болезни HD являются исключениями, т.е. им соответствует более чем случаев заболевания, где они не экспрессируют. Следовательно, предложенный алгоритм способен определить подмножество генов, которые включают редко экспрессирующие гены, являющиеся важными для изучения заболевания, так как соединяют регионы сети, характеризующиеся высоким уровнем дифференциальной экспрессии.

На рисунке 2.1 изображен наибольший функциональный модуль, извлеченный из биологической сети.

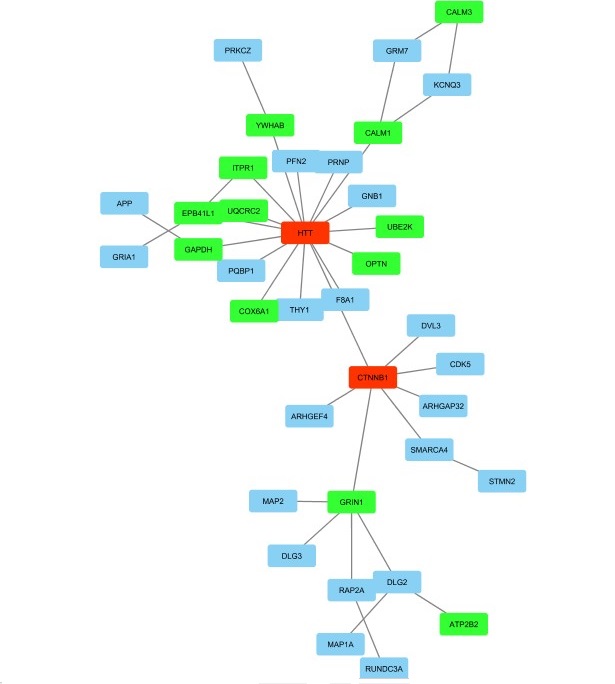


Рисунок 2.1 – Функциональный модуль или подмножество генов, выделенных на биологической сети и характеризующий заболевание HD. Зеленым цветом посечены гены, участвующие в биологическом пути из базы данных KEGG. Красным цветом помечены белок гентингтин и ген CTNNB1, являющийся редко экспрессирующим геном.

Наряду с текстовым файлом, описывающим структуру выделенных модулей предложенный алгоритм возвращает файл описания всех модулей, пример которого представлен на рис. 2.2.

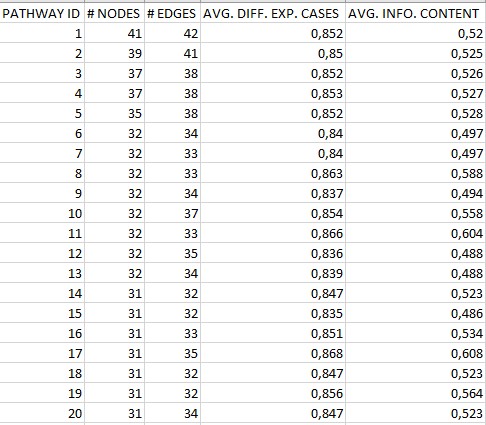


Рисунок 2.2 – Пример выходного файла с описанием характеристик 20 извлеченных модулей.

**3 Алгоритм спектральной кластеризации биомедицинских данных на основе диффузионных карт и сетей**

Многие биомедицинские наборы данных являются данными высокой размерности в контексте числа признаков в векторе данных. При этом в данных наблюдается зачастую зашумленность, но в тоже время присутствует определенная закономерность. По мере роста размерности данных возникает почти универсальная проблема: как можно исследовать данные достаточно объективно, чтобы выявить лежащую в основе кластерную структуру. На сегодняшний день, к сожалению, не хватает устойчивых методов для анализа кластерной структуры биомедицинских данных.

Мы предлагаем алгоритм или метод кластеризации биомедицинских данных высокой размерности, на основе методов спектрального анализа, диффузионных карт и сетей. Спектральные методы основаны на математическом аппарате собственных векторов матриц смежности или аффинности графов данных [17]. Эти методы способны получать гладкое представление данных и устойчивы к шуму и выбросам. Способность представлять исходное множество данных в виде низко размерной структуры данных также делает эти методы привлекательными для применения. В отличие от классических методов уменьшения размерности, таких как метод главных компонент (PCA), эти методы являются нелинейными, что важно, поскольку данные реального мира обычно не лежат на гиперплоскости. Кроме того, они сохраняют локальные структуры в данных, игнорируя расстояния между точками, которые находятся далеко друг от друга, что, как правило, не имеет смысла в многомерных данных. Эти подходы применяются в области машинного обучения, обработки сигналов и интеллектуального анализа данных. Данные методы особенно актуальны и эффективны для анализа биомедицинских данных [18-20].

3.1 Диффузионные карты

Кластеризация и низко размерное представление многомерных наборов данных являются важными проблемами во многих различных областях. В последние годы были разработаны различные спектральные методы для решения этих задач, основанные на собственных векторах матриц смежности графов данных [20-22], среди которых особо выделяются диффузионные карты.

*Диффузионная карта* – это алгоритм спектрального отображения исходного множества многомерных данных в низко размерное многообразие, позволяющий определить структуру исходного множества данных и снизить его размерность [22]. Он вычисляет семейство вложений исходного многомерного набора данных в низко размерное евклидово многообразие на основе собственных векторов и собственных значений оператора диффузии и матрицы смежности графа, построенного для исходного множества данных. В отличие от методов уменьшения линейной размерности, таких как анализ главных компонентов PCA и многомерное масштабирование MDS, диффузионные карты являются частью семейства нелинейных методов спектрального анализа данных, направленных на обнаружение низко размерного ортогонального базиса анализируемых данных. Интегрируя локальные сходства данных на исходной метрике, диффузионные карты дают глобальное описание набора данных, при том устраняя шумы и выбросы. Данный метод эффективно использовался в различных приложениях для обработки биомедицинских данных, сигналов, обработки изображений и машинного обучения [23-25]. Ниже опишем основные этапы построения диффузионной карты и кластеризации.

3.1.1 Построение графа и весовой матрицы исходного пространства данных

Для многомерного множества данных – *n-*мерный вектор анализируемого нормированного векторного пространства данных размерности *n*,строим конечный взвешенный граф c точками данных в качестве узлов, где матрица весов определяет меру сходства или расстояние между двумя точками данных и удовлетворяет следующим условиям:

- симметричности: ;

- неотрицательности: .

В качестве графа применяем нормализованный граф Лапласиана, а в качестве матрицы весов или схожести используем ядро Гауссиана или же ядро радиальной базисной функции (RBF – radial bases function)

, (3.1)

где – глобальный параметр масштабирования.

3.1.2 Нормализация весовой матрицы графа исходного пространства данных

Граф *G* с весами *W* представляет наши знания о локальной геометрии множества. Выполняем нормализацию весовой матрицы графа данных , чтобы получить приближение оператора Лапласа – Бельтрами на исходном пространстве данных, в силу чего вложение в многообразие не будет зависеть от распределения точек [25]. Нормализуем ядро по степени каждой вершины графа , соответствующей точке в пространстве данных

. (3.2)

Для получения вложения в многообразие во многих работах предлагается использовать собственные вектора нормализованной задачи собственных значений или следующую отсюда нормализованную матрицу

. (3.3)

Здесь также применимы другие близкие к этому варианты получения нормализованной матрицы, которые подробно исследованы в [25]. Мы предполагаем также использование следующего способа нормализации

. (3.4)

Использование первых нескольких собственных векторов нормализованной матрицы *P* в качестве низко размерного базиса многообразия , куда отображается исходное многомерное пространство данных, было обосновано в работе [21]. Они показали, что при равномерной выборке данных из низко размерного многообразия первые несколько собственных векторов матрицы *P* являются дискретными приближениями собственных функций оператора Лапласа-Бельтрами на многообразии, обеспечивая тем самым математическое обоснование их использования в этом случае. Заметим, что компактное вложение многообразия в Гильбертово пространство через собственные функции оператора Лапласа-Бельтрами было предложено в дифференциальной геометрии и использовано для определения расстояний между многообразиями [26].

