Классификация последовательности DNA Seq

Работу выполнили
студенты группы ПМИ-1,2-2021:
Андрюков Анатолий,
Ковбаснюк Владимир,
Лис-Граундер Алиса

Постановка задачи

X - множество последовательностей ДНК человека;
Y - множество классов объектов, объединенных общей информацией о последовательностях нуклеотидов в молекулах ДНК человека;
$\hat{y} : X \to Y$ - неизвестная зависимость.

Дано
$x_1, \ldots, x_n \subset X$ - обучающая выборка;
$y_i = \hat{y}(x_i), i = 1, \ldots, n$ - известные метки классов.

Найти
модель с алгоритмом $a : X \to Y$, который наилучшим образом будет классифицировать произвольный объект $x \in X$.

Описание набора данных

Набор данных состоит из последовательностей ДНК человека и меток классов для них.

Последовательности ДНК представлены в формате FASTA - текстовом формате, в котором нуклеотиды представлены в виде однобуквенных символов [A,C,G,T,N].

sequence	class
ATGCCCCAACTAAATACTACCGTATGGCCCACCATAATTACCCCCA...	4
ATGAACGAAAATCTGTTCGCTTCATTCATTGCCCCCACAATCCTAG...	4
ATGTGTGGCATTTGGGCGCTGTTTGGCAGTGATGATTGCCTTTCTG...	3
ATGTGTGGCATTTGGGCGCTGTTTGGCAGTGATGATTGCCTTTCTG...	3
ATGCAACAGCATTTTGAATTTGAATACCAGACCAAAGTGGATGGTG...	3

Описание набора данных

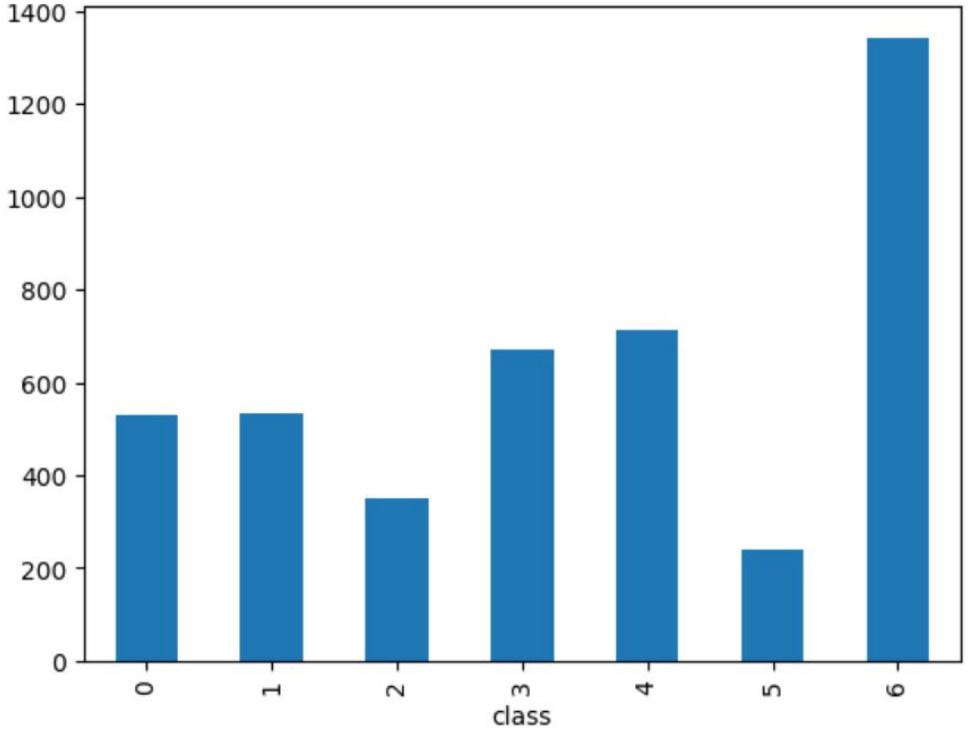
В наборе данных нет пропусков. Для каждой из 4380 последовательностей определены классы.

```
Data columns (total 2 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   sequence  4380 non-null   object
 1   class     4380 non-null   int64
dtypes: int64(1), object(1)
```

Описание набора данных

Набор данных не является сбалансированным, классы 5, 6 имеют большой разброс в количестве объектов.

Учитывая это, в качестве критериев оценки моделей будут использованы матрица ошибок и взвешенная F-мера.

