АНАЛИЗ АЛГОРИТМОВ ВЫБОРА ПРИЗНАКОВ ЭКЗОНОВ

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

Рассмотрена задача сокращения размерности пространства признаков экзонов человека с целью увеличения точности определения их генной принадлежности. Выполнен сравнительный анализ алгоритмов отбора признаков при варьировании алгоритмов индуктивного обучения.

Введение

Исследование организации и функционирования генов человека является важной задачей биоинформатики [1]. Гены состоят из экзонов и интронов. Экзон характеризуется большим количеством признаков (более 1000), в то же время число экзонов, принадлежащих гену, невелико (как правило, менее 200). Проблема большого числа признаков и относительно малого числа объектов наблюдений характерна для всей области биоинформатики в целом в связи с трудностями для алгоритмов индуктивного обучения.

Использование алгоритмов автоматического отбора признаков позволяет снизить размерность решаемой задачи [2]. Эффективность упомянутых алгоритмов определяется свойствами исследуемых наборов данных и числом объектов наблюдения [3], что формирует необходимость исследования алгоритмов отбора признаков в контексте экзонных данных. Так в настоящей работе исследованы алгоритмы отбора признаков семейства Relief [5].

Целью работы является выяснение принципиальной возможности предсказания генной принадлежности экзонов, а также исследование эффективности алгоритмов семейства Relief в задачах отбора признаков экзонов.

I. Методология

Экспериментальные данные получены из базы данных Ensembl [4] и содержат 1762 уникальных экзона, принадлежащих 14 произвольно отобранным генам. Каждый экзон дополнительно охарактеризован с помощью 1198 численных признаков: 429 признаков непосредственно самих экзонов и 769 признаков фланкирующих участков нуклеотидных последовательностей (длина цепи составляет 100 нуклеотидов).

Среди алгоритмов отбора признаков широкое распространение получили методы-фильтры [2], что обусловлено простой структурой, вычислительной эффективностью и независимостью от типа используемого алгоритма индуктивного обучения. Абсолютное большинство методов-фильтров являются унивариативными [2]. Исключение составляет семейство алгоритмов на основе метода Relief [5], способные учитывать зависимости между признаками. Алгоритмы SURF [5], MultiSURF [5], SURFStar [5] являются наиболее популярными Relief-алгоритмами в приложениях биоинформатики и включены в состав фрэймворка ReBATE (англ. Relief-Based Algorithm Training Environment) [5].

В работе рассмотрены три алгоритма индуктивного обучения: метод k-ближайших соседей [6], машина опорных векторов с линейным ядром [7] и наивный байесовский классификатор [8]. Выбор алгоритмов обусловлен их популярностью, разнородностью подходов к индуктивному обучению, отсутствием встроенных механизмов отбора признаков.

В работе использовано скоринговое правило Брайера [9] для оценки качества вероятностного прогноза алгоритма индуктивного обучения.

Блок-схема организации вычислительного эксперимента представлена на рис. 1.

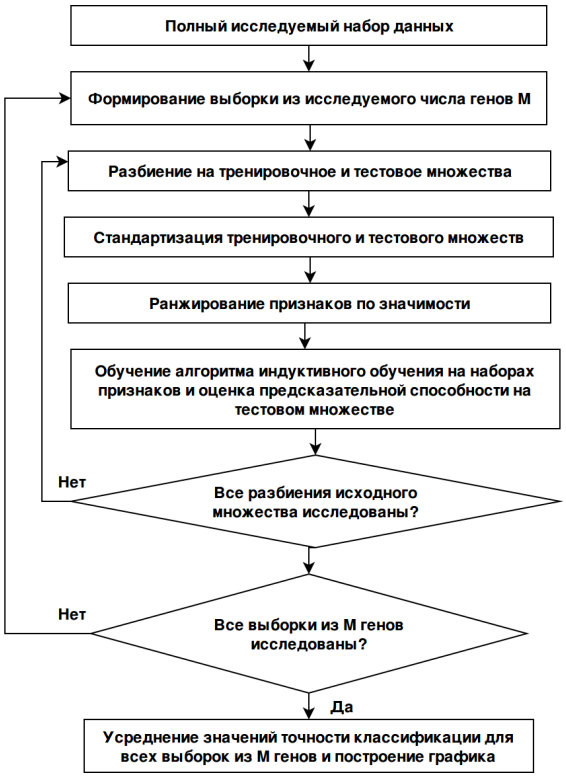

Рис. 1 – Блок-схема организации вычислительного эксперимента

Представленный подход позволяет исследовать зависимость оценки предсказательной способности алгоритмов индуктивного обучения от числа ранжированных по информативности признаков.