3.1.3 Построение низко размерного многообразия , в которое вкладывается исходное многомерное пространство данных

Стохастическая матрица *P* удовлетворяет следующим условиям:

* ;

Поэтому нормализованная стохастическая матрица *P* может рассматриваться как матрица переходов цепи Маркова на наборе данных *X*. Собственное разложение *P* дает последовательность

биортогональных левых и правых собственные векторов и соответственно и последовательность собственных значений: 1 = | *λ*0 | ≥ | *λ*1 | ≥ .... Тогда *t* шагов цепи Маркова можно рассчитать, как

(3.5)

В этом случае диффузионное расстояние между двумя точками из исходного *n-*мерного пространства определяется как

, (3.6)

где – стационарное вероятностное распределение на графе данных.

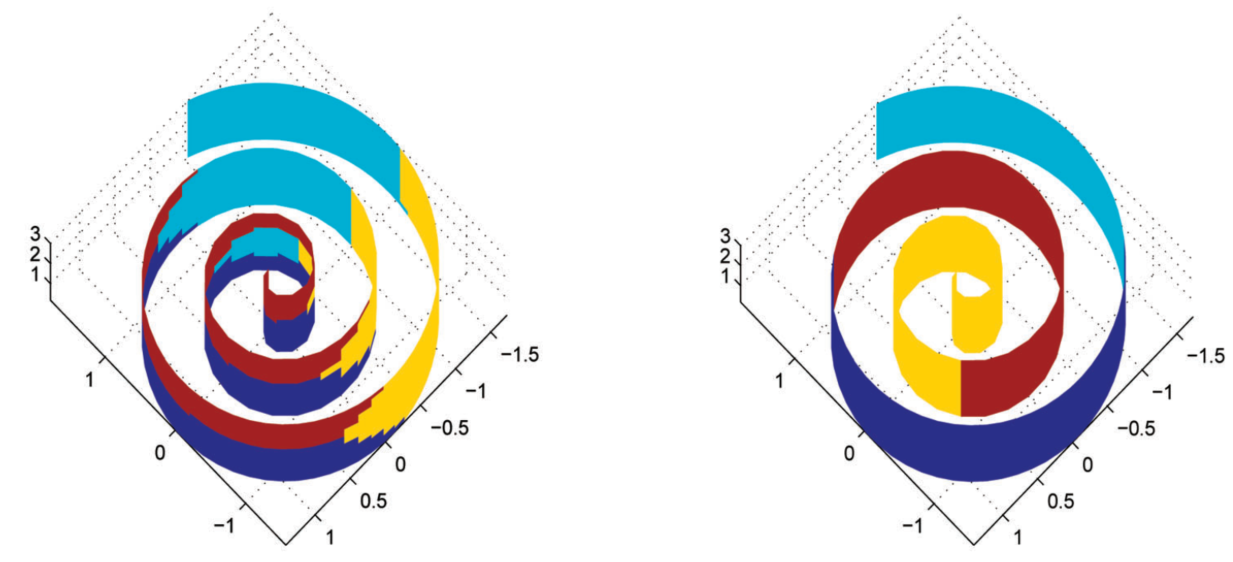
Полученная диффузионная метрика является устойчивой к шумам в исходных данных, поскольку расстояние между двумя точками зависит от всех возможных путей длины *t* между точками. Благодаря затуханию спектра матрицы *P* расстояние диффузии может быть аппроксимировано с использованием только первых *d* собственных векторов. Формула (3.6) подразумевает, что отображение, диффузионная карта, может быть определено между исходным *n-*мерным пространством данных *X* и собственными векторами . Сохраняя только первые собственные векторы, отображение встраивает исходный *n-*мерный набор данных *X* в евклидово пространство , где расстояние диффузии равно евклидову расстоянию в этом новом вложении:

. (3.7)

3.1.4 Кластеризация данных в итоговом евклидовом пространстве

Известно, что биомедицинские данные обладают сильно нелинейными структурами. В таких случаях линейные методы не будут очень эффективными для представления данных. Однако с помощью диффузионных координат можно узнать внутреннюю геометрию набора данных, а затем спроецировать точки данных в нелинейное координатное пространство с диффузионной метрикой, как показано выше. Показано, что в диффузионном евклидовом пространстве можно использовать классические геометрические алгоритмы [23]. Мы предполагаем возможное применение алгоритма *k*-средних или алгоритма дискретизации [27].

Чтобы проиллюстрировать применимость, например, алгоритма *k-*средних на полученном диффузионном вложении, рассмотрим пример кластеризации данных с нелинейной структурой, отображаемой геометрически трехмерной поверхностью известного «Швейцарского ролла» (рисунок 3.1). В этой исходной системе координат глобальные внешние расстояния, такие как евклидово расстояние, часто не имеют смысла, поскольку они не содержат никакой информации о структуре или форме набора данных. Например, если мы запускаем алгоритм *k*-средних для кластеризации с , полученные кластеры не отражают естественную геометрию множества. Как показано на рисунке 3.1-а, существует некоторая «утечка» между различными частями спирали из-за выпуклости кластеров *k*-средних в пространстве данных. В качестве сравнения мы также показываем на рисунке 3.1-б результат работы алгоритма *k*-средних в диффузионном пространстве и в этом случае мы получаем значимые кластеры, которые учитывают внутреннюю геометрию набора данных.



|  |  |
| --- | --- |
| а) | б) |

Рисунок 3.1 – Кластеризация алгоритмом *k*-средних (*k=*4) данных с нелинейной геометрической структурой «Швейцарский ролл»: а) результаты кластеризации в исходном пространстве данных, б) результаты кластеризации в диффузионном многообразии.

3.2 Обучение диффузионных отображений на основе модели диффузионных сетей

3.2.1 Искусственные нейронные сети

*Искусственной нейронной сетью* называется система, состоящая из совокупности связанных между собой по типу узлов направленного графа элементарных процессоров, называемых *искусственными или формальными нейронами*, и способная генерировать выходную информацию в ответ на входное воздействие.

Искусственные нейронные сети, в частности, глубокие сети успешно применяются к различным задачам, таким как регрессия для изучения функции по заданному набору данных, классификация, обработка изображений и сигналов и т.д. [28]. Задача, которую выполняет сеть, определяется выходным уровнем и функцией стоимости, минимизированной по сети. В обучении с учителем цель состоит в том, чтобы предсказать функцию или метки на входных данных. Функция стоимости сети состоит из функции потерь, и обычно добавляется член регуляризации веса, чтобы предотвратить переобучение.

