ПЕРЕДБАЧУВАЛЬНА АНАЛІТИКА В МЕДИЦИНІ

Є.Б. ФІСУН

Передбачувальна аналітика має величезний потенціал для розвитку в галузі медицини. Аналізуючи складні закономірності в медичних даних(генетичних профілях), прогностичні моделі можуть ідентифікувати людей із високим ризиком розвитку захворювань ще до появи симптомів, визначити які гени більш за все провокують хворобу.

Незважаючи на те що передбачувальна аналітика дає багато можливостей для розвитку індустрії охорони здоров'я, вона стикається з низкою проблем пов'язаних з даними, а саме складність та варіативність, оскільки лікарі збирають повний анамнез, що включає багато різних атрибутів, які можуть впливати на стан хворого. Медичні дані дуже часто можуть містити помилки, бути неповними, окрім цього відсутні дані для конкретних хвороб, груп (stratified sampling problem), що у свою чергу впливає на точність передбачувальної моделі.

Одними з способів як можна вирішити проблему маленьких вибірок є генерація нових семплів, за допомогою bootstrapping або jackknife. Оскільки медичні дані мають багато предикторів, то застосовуються feature selection алгоритми (filter or wrapper methods) завдяки яким можна виділити найважливіші показники, що пояснюють залежну змінну. В охороні здоров'я частою задачею є аналіз виживання, тобто нас цікавить час до визначеної події (рецидив, смерть), такі дослідження характеризуються наявністю censored data що робить некоректним використання традиційних методів (наприклад логістична регресія) для передбачення скільки часу ще проживе індивідум, для даної задачі варто застосовувати Kaplan-Meier analysis, Cox regression.

У роботі було проаналізовано комбінований датасет експресії генів людей та мишей, двох груп (хворі на ангельман синдром та контрольна група). Загалом вибірка складається з 60 семплів та 2399 генів. Для візуалізації експресії генів було використано heatmap. За допомогою heatmap ми можемо, які гени мають більш високу експресію (червоний колір), а які більш низьку (синій). Було проведено ряд експериментів, для покращення точності класифікатора, а саме застовування фільтраційних методів коефіцієнт взаємної інформації

$$\sum_{x \in X} \sum_{y \in Y} p(x,y) log \frac{p(x,y)}{p(x)p(y)}$$

та тест Колмогорова-Смірнова щоб визначити чи є значні відмінності між генами хворих та здорових кандидатів [1].

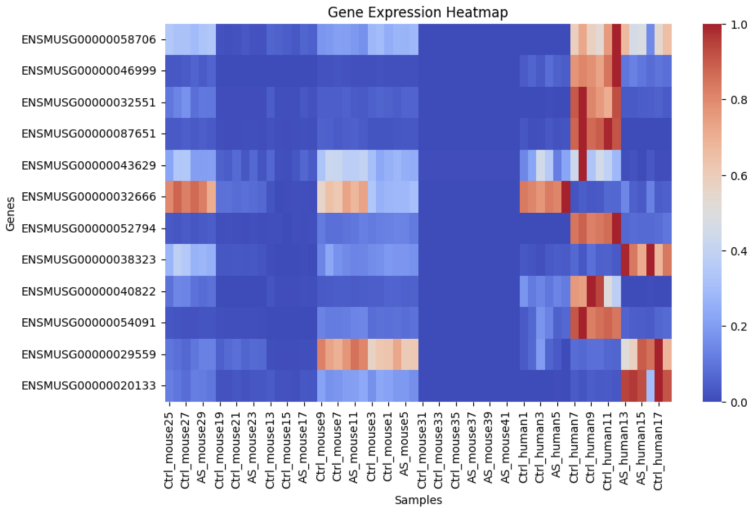

Рис. 1. Heatmap of gene expression data

Також був застосований wrapper method (рекурсивне усунення ознак) для виявлення найбільш вагомих ознак. Окремо від алгоритмів для визначення значних ознак було зменшено розмірність датасету за допомогою PCA [2]. Потім було вирішено підійти до цієї проблеми зі сторони даних та збільшити кількість семплів за допомогою bootstrapping, а потім зробити semi-supervised learning для нерозмічених даних та натренувати фінальний класифікатор.

Підсумовуючи у роботі було показано, що класичні методи машинного навчання не працюють для передбачувальної аналітики, це характеризується специфікою даних, оскільки ми маємо малу вибірку та багато незалежних змінних ми стикаємось з проблемою перенавчання.

Література

[1] Aliaa El-Gawady, Mohamed A. Makhlouf та інші. *Machine Learning Framework for the Prediction of Alzheimer's Disease Using Gene Expression Data Based on Efficient Gene Selection* // MDPI — 2022.

[2] Olutomilayo Olayemi Petinrin, Faisal Saeed, Naomie Salim та інші. *Dimension Reduction and Classifier-Based Feature Selection for Oversampled Gene Expression Data and Cancer Classification* // MDPI — 2023.

Національний університет Києво-Могилянська Академія, Київ, Україна
Email address: `yelyzaveta.fisun@ukma.edu.ua`