РАЗРАБОТКА ИМИТАЦИОННОЙ МОДЕЛИ И ИССЛЕДОВАНИЕ АЛГОРИТМОВ АВТОМАТИЧЕСКОГО ВЫБОРА АТРИБУТОВ ЭКЗОНОВ ГЕНОВ

Волков А. В.[1], Яцков Н. Н.[1], Гринев В. В.[2]
[1]Кафедра системного анализа и компьютерного моделирования, [2]кафедра генетики, Белорусский государственный университет
Минск, Республика Беларусь
E-mail: andrei@cybergizer.com, yatskou@bsu.by, grinev_vv@bsu.by

В работе предложен алгоритм для генерации многомерных наборов данных, учитывающий информативность подгрупп признаков. Проведено исследование алгоритмов отбора признаков на смоделированных данных. Представлены результаты исследования точности классификации экзонов генов в зависимости от числа отобранных признаков.

Введение

Онкологические болезни определяются наличием экспрессированых онкогенов. Гены состоят из экзонов и энтронов. Особый интерес представляют экзоны. Из экзонов формируются транскрипты РНК. На основе транскриптов РНК происходит синтез белка в клетке. В качестве признаков экзонов могут выступать как свойства нуклеотидных последовательностей, так и свойства экзонов измеренные экспериментальным путем. Каждый экзон характеризуется большим количеством признаков. Для увеличения точности анализа экзонов (например, для решения задач классификации и кластеризации) следует оставить только информативные признаки. Выбор наиболее значимых признаков экзонов гена, в том числе и онкогена, является малоисследованной задачей.

Вычислительные возможности современной техники позволили разработать большую базу алгоритмов отбора признаков объектов данных [1]. Однако выбор оптимального алгоритма анализа не является тривиальной задачей и часто делается интуитивно, что, в свою очередь, приводит к ошибкам интерпретации результатов анализа. Задачу выбора оптимального алгоритма можно упростить, используя имитационное моделирование и синтетические данные. Существующие имитационные модели генерации кластеров многомерных данных имеют ряд ограничений [2], среди которых наиболее существенным является невозможность явной задачи степени информативности признака или подгрупп признаков.

Цель работы – разработка имитационной модели учитывающей информативности признаков объектов данных и исследование алгоритмов автоматического выбора атрибутов экзонов генов человека.

I. Методология

Входные параметры имитационной модели: число кластеров в генерируемом наборе данных; число и размеры подгрупп признаков с заданной степенью разделимости между кластером и ближайшими к нему кластерами; число шумовых признаков; число выбросов в данных; размеры кластеров; параметр формы кластеров (эллиптическая и сферическая формы). Блок схема алгоритма имитационного моделирования представлена на рис. 1.

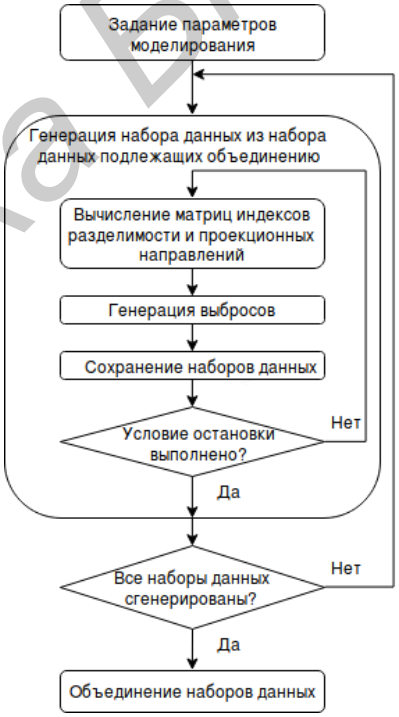

Рис. 1 – Блок схема алгоритма имитационного моделирования

Среди алгоритмов отбора признаков широкое рапространение получили методы-фильтры [1], что обусловлено легкостью их проектирования и простой структурой. В настоящей работе были выбраны популярные и универсальные методы-фильтры: алгоритм счёта Фишера [1], алгоритм Relief-F [2] и алгоритм минимальная избыточность-максимальная релевантность [3].

Оценка эффективности алгоритмов отбора признаков выполена по методу к-ближайших соседей [4].

Экспериментальные данные взяты из базы данных Ensembl [5] и содержат 1762 уникальных экзона. Каждый экзон характеризуется 178 численными признаками. Для каждого из экзонов указана принадлежность к модельному гену человека. Совокупное число генов — 14. Разработанные алгоритмы реализованы на языках программирования R и Python.

II. Результаты

Исследована эффективность алгоритмов отбора признаков на смоделированных данных в условиях большого числа шумовых признаков, для различного числа кластеров данных и различной разделимости между кластерами. Все исследуемые алгоритмы отбора признаков продемонстрировали способность выделять значимые признаки на фоне шумовых признаков. Некоторые результаты представлены на рис. 2 и рис. 3.