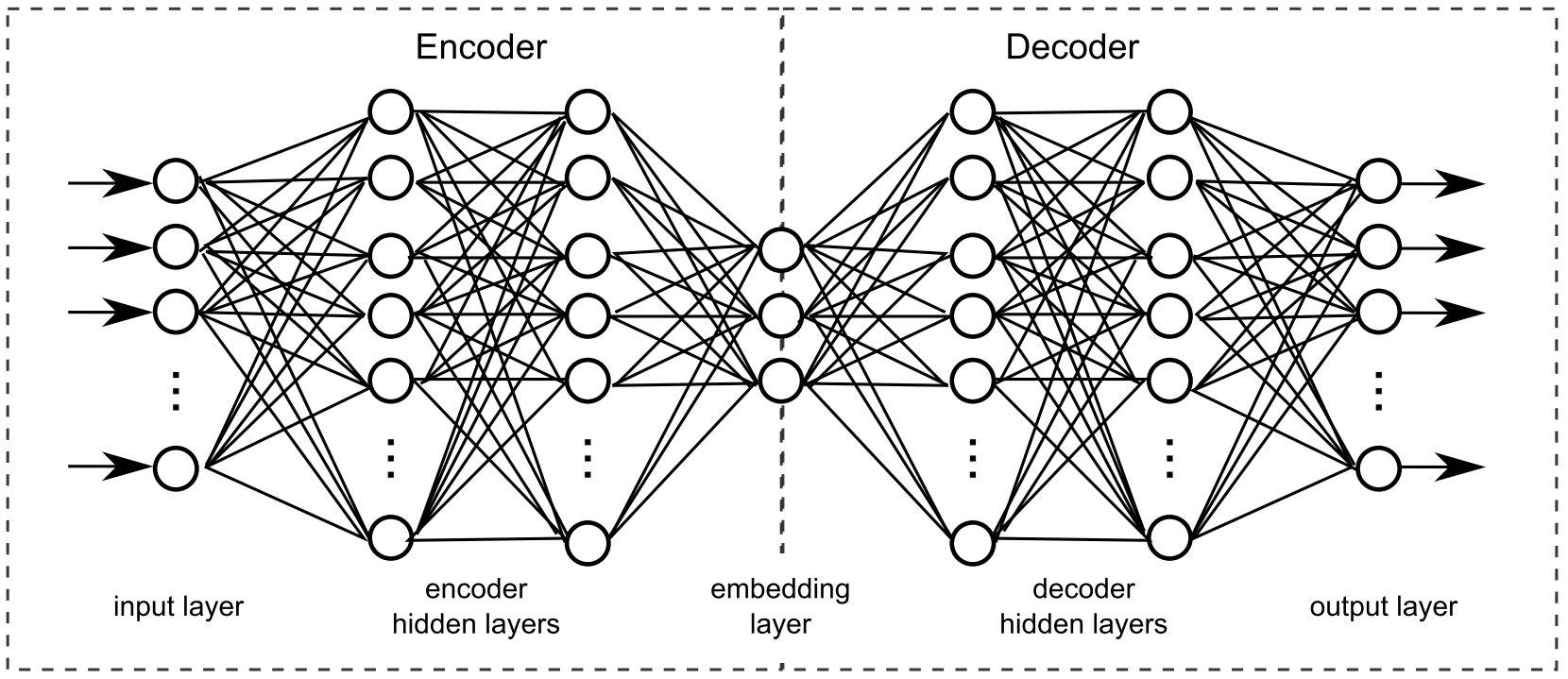
Нейронные сети, в том числе модели сетей глубокого обучения, могут также применяться в случаях обучения без учителя. К такого рода нейронным сетям относится такая модель глубокого обучения как автокодировщик – autoencoder. Автокодировщик состоит из двух частей функции кодировщика (encoder) *f*(·) и функции декодировщика (decoder) *g*(·), где размерность функции *f* кодировщика, как правило, меньше размерности входных данных (рисунок 3.2).

Рисунок 3.2 – Модель и структура автокодировщика

Цель модели автокодировщика с одной стороны – получить представление исходного множества данных X меньшей размерности на этапе кодировки, с другой стороны – поскольку мы не имеем эталонных значений, то в качестве таковых на этапе декодирования мы используем сами входные вектора X и обучаем модель автокодировщика с помощью методов обратного распространения ошибки обучения с учителем. Восстановление входного вектора задается путем последовательного объединения или суперпозиции функции декодировщика на функцию кодировщика: . При этом автокодировщик обучается минимизировать ошибки восстановления входных векторов , т.е. аппроксимировать тождественную функцию . Автокодировщики были популяризированы Хинтоном [29]. Классические автокодировщики состоят из одного скрытого слоя, понижающего размерность [30], но были предложены глубокие автокодировщики со множеством скрытых слоев (рисунок 3.2), применительно к задачам классификации, шумоподавления, анализа изображений, улучшения и распознавания речи [29].

3.2.2 Диффузионные сети

В результате выполнения первых трех шагов алгоритма кластеризации, изложенного в пункте 3.1, мы строим для входного множества данных *n*-мерных векторов *X* диффузионную карту , которая отображает исходное многомерное пространство данных в гладкое, *d-*размерное Риманово многообразие в пространстве . Наша задача состоит в том, чтобы построить обученную нейросетевую модель для дальнейшего анализа векторов *X* тестовых выборок, пропуская при этом этап построения диффузионной карты для тестовых наборов данных и заменяя его обученной нейросетевой моделью.

Предлагается решение данной задачи на основе нейросетевой модели глубокого обучения, поэтапного обучения нейронных сетей кодировщика и декодировщика, и объединения их в глубокий многослойный автокодировщик [31]. Таким образом, вместо обучения модели автокодировщика как целой сети, мы включаем построенную диффузионную карту в качестве эталонного набора векторов для встроенного низкоразмерного слоя в поэтапную процедуру обучения автокодировщика. Кодировщик и декодировщик оба являются многосвязными нейронными сетями прямого распространения, состоящими из *L* уровней, выходной уровень которых является регрессионным уровнем. Кодировщик анализирует отображение из пространства данных векторов *X* в диффузионное пространство, минимизируя квадратичную ошибку между выходом кодировщика и эталонными векторами диффузионной карты . Кроме того, накладывается новое ограничение для сохранения свойств вложения. Декодировщик изучает обратное отображение из низкоразмерного диффузионного пространства обратно в многомерное пространство данных векторов *X*, решая проблему аппроксимации тождественной функции . Объединяя кодировщик и декодировщик через операцию суперпозиции, мы получаем автокодировщик, внутренний слой которого вычисляет диффузионную карту исходных данных. Такой автокодировщик называется *диффузионной сетью* [31] и может также использоваться как для обнаружения выбросов, так и для шумоподавления.

3.3 Программная реализация алгоритма спектральной кластеризации биомедицинских данных

3.3.1 Описание программного модуля

Программный модуль, реализующий алгоритма спектральной кластеризации биомедицинских данных реализован как набор функций в среде RStudio на языке программирования R [3]. Программный модуль включает функции для предобработки исходных данных, и тем самым сокращения размерности исходного пространства признаков-генов, выполнения спектрального кластерного анализа с автоматическим определением числа кластеров и оптимальной подвыборки данных генной экспрессии в контексте решения задачи совпадения меток кластеров с метками эталонных классов, валидации результатов кластерного анализа, трехмерной визуализации результатов спектрального кластерного анализа (таблица 10).

