УДК 621.311

КЛАССИФИКАЦИЯ ЭКЗОНОВ ГЕНОВ ЧЕЛОВЕКА С ИСПОЛЬЗОВАНИЕМ АЛГОРИТМОВ АВТОМАТИЧЕСКОГО ВЫБОРА АТРИБУТОВ ЭКЗОНОВ

аспирант, А.В. ВОЛКОВ,
канд. физ.-мат. наук, доц. Н.Н. ЯЦКОВ,
канд. биол. наук, доц. В.В. ГРИНЕВ
(Белорусский государственный университет, Беларусь)

Исследование онкогенов человека является важной задачей биоинформатики [1]. Онкологические болезни определяются наличием экспрессированных онкогенов [2]. Гены состоят из экзонов и интронов. Особый интерес представляют экзоны. Из экзонов формируются транскрипты РНК. На основе транскриптов РНК происходит синтез белка в клетке [3]. В качестве признаков экзонов могут выступать длины нуклеотидных последовательностей, биофизические свойства экзонов, измеренные экспериментальным путем, признаки фланкирующих нуклеотидных участков последовательностей [4]. Каждый экзон характеризуется большим количеством признаков, в то же время число наблюдений невелико. Проблема большого числа признаков и относительно малого числа наблюдений характерна для всей доменной области биоинформатики в целом [5].

Обучение алгоритмов классификации является сложной задачей, что связано с «проклятием размерности» [6]. Возможным решением данной проблемы является использование алгоритмов отбора признаков, которые осуществляют непосредственный отбор наиболее релевантных признаков из исходного множества признаков. Отсутствие каких-либо преобразований над исходными признаками позволяет сохранить физический смысл признаков. Данное свойство является особенно важным в биоинформатических приложениях поскольку каждый из признаков имеет уникальный биологический смысл, важный для эксперта в доменной области. В настоящее время применение алгоритмов отбора признаков особенно необходимо в приложениях биоинформатики, связанными с построениями прогностических моделей машинного обучения.

Целью работы является исследование влияния алгоритмов автоматического выбора атрибутов на точность классификации экзонов генов человека.

В работе выполнен сравнительный анализ наиболее эффективных алгоритмов автоматического отбора признаков на примерах наборов экзонов 14 генов организма человека [7].

Методология. Блок-схема организации вычислительного эксперимента для исследования алгоритмов автоматического выбора признаков экзонов представлена на рисунке 1.

Рисунок 1 – Блок-схема организации вычислительного эксперимента

В блоке 1 осуществляется загрузка анализируемых данных. Экспериментальные данные получены из базы данных Ensemble [6] и содержат 1762 уникальных экзона. Каждый экзон характеризуется 1198 численными признаками (429 признаков непосредственно самих экзонов и 769 признаков фланкирующих нуклеотидных последовательностей). Для каждого из экзонов указана принадлежность к модельному гену человека. Совокупное число генов 14.

В блоке 2 производится разбиение данных на два подмножества, эталонную и тестируемую выборки, использующиеся далее для перекрестной проверки.

В блоке 3 выполняется ранжирование признаков по информативности. Среди алгоритмов отбора признаков широкое распространение получили методы-фильтры [8], что обусловлено легкостью их проектирования и простой структурой. В настоящей работе рассмотрены наиболее популярные и универсальные методы-фильтры: алгоритм счета Фишера [9], алгоритм Relief-F [10] и алгоритм отбора признаков на основе индекса Джини [11].

В блоке 4 оценивается релевантность выбранного набора признаков с помощью оценки точности классификации алгоритмов индуктивного обучения на тестовом наборе данных. Реализованы три типа классификаторов: наивный байесовский классификатор [12], метод k-ближайших соседей [13] и машина опорных векторов [14]. Алгоритмы представляют три совершенно разных подхода к индуктивному обучению и не содержат встроенных механизмов отбора признаков.

В блоке 5 оценивается точность классификации экзонов в зависимости от количества наиболее информативных признаков.

Результаты. Исследована эффективность алгоритмов отбора признаков на примерах классификации экзонов генов человека. На рисунке 2 представлена зависимость точности бинарной классификации экзонов от количества признаков для алгоритмов отбора признаков и метода k-ближайших соседей. Результаты получены усреднением по 10 парам различных генов. По характеру полученных зависимостей установлен факт значимой разделимости между экзонами принадлежащими различным генам.