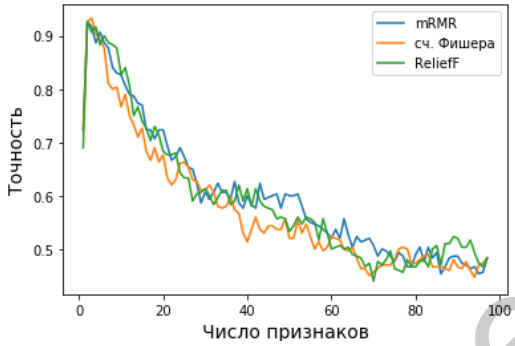

Рис. 2 – Зависимость точности классификации объектов от количества признаков для 3 классов в условиях плохой разделимости

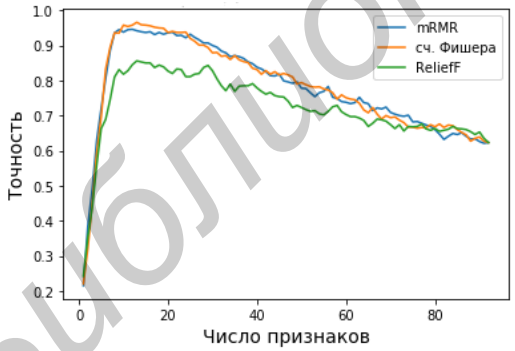

Рис. 3 – Зависимость точности классификации объектов от количества признаков для 10 классов в условиях хорошей разделимости

Исследована эффективность алгоритмов отбора признаков на примерах классификации экзонов генов человека. Установлен факт значимой разделимости между экзонами принадлежащими различным генам. Наилучшая точность классификации достигается при классификации наборов, состоящих из двух генов и принимает значение 0.95. Некоторые результаты представлены на рис. 4 и рис. 5.

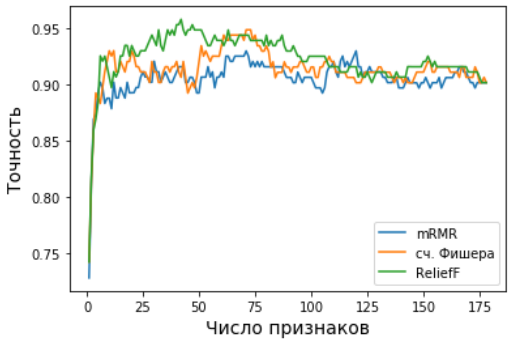

Рис. 4 – Зависимость точности классификации экзонов от количества признаков для 2 генов

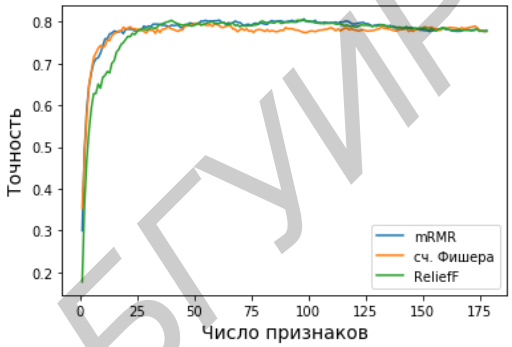

Рис. 5 – Зависимость точности классификации экзонов от количества признаков для 14 генов

III. Выводы

В работе реализован алгоритм имитационного моделирования многомерных наборов данных с учетом информативности признаков объектов. Алгоритм позволяет осуществлять оптимальный выбор наиболее эффективных алгоритмов отбора признаков для решения задач классификации различной сложности.

Разработанные алгоритмы позволяют классифицировать экзоны 14 генов на небольшом наборе наиболее информативных признаков с точностью 0.78.

IV. Список литературы

1. Feature Selection: A Data Perspective [Electronic resource] / J. Li, K. Cheng, S. Wang, F. Morstatter, R. Trevino, J. Tang, H. Liu, 2016. – Mode of access: https://arxiv.org/abs/1601.07996. – Date of access: 3.09.2017.
2. Qiu, W.L. Generation of Random Clusters with Specified Degree of Separaion / W.L. Qiu, H. Joe // Journal of Classification. – Vol. 23(2). – 2005. – P. 315–334.
3. Peng H. C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. / H. C. Peng, F. Long, C. Ding // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2005. – Vol. 27, № 8. – P. 1226–1238.
4. Novaković, J. Toward optimal feature selection using ranking methods and classification algorithms / J. Novaković // Yugoslav Journal of Operations Research. – 2016. – Vol. 21, № 1. – P. 132.
5. Aken, B. L. The Ensembl gene annotation system. B.L. Aken, S. Ayling, D. Barrell, L. Clarke, V. Curwen, S. Fairley, J. Fernandez Banet, K. Billis, C. Garcia Giron, T. Hourlier, et al. (2016) Database (Oxford), doi: 10.1093/database/baw093