Таблица 10 – Описание основных функций программного модуля

|  |  |
| --- | --- |
| Функций | Описание |
| DataGE.Preprocess | Выполняет исключение генов с отсутствием названия в анализируемом наборе данных генной экспрессии; исключение генов у которых более 25% значений NA; заполнение пропусков методом ближайших соседей (kNN); ранжирование признаков-генов согласно эталонному классу из клинического набора данных на основе *p*-значений и методов множественных сравнений и тестирования коррекций Бенджамина-Хохберга [32]; исключение признаков, *p*-значения которых превышают порог *P*; определение и удаление генов с высоким уровнем корреляции; нормализация значений полученного набора данных генной экспрессии на отрезок [0;1]. |
| Diff.Clust | Выполняет спектральную кластеризацию на основе диффузионных отображений с Гауссовым ядром (пункт 3.1): построение весовой матрицы расстояний графа исходного пространства данных; нормализация построенной матрицы; построение диффузионной карты, включая Лапласиан и систему собственных значений и векторов; автоматическое определение числа кластеров методом максимизации разрыва между значениями *Z* статистик собственных векторов в соответствии с алгоритмом спектральной кластеризации [33] для Гауссовых кластеров; итоговая кластеризация методом GMM [34]. Выполняется циклический перебор подвыборок данных генной экспрессии, включающий спектральную кластеризацию подвыборки, расчет индекса совпадения меток кластеров с эталонными метками классов из набора клинических данных. |
| Val.Data | Выполняет валидацию результатов спектрального кластерного анализа путем вычисления значений 23 показателей качества кластеризации для полученной оптимальной подвыборки данных генной экспрессии, в результате которого по «правилу большинства» автоматически определяется число кластеров. |
| Vis.Data | Выполняет трехмерную визуализацию полученных результатов кластеризации *L*-мерного пространства данных генной экспрессии (Гауссовых кластеров) на основе t-SNE вложения в трехмерное пространство [35] и в пространство первых трех признаков-генов. |

3.3.2 Тестирование программного модуля

Для тестирования работы функций и сравнительного анализа различных подходов к отбору признаков были проведены эксперименты на реальных данных генной экспрессии и клинических данных из базы данных TCGA [36].

Биологические данные представляют собой набор данных генной экспрессии по раку груди, в котором объекты-пациенты классифицированы по данным из клинических наблюдений (таблица 11). Данные генной экспрессии делятся по признаку Class\_meta на следующие два класса: 1 – выживаемость без метастаз как первое событие, 2 – были метастазы как первое событие.

В экспериментах по кластеризации данных генной экспрессии мы будем сравнивать результаты кластеризации на совпадение с метками эталонных классов по следующему интуитивно понятному индексу:

, (3.8)

где – индекс совпадения меток кластеризации и эталонных меток класса 1, равный отношению числа совпавших меток кластеризации и эталонных меток класса 1 к мощности класса 1; – индекс совпадения меток кластеризации и эталонных меток класса 2, равный отношению числа совпавших меток кластеризации и эталонных меток класса 2 к мощности класса 2.

Таблица 11 – Описание реального набора данных

|  |  |  |  |
| --- | --- | --- | --- |
| Набор данных | Количество объектов | Количество признаков | Количество и состав классов |
| Рак груди (Entrez) | 295 | 24481 | 2 (1 – выживаемость без метастаз как первое событие, 2 – были метастазы как первое событие) |

Общая схема эксперимента состоит из выполнения следующих шагов.

1. Предобработка данных генной экспрессии функцией DataGE.Preprocess:
   1. исключение генов с отсутствием названия в базе данных Entrez;
   2. исключение генов, у которых более 25% значений NA;
   3. заполнение пропусков методом ближайших соседей (kNN);
   4. ранжирование признаков-генов согласно классу Class\_meta на основе *p-*значений и методов коррекций на множественные сравнения Бенджамина-Хохберга;
   5. исключение признаков, *p-*значения которых превышают экспериментально определенный порог 0.05;
   6. определение и удаление генов с высоким уровнем корреляции (>0.75);
   7. нормализация значений полученного набора данных генной экспрессии на отрезок [0;1].

Конечной целью выполнения предобработки данных является выделение признаков-генов, которые существенно влияют на соответствие эталонным классам, и, таким образом сокращение размерности пространства признаков-генов. В результате выполнения пунктов 1)-3) предобработки данных размерность пространства признаков-генов сократилась с 24481 до 13100. Выполнение же пунктов 3)-6) сократило размерность признаков-генов до 270. Таким образом, на вход функции кластеризации на основе диффузионных отображений Diff.Clust подается набор данных из 295 объектов, характеризующихся профилями 27 генов.

1. Выполнение кластерного анализа, полученного сокращенного оптимизированного набора данных генной экспрессии функцией спектральной кластеризации на основе диффузионных отображений Diff.Clust.

Циклическое выполнение следующих шагов:

* 1. формирование подвыборки данных генной экспрессии размерностью *L*x 295, где – число первых признаков-генов в ранжированном наборе данных;
  2. диффузионная (спектральная) кластеризация полученной подвыборки данных генной экспрессии с автоматическим определением количества кластеров;
  3. вычисление индекса совпадения (3.8) результатов кластеризации с эталонными классами.

В результате выполнения данного цикла определяется максимальное значение индекса , которое в свою очередь определяет соответствующую ему оптимальную подвыборку данных и соответствующую ей оптимальную кластерную разметку объектов-пациентов. Косвенным критерием оптимального результата в данном случае также является совпадение автоматически определенного числа кластеров с числом классов, то есть со значением 2.

1. Валидация результатов кластерного анализа функцией Val.Data путем вычисления значений 23 показателей качества кластеризации для полученной оптимальной подвыборки данных генной экспрессии, в результате которого по «правилу большинства» автоматически определяется оптимальное число кластеров.
2. Трехмерная визуализация функцией Vis.Data полученных результатов кластеризации *L* -мерного пространства данных генной экспрессии (Гауссовых кластеров) на основе t-SNE вложения в трехмерное пространство и в пространство первых трех признаков-генов.

Функция кластерного анализа Diff.Clust на шаге 3 выполняет перебор подвыборок данных размерности *L*x 295. При этом *L –* число признаков-генов в подвыборке меняется на отрезке [10;270] с шагом 5. Начиная с выборки с числом генов 10 на каждой последующей *i-*й итерации цикла рассматривается подвыборка с генов, для которой выполняется диффузионная кластеризация с автоматическим определением числа кластеров *K* и расчётом индекса совпадения (3.8) результатов кластеризации с эталонными классами. В итоге выполнения данного перебора определяется подвыборка данных размерности (*L=*50)x 295, для которой индекс совпадения является максимальным среди проанализированных подвыборок. Для полученной подвыборки данных оптимальное число кластеров *K=*2, что совпадает с числом эталонных классов (рисунок 3.3).

|  |  |
| --- | --- |
|  |  |

Рисунок 3.3 – Результаты определения оптимального числа кластеров по числу собственных векторов в диффузионном пространстве по критерию максимального разрыва в собственных значениях собственных векторов

Выполнение валидации функцией Val.Data путем вычисления значений 23 показателей качества кластеризации для полученной оптимальной подвыборки данных генной экспрессии по «правилу большинства» на четвертом шаге дает такой же результат как и функция кластерного анализа *K=*2 (рисунок 3.4).

