Передбачувальна аналітика в медицині

Фісун Єлизавета

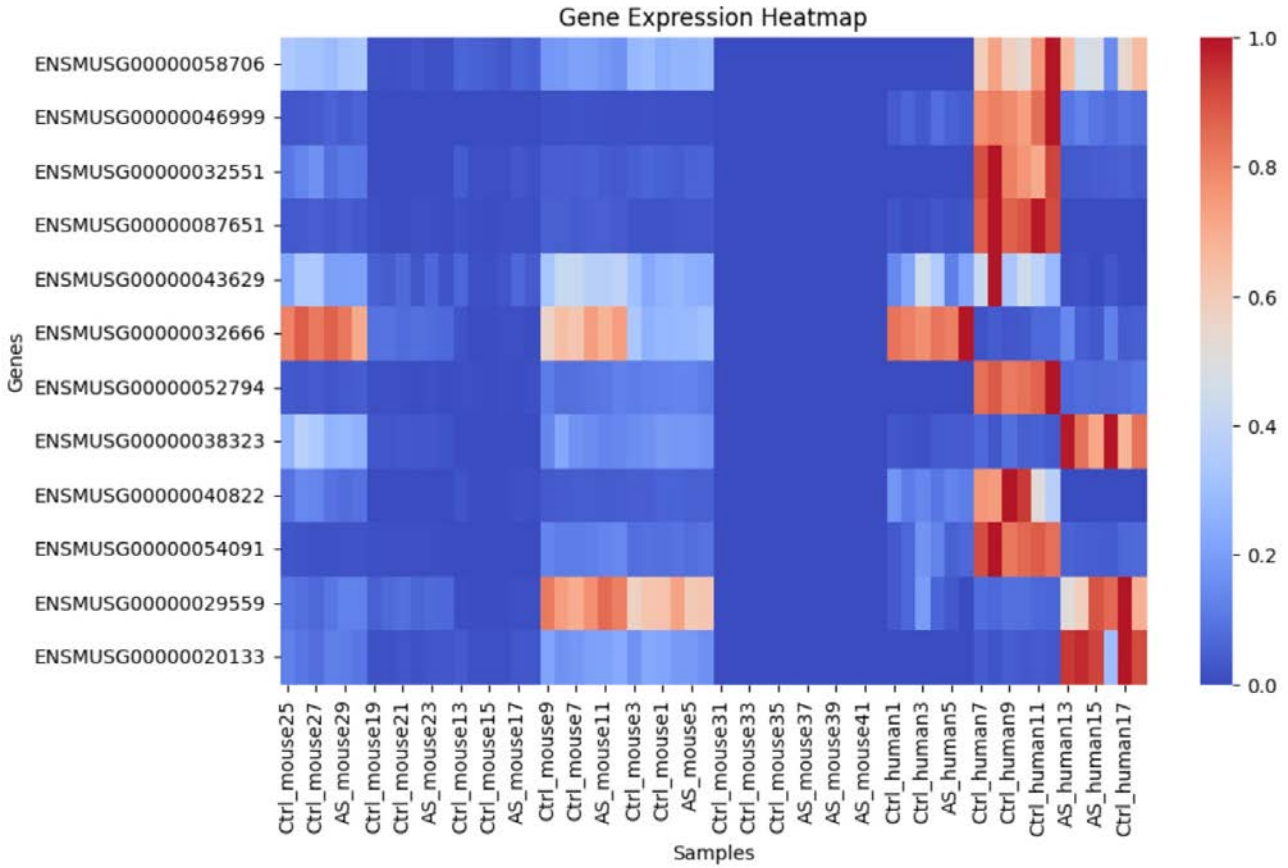
Актуальність

- 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.

Мета

- Показати особливості моделювання в медицині, класичні підходи машинного навчання не завжди працюють через специфіку даних
- У практичній задачі зробити класифікатор, який в якості вхідних даних приймає експресію певного списку генів і передбачає чи даний індивідум хворий на Ангельман синдром чи ні

Візуалізація даних



Проблеми

- Дані високої розмірності
- Мала вибірка даних для окремих хвороб
- Помилки вимірювань в даних
- Наявні викиди в даних, шум
- Різні шкали вимірювань у різних експериментах
- У підсумку ми стикаємось з проблемою перенавчання моделі

Обробка даних

- Датасет: 9 різних експериментів, 60 семплів (42 - миші, 18 - люди) та 2399 спільних генів (або 3236 генів у випадку тільки мишей)
- Приведення генів до одного неймспейсу та/або конвертація генів між різними організмами
- Визначити спільні гени між експериментами
- Нормалізація даних
- Детекція аутлаєрів