II. Результаты и обсуждение

Исследована эффективность алгоритмов отбора признаков в задачах предсказания генной принадлежности экзонов человека при варьировании алгоритмов индуктивного обучения. Установлен факт значимой разделимости между экзонами, принадлежащими различным генам. Пример зависимости представлен на рис. 2.

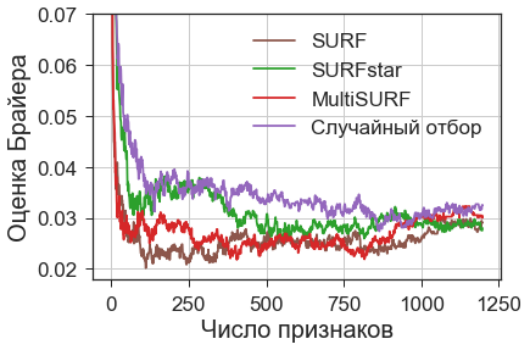

Рис. 2 – Зависимость оценки по скоринговому правилу Брайера от числа ранжированных по информативности признаков (2 гена, метод 1 ближайшего соседа)

Тренировка алгоритмов индуктивного обучения на признаках фланкирующих интронов обеспечивает более высокую предсказательную способность классификаторов по сравнению с тренировкой алгоритмов индуктивного обучения на признаках экзонных нуклеотидных последовательностей (рис. 3). Это наблюдение само по себе представляет большой интерес и требует дальнейшего детального изучения с помощью методов биоинформатики, а также экспериментальных методов молекулярной биологии.

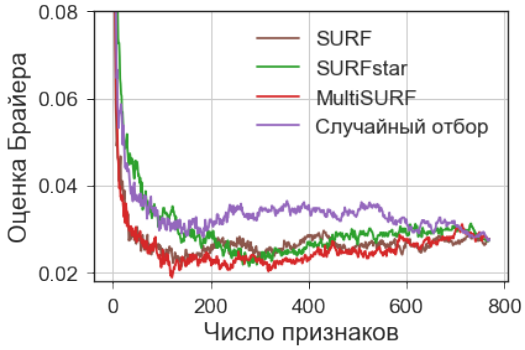

Рис. 3 – Зависимости оценок по скоринговому правилу Брайера для ранжированного ряда признаков фланкирующих интронов

III. Заключение

Установлен факт значимой разделимости между экзонами, принадлежащими разным генам. Показано, что использование алгоритмов автоматического отбора в сочетании с методом k-ближайших соседей (1 ближайший сосед) позволяет уже на 15 признаках достигать 96% точности предсказания генной принадлежности, что на 3.6% выше, чем случайный отбор признаков и значительно вычислительно эффективнее анализа полного набора из 1198 признаков. Обнаружено, что более низкие значения счетов Брайера (0,02) соответствуют обучению алгоритмов на признаках фланкирующих интронов, в сравнении с величиной 0,07 для признаков экзонных нуклеотидных последовательностей (метод одного ближайшего соседа в сочетании с алгоритмом MultiSURF).

IV. Список литературы

1. Grinev V. V., Migas A. A., Kirsanava A. D., Mishkova O. A., Siomava N., Ramanouskaya N. V., Vaitsiankova A. V., Ilyushonak I. M., Nazarov P. V., Vallar L., Aleinikova O. V. Decoding of exon splicing patterns in the human RUNX1-RUNX1T1 fu- sion gene // Int. J. Biochem. Cell Biol. 2015. Vol. 68. P. 48–58

2. Li JD, Cheng KW, Wang SH, Morstatter F, Trevino RP, Tang JL, Liu H (2016) Feature selection: a data perspective, vol 3, pp 1–73. arXiv:1601.07996

3. Oreski D, Oreski S, Klicek B. Effects of dataset characteristics on the performance of feature selection techniques. Appl Soft Comput. 2017;52:109119.

4. 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

5. Urbanowicz, R. J., Olson, R. S., Schmitt, P., Meeker, M., Moore, J. H., 2018. Benchmarking relief-based feature selection methods for bioinformatics data mining. Journal of biomedical informatics.

6. Altman, N. S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression". The American Statistician. 46 (3): 175–185.

7. Cortes, Corinna; Vapnik, Vladimir N. (1995). "Support-vector networks". Machine Learning. 20 (3): 273–297.

8. Russell, Stuart; Norvig, Peter (2003) [1995]. Artificial Intelligence: A Modern Approach (2nd ed.). Prentice Hall. ISBN 978-0137903955.

9. Gneiting, T. and Raftery, A. E. (2007) Strictly proper scoring rules, prediction, and estimation. J. Am. Statist. Ass., 102, 359–378.