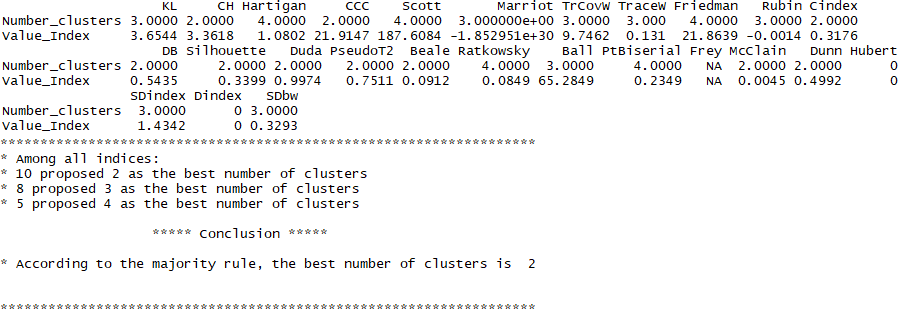


Рисунок 3.4 – Результаты валидации оптимального числа кластеров на основе вычисления 23 индексов качества кластеризации

В результате выполнения пункта 4 получаем трехмерную визуализацию полученных результатов кластеризации 50-мерного пространства данных генной экспрессии (Гауссовых кластеров) в трехмерном пространстве t-SNE вложения в (рисунок 3.5). Из рисунка 3.5 видно, что полученные кластеры достаточно компактны, что также качественно подтверждает хороший результат кластеризации.

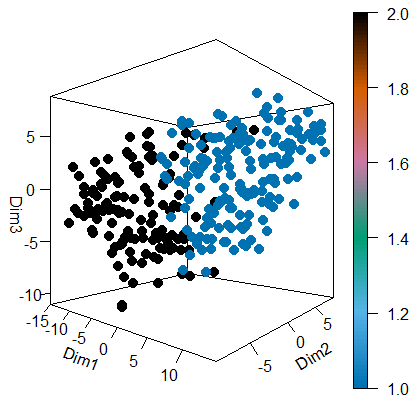


Рисунок 3.5 – Трехмерная визуализация Гауссовых кластеров в трехмерном пространстве t-SNE вложения для полученной подвыборки

Трехмерная визуализация полученных результатов кластеризации 50-мерного пространства данных генной экспрессии в трехмерном пространстве первых трех признаков-генов показана на рисунке 3.6 и качественно свидетельствует о хорошем результате кластеризации. Визуализация выполнялась в пространстве первых трех компонент полученной подвыборки, так как эти компоненты имеют наименьшие *p-*значения.

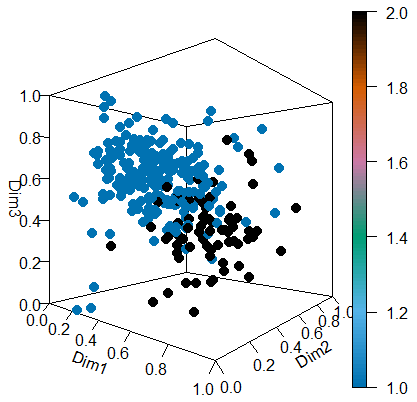


Рисунок 3.6 – Трехмерная визуализация Гауссовых кластеров в пространстве первых трех признаков-генов с наименьшими *p-*значениями

**4 Разработка гибридной модели классификации на основе ансамбля классификаторов**

В связи с тем, что различные источники биологической информации характеризуют различные изменения, происходящие в организме на клеточном уровне при развитии сложного заболевания, то предполагается, что их объединении позволит повысить точность диагностики подтипа заболевания, надежность прогноза развития заболевания и отклика на терапию. Кроме того, объединение разнородных данных позволит обнаружить взаимосвязи между различными биомедицинскими сущностями (гены, белки, метаболиты и т.д.), напрямую связанными с развитием заболевания, компенсировать зашумленность и ошибки, имеющиеся в отдельных источниках данных и тем самым получить более достоверные результаты. Общей проблемой при решении данной задачи является способ комбинации информации из различных источников данных. В нашем исследовании интерес представляют методы построения классификаторов на различных источниках многомерных данных, которые, как правило, имеют разнородное представление. Следовательно, задачей является унификация этого представления, определение базового классификатора, построение классификационных моделей на каждом источнике данных и выбор способов комбинации прогнозных значений, полученных с использованием построенных моделей.

Нами разработана общая схема гибридной модели классификации, которая позволяет объединить несколько источников биологической информации о пациентах с целью построения классификационной модели, позволяющей диагностировать подтипы сложных заболеваний, характеризующихся генетическими нарушениями. Предлагаемая гибридная модель представляет собой классификационный ансамбль со следующими отличительными особенностями:

1. унифицированное представление информации из различных источников данных путем построения матрицы объект-объектных расстояний с использованием различных ядерных функций (функций плотности), включая гауссову, полиномиальную функцию, скалярное произведение векторов и т.д.;
2. выполнение процедуры отбора классификационных признаков для каждого отдельного источника данных;
3. построение базового или индивидуального классификатора гибридной модели, который может представлять собой как единичный классификатор, так и ансамбль классификаторов, построенный на одном источнике данных;
4. реализация нескольких способов интеграции отдельных классификаторов модели;
5. анализ информативности отдельных классификаторов с использованием оценки их весовых коэффициентов.

Общая схема гибридной модели представлена на рисунке 4.1.

Разнородные источникиданных

Источник данных 1

Источник данных 2

Источник данных *R*



Классификатор 1

Классификатор 2

Классификатор R



Агрегирование выходов

Класс

объекта

Рисунок 4.1 ‑ Схема гибридной модели классификации

На рисунке 4.2 представлена схема индивидуального классификатора гибридной модели в случае использовании для его построения процедуры бэггинга.

Источник данных M

Подвыборка данных/признаков 1

Подвыборка данных/признаков 2

Подвыборка данных/признаков N



Классификатор 1

Классификатор 2

Классификатор N



Агрегирование выходов

Класс

объекта

Рисунок 4.2 ‑ Схема индивидуального классификатора, представленного в виде ансамбля

В области машинного обучения исследования в области ансамблей классификаторов имеют достаточно богатую историю и направлены в основном на построение ансамбля на одном наборе данных с целью повышения точности классификации [37]. Теоретическим обоснованием повышения точности классификации с использованием ансамбля является теорема Кондорсье [38]. Согласно этой теореме для задачи бинарной классификации и  базовых классификаторов, ошибка которых ниже 0.5, ансамбль большинством голосов имеет ошибку ниже, чем отдельный классификатор в том случае, если ошибки отдельных членов ансамбля не коррелированы. Например, если мы имеем 21 классификатор, и вероятность ошибки для каждого базового классификатора равна , причем ошибки являются независимыми, вероятность ошибки ансамбля большинством голосов рассчитывается с использованием биномиального распределения, для случая, когда более чем  классификаторов дали ошибочный результат

(4.1)

Однако важной составляющей получения такого результата является независимость отдельных базовых классификаторов, иначе нельзя гарантировать повышение точности классификации. Таким образом, важной составляющей построения ансамблей классификаторов является поиск компромисса между точностью и независимостью базовых классификаторов, так как более точные классификаторы, как правило, являются более зависимыми [39]. С этой точки зрения ансамбли, построенные на разнородных источниках данных, могут достаточно гарантированно повысить точность классификации и прогноза подтипов сложных заболеваний, так как каждый из источников данных описывает изучаемый организм в различных плоскостях: данные генной экспрессии, РНК-секвенирования, метаболические данных, данные числа копий генов, а также клинические данные.