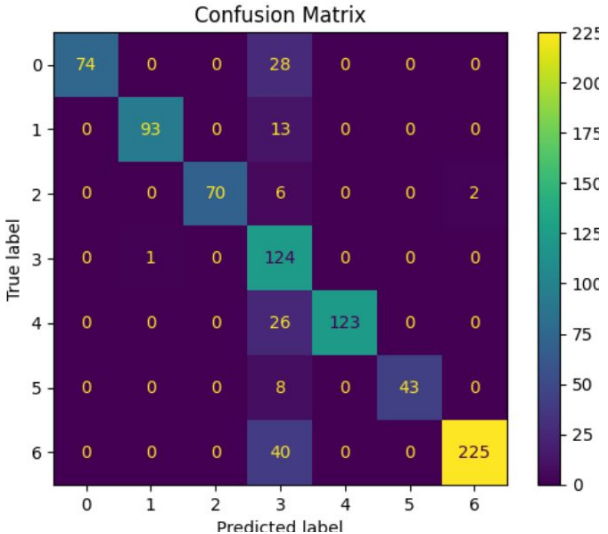


Модели. KNN

Метод k-ближайших соседей предполагает, что расположенные близко друг к другу объекты в пространстве признаков принадлежат к одному классу. Используется со стандартной Евклидовой метрикой и k в диапазоне от 1 до 5. По оценке взвешенной F-меры в итоге был выбран k = 1.

Предобработка входных данных

Для задач классического машинного обучения последовательности разбиваются на слова по 6 символов. Для каждой последовательности слов строится частотный словарь, на основе которого последовательность преобразуется в вектор.



Оценки работы алгоритма

```
accuracy = 0.858
precision = 0.926
recall = 0.858
f1 = 0.874
```

Модели. Дерево решений

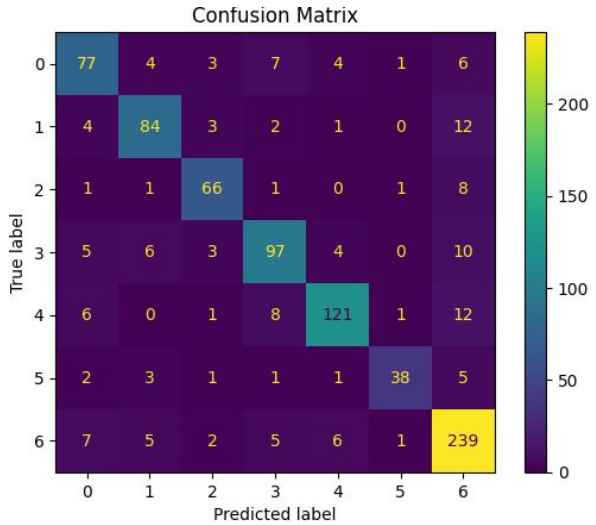
Дерево решений — это алгоритм машинного обучения для классификации и регрессии, представляющий данные в виде структуры узлов и листьев. Основные параметры включают:

```
'criterion': ['gini', 'entropy']
'max_depth': [None, 10, 20, 30],
'min_samples_split': [2, 5, 10],
'min_samples_leaf': [1, 2, 4],
```

По итогу был выбран: min_samples_split=5, random state=42.

Оценки работы алгоритма:

```
accuracy = 0.824
precision = 0.826
 recall = 0.824
 f1 = 0.824
```



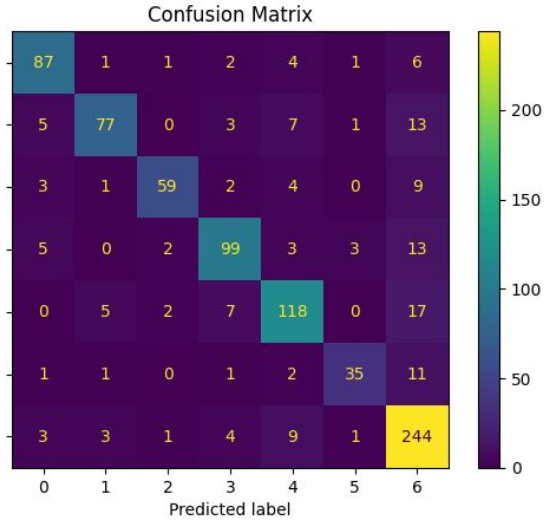
Модели. Модель глубокого обучения

Описание. Модель глубокого обучения использует представление последовательностей в виде k-меров (триплеты нуклеотидов), применяет встраивания (embeddings) для преобразования k-меров в плотное представление и использует свёрточные слои (CNN) для извлечения локальных признаков и завершает обработку полносвязными слоями для окончательной классификации.

Предобработка входных данных. Каждая последовательность ДНК разбивается на перекрывающиеся подстроки длины 3 (AGT). Все уникальные k-меры из всех последовательностей извлекаются и добавляются в словарь, где каждому k-меру назначается уникальный индекс.Затем каждая последовательность представляется как последовательность индексов k-меров на основе словаря. После уравнивается длина последовательностей для корректного ввода нейронной сети.

Оценки работы алгоритма.

```
accuracy = 0.821
precision = 0.826
recall = 0.821
f1 = 0.819
```



Результаты исследований

По оценке F-меры (f1):

1	KNN
2	Дерево решений
3	Модель глубокого обучения

По оценке точности (accuracy):

1	KNN
2	Дерево решений
3	Модель глубокого обучения

По оценке Confusion Matrix:

1	KNN
2	Дерево решений
3	Модель глубокого обучения