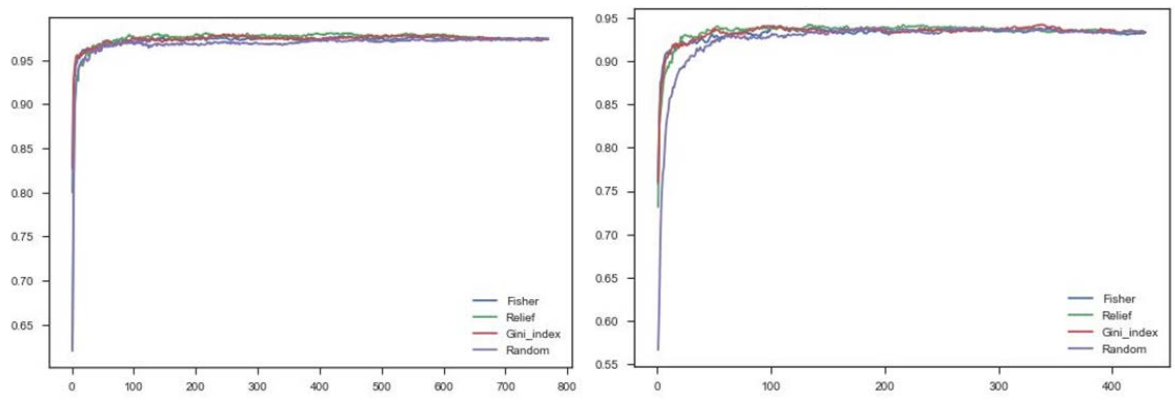

Рисунок 2. – Зависимость точности бинарной классификации экзонов по методу k-ближайших соседей в зависимости от числа признаков фланкирующих нуклеотидных последовательностей (слева) и числа признаков экзонов (справа)

Наилучшую точность классификации среди демонстрирует метод k ближайших соседей – 0.96. При этом использование алгоритмов отбора признаков не дает преимущества над случайным отбором признаков (кривые Random на рисунке 2.).

Выводы. Показана принципиальная применимость подхода выделения наиболее информативных признаков экзонов для решения задачи классификации экзонов генов и последующего предсказания альтернативных вариантов транскриптов РНК генов человека.

Установлен факт хорошей разделимости между экзонами принадлежащими различным генам (метод k-ближайших соседей, 2 класса – точность 0.96), что свидетельствует о принципиальной возможности классификации экзонов.

Проведено исследование алгоритмов отбора признаков на признаках экзонов генов. Алгоритм счета Фишера в контексте отбора признаков экзонов демонстрирует наивысшую вычислительную эффективность, при схожих показателях точности классификации.

Литература

1. Kashyap A., Naresh Babu M., Bujjamma D. (2015) Bioinformatics of Non Small Cell Lung Cancer and the Ras Proto-Oncogene. In: Bioinformatics of Non Small Cell Lung Cancer and the Ras Proto-Oncogene. SpringerBriefs in Applied Sciences and Technology. Springer, Singapore
2. Croce CM (January 2008). "Oncogenes and cancer". The New England Journal of Medicine. 358 (5): 502–11. doi:10.1056/NEJMra072367. PMID 18234754.
3. Lodish H, Berk A, Zipursky SL, et al. Molecular Cell Biology. 4th edition. New York: W. H. Freeman; 2000. Section 9.1, Molecular Definition of a Gene.
4. M. Q. Zhang; Statistical Features of Human Exons and Their Flanking Regions, Human Molecular Genetics, Volume 7, Issue 5, 1 May 1998, Pages 919–932
5. Van Hulse J, Khoshgoftaar TM, Napolitano A, Wald R (2012) Threshold-based feature selection techniques for high-dimensional bioinformatics data. Netw Model Anal Health Inf Bioinformatics 1(1–2):47–61
6. Bellman R.E. Dynamic Programming / R.E. Bellman // Courier Dover Publications, 2003 – 384p.
7. 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. GarciaGiron, T. Hourlier, et al. (2016) Database (Oxford), doi:10.1093/database/baw093
8. Sanchez-Marono N, Alonso-Betanzos A, Tombilla-Sanroman M. Filter Methods for Feature Selection – A Comparative Study.Intelligent Data Engineering and Automated Learning. Springer Berlin Heidelberg, pp.178-187(2007).
9. Richard O Duda, Peter E Hart, and David G Stork. Pattern classification. John Wiley & Sons, 2012.
10. Kenji Kira and Larry A Rendell. The feature selection problem: Traditional methods anda new algorithm. In AAAI, volume 2, pages 129–134, 1992.
11. Gini, Corrado (1912). Variabilità e mutabilità. Reprinted in Pizzetti, E.; Salvemini, T., eds. (1955). Memorie di metodologia statistica. Rome: Libreria Eredi Virgilio Veschi.
12. Russell, Stuart; Norvig, Peter (2003) [1995]. Artificial Intelligence: A Modern Approach (2nd ed.). Prentice Hall. ISBN 978-0137903955.
13. Altman, N. S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression". The American Statistician. 46 (3): 175–185.
14. Cortes, Corinna; Vapnik, Vladimir N. (1995). "Support-vector networks". Machine Learning. 20 (3): 273–297.