4.1 Разнородные источники биологических данных

В качестве сложных заболеваний для исследований, проводимых в рамках проекта были выбраны онкологические заболевания. Описано более 200 форм онкологических заболеваний и каждый тип заболевания имеет отличающийся молекулярный профиль, что в свою очередь требует применения индивидуальных терапевтических мероприятий. Для каждого типа заболевания характерна уникальная комбинация соматических мутаций, числа копий генов, профилей экспрессии генов, различных эпигенетический вариаций. Таким образом необходимость повышения точности диагностики и лечения напрямую связана с необходимостью понимания генетических изменения в раковых клетках. Последние достижения высокопроизводительных технологий позволили расшифровать геном человека и продвинуть исследования, позволяющие понять источник поломок в организме на клеточном уровне. Наиболее важным проектом, направленным на понимание генетики онкологических заболеваний является Атлас ракового генома (TCGA)  целью которого является систематизация данных о [генетических мутациях](https://ru.wikipedia.org/wiki/%D0%9C%D1%83%D1%82%D0%B0%D1%86%D0%B8%D1%8F), приводящих к возникновению [рака](https://ru.wikipedia.org/wiki/%D0%97%D0%BB%D0%BE%D0%BA%D0%B0%D1%87%D0%B5%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D0%B0%D1%8F_%D0%BE%D0%BF%D1%83%D1%85%D0%BE%D0%BB%D1%8C) [[36]](https://ru.wikipedia.org/wiki/%D0%90%D1%82%D0%BB%D0%B0%D1%81_%D1%80%D0%B0%D0%BA%D0%BE%D0%B2%D0%BE%D0%B3%D0%BE_%D0%B3%D0%B5%D0%BD%D0%BE%D0%BC%D0%B0#cite_note-1). Данный проект на базе открытых архивов, как например TCGA портал, предоставляет для открытого доступа данные по более чем 30 онкологическим заболеваниям человека. В нашем исследовании используются различные TCGA данные, а именно

1. Данные РНК секвенирования (RNAseq), которые предоставляют разноплановую биологическую информацию, включая экспрессию генов, информацию о гибридных генах, полученных слиянием двух отдельных генов и т.д.
2. Данные секвенирования экзонов, представляющие собой данные экспресии кодирующих участков ДНК.
3. Данные микроРНК секвенирования. МикроРНК являются специфическими короткими, некодирующими РНК, которые регулируют сотни ген внутри или между разнообразными сигнальными путями. МикроРНК секвенирование определяет соответствующие тканеспецифичные профили экспрессии, микроРНК изоформы, связанные с заболеванием.
4. Данные ДНК секвенирования описывают вариации на уровне ДНК, включая вставки, делеции, полиморфизм также как и вариации числа копий ген, частоту мутаций и т.д.
5. Данные однонуклеотидных мутаций (SNP).
6. Данные профилей ДНК метилирования, предоставляющие информацию о эпигенетических изменениях в геноме.
7. Данные широкомасштабных профилей экспрессии белков, содержащие информацию об экспресии и концентрации белков.

На рисунке 4.3 представлен пример упрощенной схемы объединения двух источников данных для построения классификационной модели или модели диагностики сложного заболевания.

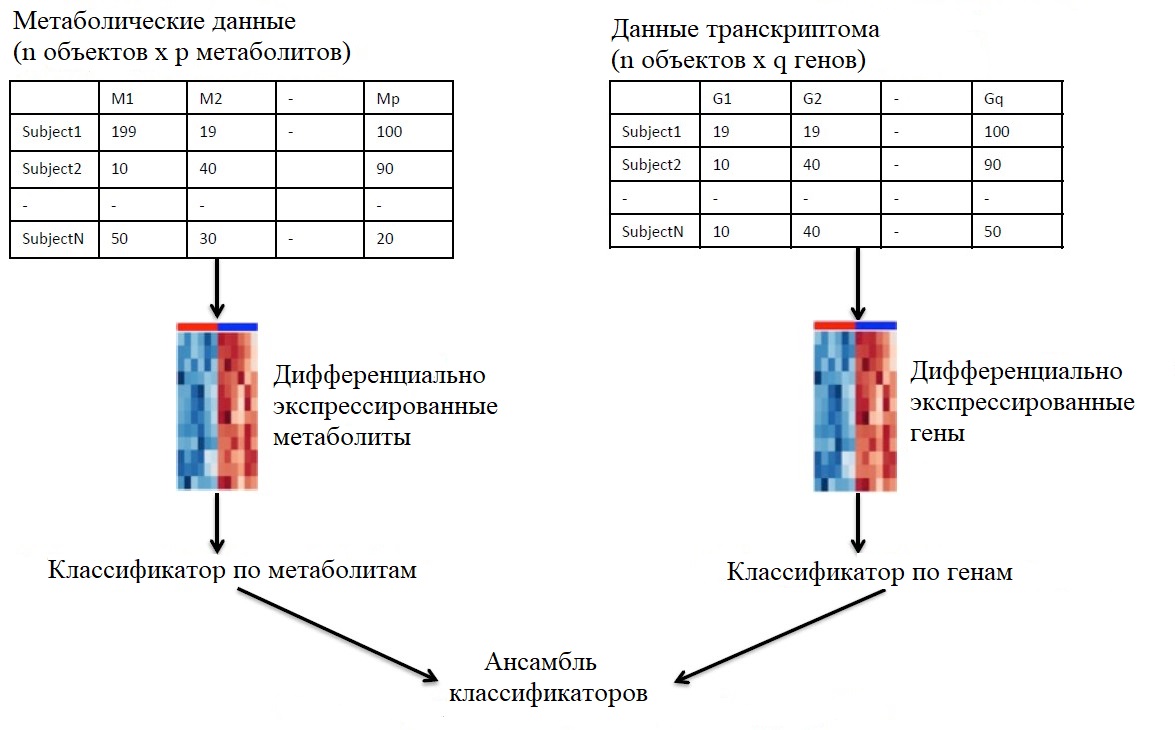


Рисунок 4.3 – Упрощенная схема объединения двух источников данных в ансамбль

4.2. Различные схемы интеграции разнородных данных

В нашем исследовании под обучением ансамбля классификаторов понимается процедура обучения конечного набора базовых классификаторов, частные решения которых затем объединяются с целью формирования результирующего классификационного решения на основе агрегированного классификатора. Существуют различные подходы и к правилам объединения частных классификаторов в ансамбль и к стратегиям формирования результирующего классификационного решения.

Обоснованием применения ансамбля классификаторов является то, что он предоставляет возможность интеграции информации не на уровне данных, которые имеют разнородное представление, а на уровне принятия решений, объединяя выходы отдельных классификаторов, каждый из которых построен на отдельном источнике данных. Каждый из типов данных содержит отличительные характеристики объекта исследования, которые дополняя друг друга, позволяют повысить качество результатов предсказания. Каждый источник данных используется для обучения базового классификатора.