Класифікатори

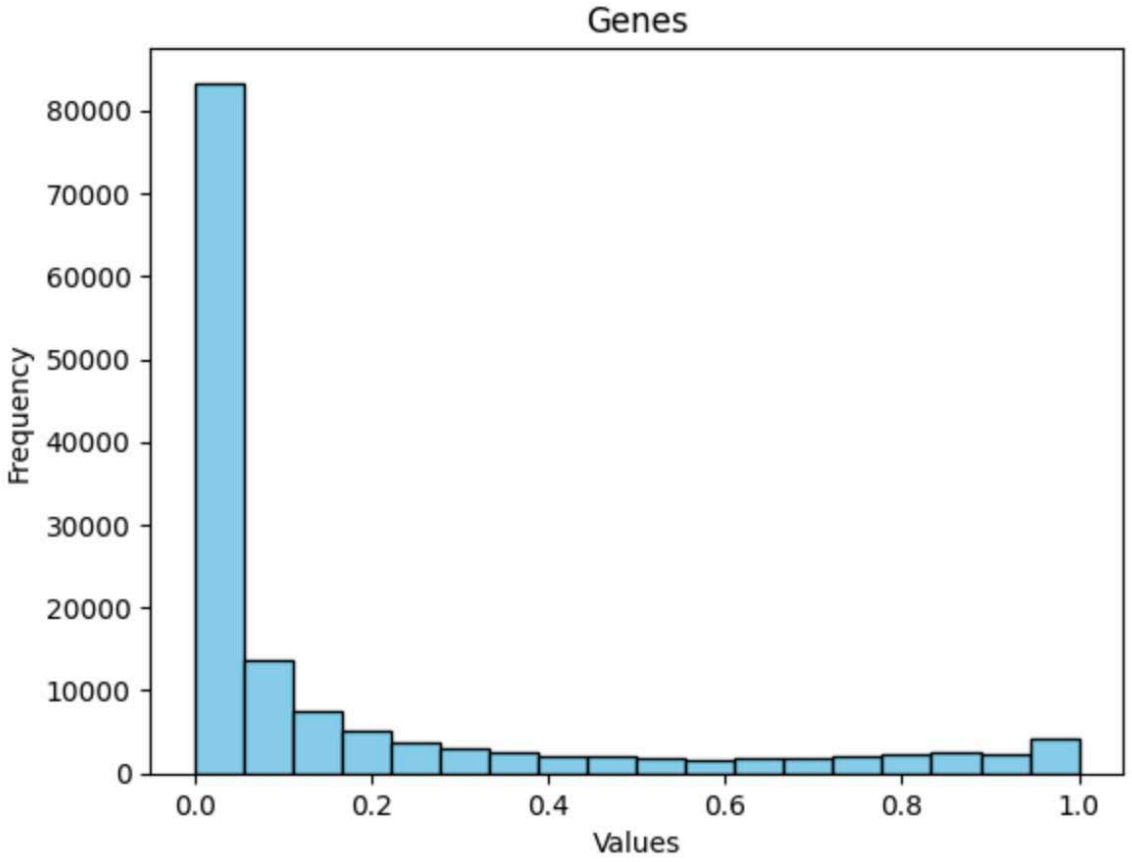
- Логістична регресія
- Random Forest
- SVM

Model	Dataset	Accuracy
Logistic regression	human/mouse/combined	0.25/0.44/0.33
SVM (sigmoid)	human/mouse/combined	0.5/0.33/0.42
Random forest	human/mouse/combined	0.5/0.44/0.25

Визначення важливих генів (комбінований датасет)

- Взаємна інформація (Mutual information)

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



Визначення важливих генів (тільки миші)

- Тест Колмогорова-Смірнова

$$F(x)=\begin{cases} 0, & x < x_{(1)} \\ \frac{k}{n}, & x_{(k)} \le x \le x_{(k+1)}, \quad k = 1, \dots, n-1 \\ 1, & x > X_{(n)}. \end{cases}$$

$$D = \max_x |F_1(x) - F_2(x)|.$$

D critical = 0.20985, α=0.05, якщо D < D critical, то немає значної різниці між негативними і позитивними семплами. З 3236 генів вийшло всього 136 важливих. Cross-validation, Random Forest, точність 0.7088.

Резултати

Model	FS method	Dataset	Accuracy
Logistic regression	Mutual information (top 10)	Combined	Cross-validation 0.28
Random forest	Mutual information (top 10)	Combined	Cross-validation 0.44
SVM	Mutual information (top 10)	Combined	Cross-validation 0.38
Random forest	K-S test	Mouse	Cross-validation 0.71

Висновки

- Через малу вибірку та велику кількість фічей ми стикаємось з проблемою overfitting. З однієї сторони спробувати вирішити цю проблему можна за допомогою фільтрації генів та відбору більш вагомих. Також у перспективі можна спробувати збільшити розмір вибірки за допомогою генерації синтетичних даних та проаналізувати чи вплине це на точність моделі.

Джерела

1. Machine Learning Framework for the Prediction of Alzheimer’s Disease Using Gene Expression Data Based on Efficient Gene Selection. Aliaa El-Gawady, Mohamed A. Makhlouf and others. 2022
2. A Cancer Gene Selection Algorithm Based on the K-S Test and CFS. Qiang Su, Yina Wang and others 2017
3. Dimension Reduction and Classifier-Based Feature Selection for Oversampled Gene Expression Data and Cancer Classification. Olutomilayo Olayemi Petinrin, Faisal Saeed and others. 2023