Пусть каждый из типов биомолекулярных данных  характеризуются различными признаками , где  ‑ количество имеющихся источников информации. Таким образом, объект данных  характеризуется различными множествами признаков

, (4.2)

где  представляет собой данные, соответствующие признакам  специфического набора данных . Классификатор обучается на данных  с целью построения функции , которая оценивает степень «поддержки» того, что данный объект  относится к классу . Выход классификатора представляет собой ‑мерный вектор , где  ‑ количество классов. В большинстве классификаторов  является оценкой апостериорной вероятности . Комбинация классификаторов заключается в комбинировании выходов  классификаторов  для получения “мягкой” метки класса для . В случае, когда для объекта  требуется точная метка, используется правило максимума принадлежности: присвоить объекту  класс  если

 (4.3)

Выходы  базовых классификаторов, каждый из которых обучен на отдельном источнике данных, комбинируются с использованием некоторой функции  для получения общей величины «поддержки»  (вероятности) для класса :

. (4.4)

В качестве функции  может быть использованы функции максимума, минимума, среднего значения, взвешенного среднего значения, произведения.

Для метода большинством голосов метка класса для объекта  соответствует классу, который представлен большинством точных меток класса, полученных согласно выходам   классификаторов.

При использовании правила взвешенного среднего выражение для расчета значения  класса  для объекта  следующее

. (4.5)

Весовые коэффициенты при этом могут быть рассчитаны следующими двумя способами:

. (4.6)

В обоих случаях для оценки весовых коэффициентов  используется *F*-мера , которая является гармоническим средним между точностью и чувствительностью результатов классификации. В отличие от классической точности классификатора данная мера позволяет учесть дисбаланс между подтипами заболеваний или вариантом случай/контроль. Значение  оценивается для каждого базового классификатора с использованием процедуры «внутренней» перекрестной проверки на обучающей выборке. Результирующая метка класса  выбирается согласно оценке вероятности  в (4.3) следующим образом

, (4.7)

где значение 1 соответствует положительному предсказанию для класса , а 0 – отрицательному. В качестве порога  значения обычно выбирается значение 0.5 (если  является оценкой вероятности). Таким образом, в такой постановке объект  может быть отнесен более чем к одному классу, что позволит в наших дальнейших исследованиях рассматривать задачу предсказания подтипа заболевания как задачу многоклассовой классификации.

Неверные предсказания класса (подтипа) заболевания для ряда объектов данных, полученные с использованием базовых классификаторов могут быть причиной того, что некоторые типы биомолекулярных данных могут быть более информативны для определенных классов и совершенно неинформативны для других классов. Для того, чтобы принять во внимание систематические неверные предсказания определенных базовых классификаторов для их комбинации предлагается использовать метод решающих шаблонов [40]. В этом методе решающий профиль  для объекта  представляет собой матрицу, состоящую из элементов , которые определяют уровень «поддержки» ‑ым классификатором класса . Решающие шаблоны  представляют собой усредненные решающие профили, полученные с использованием множества обучающих данных, относящихся к классу :

. (4.8)

Мера сходства  между решающим шаблоном  для класса ,  и решающим профилем для тестового объекта  рассчитывается как

 (4.9)

и окончательное решение ансамбля принимается согласно максимуму значений меры сходства:

. (4.10)

Для предсказаний типа случай/контроль или прогноза течения заболевания, отклика на лечение мы рассматриваем задачу бинарной классификации. Положительному классу присваивается значение 1, отрицательному – значение -1. Используя следующую зависимость  мера сходства  для положительного и отрицательного класса записывается как

, (4.11)

и окончательное решение ансамбля следующее

 (4.12)

Заключение

В течении трех этапов реализации проекта № Ф19УКРГ-005 «Разработка методов, алгоритмов и интеллектуальной аналитической системы для обработки и анализа разнородных клинических и биомедицинских данных с целью совершенствования диагностики сложных заболеваний» авторами проанализированы различные подходы к предварительной обработке данных генной экспрессии, РНК-секвенирования, используемых в качестве источников биологической информации для выделения информативных маркеров сложных заболеваний. Разработаны ряд алгоритмов выделения информативных признаков и кластеризации биомедицинских данных и выполнена их программная реализация. Основные результаты выполнения проекта следующие:

* алгоритм предобработки клинических и биомедицинских данных, который позволяет ранжировать признаки по их информативности относительно классов (фенотипов заболевания или групп «случай-контроль»);
* алгоритмы выделения комбинаций биомаркеров на одном источнике данных, учитывающий коррелированность признаков и позволяющий исключить их влияние. В рамках алгоритма задача отбора информативных признаков формулируется как двухкритериальная оптимизационная задача, где первый критерий соответствует функции релевантности набора признаков по отношению к классу, второй критерий ‑ критерий избыточности определяет общую релевантность всех отобранных признаков. Предложенный эвристический метод позволяет учесть взаимное влияние отдельных признаков и определить близкое к оптимальному решение оптимизационной задачи, точное решении которой требует выполнения количества вычислений равное и является сложным в вычислительном плане;
* алгоритм выделения комбинаций биомаркеров на интегрированных источниках данных, позволяющего интегрировать данные интерактома и транскрипторные данные для определения функциональных подсетей, связанных с заболеванием. Результатом является минимальная подсеть или комбинация подсетей взаимодействующих генов, для которых нарушена регуляция в множестве случаев заболевания, однако в каждом отдельном случае могут быть дисрегулированы различные подмножества генов. Данный подход позволяет учесть индивидуальные различия генетических факторов при сложных заболеваниях и путем выделения функциональных модулей на сети объяснить появление схожих фенотипов несмотря на различные генетические причины, что является результатом нарушения функциональности одно и того же компонента клеточной системы;
* алгоритм кластеризации биомедицинских данных высокой размерности, на основе методов спектрального анализа, диффузионных карт и сетей. Алгоритм позволяет получать гладкое представление данных и устойчив к шуму и выбросам. Согласно алгоритму, исходное множество данных представляется в виде низко размерной структуры, что позволяет сократить размерность многомерных данных с одновременным выделением кластеров, характеризующий клинические показатели.
* программный модуль, объединяющий алгоритмы предобработки и выделения признаков реализован как набор функций в среде RStudio на языке программирования R. Программный модуль включает функции для ранжирования признаков, выделения подмножеств информативных признаков, а также ряд функций для выполнения классификации и оценки отобранных комбинаций признаков с использованием моделей классификации;
* программная реализация алгоритма представляет собой консольное приложение IFM, написанное на языке Java в программной среде разработки IntelliJ IDEA 2019.2. Данное приложение позволяет извлекать информативные функциональные модули или подмножества генов из биологической подсети и набора данных генной экспрессии, характеризующего подтипы заболеваний или данные типа «случай-контроль».
* программный модуль, реализующий алгоритм спектральной кластеризации представляет собой набор функций в среде RStudio на языке программирования R. Программный модуль включает функции для выполнения фильтрации генов с низкой вариабельностью относительно случаев, расчета значений различных вариантов ядерных функций, определяющих близость объектов данных, расчет собственных значений матрицы Лапласиана, выполнение кластеризации;
* гибридная модель классификации на основе ансамбля классификаторов и различных схем интеграции разнородных данных, которая позволяет учитывать особенности биомедицинских данных, а также большой объем и несбалансированность обучающей выборки.

Список использованных источников

1. Cook, N. R. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve / N.R. Cook // Clinical chemistry. – 2008. – Vol. 54, No. 1. – P. 17-23.
2. Hand, D.J. A simple generalisation of the area under the ROC curve for multiple class classification problems / D.J.Hand,.R.J.Till //Machine learning. – 2001. – Vol. 45, No. 2. – P. 171-186.
3. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. -2019.
4. Wang, Y. Gene selection from microarray data for cancer classification—a machine learning approach / Y. Wang [et al.] //Computational biology and chemistry. – 2005. – Vol. 29, No. 1. – P. 37-46.
5. Weinstein, J.N. The cancer genome atlas pan-cancer analysis project / J.N.Weinstein [et al.] //Nature genetics. – 2013. – Vol. 45, No. 10. – P. 1113.
6. Golub, T.R. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring / T.R.Golub [et al.] //Science. – 1999. – Vol. 286, No. 5439. – P. 531-537.
7. Alon, U. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays / U. Alon [et al.] //Proceedings of the National Academy of Sciences. – 1999. – Vol. 96, No. 12. – P. 6745-6750.
8. Cho, D.Y. Network biology approach to complex diseases / D.Y.Cho, Y.A.Kim, T.M.Przytycka // PLoS computational biology. – 2012. – Vol. 8, No. 12. – P. e1002820.
9. Yan, J. Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data / J.Yan [et al.] // Briefings in bioinformatics. – 2017. – Vol. 19, No. 6. – P. 1370-1381.
10. Johnson, D.S. Approximation algorithms for combinatorial problems / D.S. Johnson // Journal of computer and system sciences. – 1974. – Vol. 9, No. 3. – P. 256-278.
11. Prasad, T.S.K. Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology / T.S.K. Prasad, K.Kandasamy, A.Pandey // Reverse Chemical Genetics. – Humana Press, Totowa, NJ, 2009. – P. 67-79.
12. Ostell, J. The Entrez search and retrieval system / J. Ostell //The NCBI Handbook [Internet]. 2nd edition. – National Center for Biotechnology Information (US), 2014.
13. Edgar, R. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository / R.Edgar, M.Domrachev, A.E.Lash // Nucleic acids research. – 2002. – Vol. 30, No. 1. – P. 207-210.
14. Ideker, T. Discovering regulatory and signalling circuits in molecular interaction networks / T.Ideker [et al.] // Bioinformatics. – 2002. – Vol. 18, No. suppl\_1. – P. S233-S240.
15. Shannon, P. Cytoscape: a software environment for integrated models of biomolecular interaction networks / P. Shannon [et al.] // Genome research. – 2003. – Vol. 13, No. 11. – P. 2498-2504.
16. Kanehisa, M. KEGG: kyoto encyclopedia of genes and genomes / M.Kanehisa, S.Goto // Nucleic acids research. – 2000. – Vol. 28, No. 1. – P. 27-30.
17. Chung, F.R.K. Spectral graph theory / F.R.K.Chung, F.C.Graham. – American Mathematical Soc., 1997. – 215 p.
18. Mathews, J.C. Molecular phenotyping using networks, diffusion, and topology: soft tissue sarcoma / J.C.Mathews [et al.] // Scientific reports. – 2019. – Vol. 9, No. 1. – P. 1-9.
19. Xu, R. Applications of diffusion maps in gene expression data-based cancer diagnosis analysis / R.Xu, S.Damelin, D.C.Wunsch //2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. – IEEE, 2007. – P. 4613-4616.
20. Mishne, G. Data-driven tree transforms and metrics / G. Mishne [et al.] // IEEE transactions on signal and information processing over networks. – 2017. – Vol. 4, No. 3. – P. 451-466.
21. Belkin, M. Laplacian eigenmaps and spectral techniques for embedding and clustering / M.Belkin, P.Niyogi // Advances in neural information processing systems. – 2002. – P. 585-591.
22. Lafon, S. Data fusion and multicue data matching by diffusion maps / S.Lafon, Y.Keller, R.R.Coifman // IEEE Transactions on pattern analysis and machine intelligence. – 2006. – Vol. 28, No. 11. – P. 1784-1797.
23. Lafon, S. Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization / S.Lafon, A.B.Lee // IEEE transactions on pattern analysis and machine intelligence. – 2006. – Vol. 28, No. 9. – P. 1393-1403.
24. Farbman, Z. Diffusion maps for edge-aware image editing / Z.Farbman, R.Fattal, D.Lischinski // ACM Transactions on Graphics (TOG). – 2010. – Vol. 29, No. 6. – P. 145.
25. Talmon, R. Single-channel transient interference suppression with diffusion maps / R.Talmon, I.Cohen, S. Gannot // IEEE transactions on audio, speech, and language processing. – 2012. – Vol. 21, No. 1. – P. 132-144.
26. Bérard, P. Embedding Riemannian manifolds by their heat kernel / P.Bérard, G.Besson, S.Gallot // Geometric & Functional Analysis GAFA. – 1994. – Vol. 4, No. 4. – P. 373-398.
27. Yang, Y.A. Unified Framework for Discrete Spectral Clustering / Y.A.Yang [et al.] //Proceedings of the International Joint Conferences on Artificial Intelligence . – 2016. – P. 2273-2279.
28. Николенко, С.И. Глубокое обучение. Погружение в мир нейронных сетей / С.И. Николенко, А.А. Кадурин,Е.В. Архангельская. – СпБ.: Питер, 2018. – 480 с.
29. Hinton, G.E. Reducing the dimensionality of data with neural networks / G.E.Hinton, R.R.Salakhutdinov // Science. – 2006. – Vol. 313, No. 5786. – P. 504-507.
30. Vincent, P. Extracting and composing robust features with denoising autoencoders / P. Vincent [et al.] // Proceedings of the 25th international conference on Machine learning. – ACM, 2008. – P. 1096-1103.
31. Mishne, G. Diffusion nets / G. Mishne [et al.] // [Applied and Computational Harmonic Analysis](https://www.sciencedirect.com/science/journal/10635203). – 2019.‑ Vol. 47, No. 2. ‑P.259-285.
32. Benjamini, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing / Y.Benjamini, Y.Hochberg // Journal of the Royal statistical society: series B (Methodological). – 1995. – Vol. 57, No. 1. – P. 289-300.
33. Ng, A.Y. On spectral clustering: Analysis and an algorithm / A.Y.Ng, M.I.Jordan, Y.Weiss // Advances in neural information processing systems. – 2002. – P. 849-856.
34. Xiang, T. Spectral clustering with eigenvector selection / T.Xiang, S.Gong // Pattern Recognition. – 2008. – Vol. 41, No.3. – P. 1012-1029.
35. Van der Maaten, L.J.P. Visualizing High-Dimensional Data Using t-SNE / L.J.P. van der Maaten, G.E.Hinton // Journal of Machine Learning Research – 2008. – P. 2579-2605.
36. The Cancer Genome Atlas [Electronic resource]. – Mode of access: http://cancergenome.nih.gov/abouttcga. – Date of access: 9.03.2019.
37. Multiple Classifier Systems / J. Kittler & F. Roli (editors) // Proc. of 2nd International Workshop, MCS2001, (Cambridge, UK, 2-4 July 2001) / Lecture Notes in Computer Science. – Vol. 2096. – Springer-Verlag, Berlin.
38. Valentini, G. Ensembles of learning machines / G. Valentini, F. Masulli // in: R. Tagliaferri, M. Marinaro (eds.), Neural Nets WIRN Vietri-2002, Lecture Notes in Computer Science, Springer, Berlin. ‑ June 2002. – Vol. 2486. ‑ P. 3–19.
39. Kuncheva, L.I. Combining Pattern Classifiers // Methods and Algorithms, Wiley, 2004.
40. Kuncheva, L. Decision templates for multiple classifier fusion: an experimental comparison / L. Kuncheva, J. Bezdek, R. Duin // Pattern Recognition. – 2001. – Vol 34. – P. 299–314.