Нелинейные ML-модели

Годовой проект

Нелинейные ML-модели

В ходе эксперимента были исследованы различные модели машинного обучения для задачи классификации изображений фруктов и овощей. Для извлечения признаков использовались методы: **HOG** и **SIFT**.

Оба метода применялись **к цветным изображениям размером 64x64 пикселей**.

Извлечение признаков с помощью HOG

Модель	Гиперпараметры	accuracy	f1-macro
SVC+PCA	n_components=0.6	0.70	0.70
SVC+PCA	n_components=0.6, C=10, kernel='rbf'	0.76	0.76
RandomForest+PCA	n_components=0.6	0.76	0.76
RandomForest+PCA	n_components=0.6, criterion='entropy', max_depth=None, max_features='sqrt', n_estimators=500	0.77	0.77
LightGBM+PCA	n_components=0.6	0.74	0.74
LightGBM+PCA	n_components=0.6, min_child_samples=12, num_leaves=60, reg_alpha=2.8841108732861e-05, reg_lambda=2.4410628100010748e-08	0.78	0.78
CatBoost+PCA	n_components=0.6	0.76	0.76
CatBoost+PCA	n_components=0.6, depth=10, learning_rate=0.1, min_child_samples=44, reg_lambda=0.051712194163615596	0.79	0.79

Извлечение признаков с помощью HOG

- **CatBoost с использованием HOG показал наилучшие результаты**: **accuracy = 0.79 и f1 macro-averaged = 0.79**. Это говорит о том, что CatBoost хорошо справляется с задачей классификации на основе признаков, извлеченных с помощью HOG.
- LightGBM также показал хорошие результаты (accuracy = 0.78), что подтверждает эффективность градиентного бустинга для данной задачи.
- RandomForest показал стабильные результаты (accuracy = 0.77), но немного уступил CatBoost и LightGBM.
- SVC с ядром RBF показал accuracy = 0.76, что также является достойным результатом, но требует больше вычислительных ресурсов и времени по сравнению с деревьями и бустингом.

Во всех экспериментах с HOG **использовался метод PCA** для уменьшения размерности данных. Это позволило сохранить 60% дисперсии (n_components=0.6) и ускорить обучение моделей. Также предыдущие опыты показали, что PCA оказался полезен для избежания переобучения.

Извлечение признаков с помощью SIFT

Модель	Гиперпараметры	Размер изображения	Цветное	accuracy Test
RandomForest	n_estimators: 100, criterion: "gini", max_depth: None, max_features: "sqrt"	64px	цветное	0.6
RandomForest	n_estimators: 200, criterion: "entropy", max_depth: None, max_features: "sqrt"	64px	цветное	0.65
LightGBM	min_child_samples: 20, num_leaves: 31, reg_alpha: 0, reg_lambda': 0	64px	цветное	0.67
LightGBM	min_child_samples: 66, num_leaves: 165, reg_alpha: 0.00093, reg_lambda': 0.00074	64px	цветное	0.71
CatBoost	depth: None, learning_rate: None, min_child_samples: None, 'reg_lambda': None	64px	цветное	0.69
CatBoost	depth: 10, learning_rate: 0.1, min_child_samples: 63, 'reg_lambda': 7.77e-05	64px	цветное	0.7

Извлечение признаков с помощью SIFT

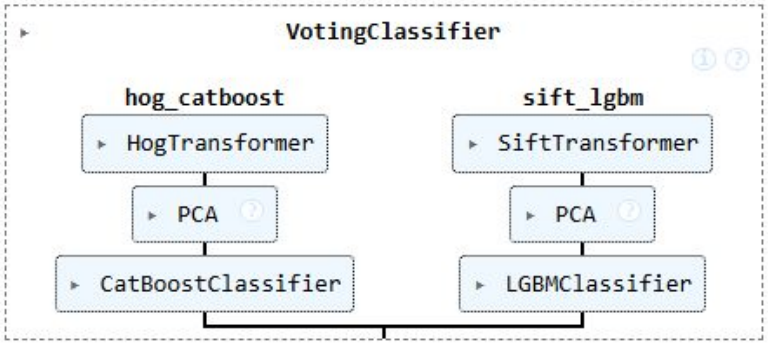
- **Для SIFT наилучшие результаты показал LightGBM (accuracy = 0.71)**
- Также как и для HOG лучшие результаты показали бустинги: LightGBM и CatBosst, что еще раз подтверждает эффективность бустинга в нашей задаче.
- В сравнении с HOG извлечение признаков с помощью SIFT менее эффективно в данной задаче.

VotingClassifier

Также в качестве эксперимента были объединены лучшие модели для HOG и SIFT с помощью VotingClassifier.

Попытка объединить лучшие модели для HOG и SIFT с помощью VotingClassifier **не привела к улучшению результатов**. Метрики немного ухудшились **accuracy = 0.76, f1 macro-averaged = 0.75** (по сравнению с лучшими результатами для HOG).

- Модели, обученные на SIFT, имеют более низкую точность, что может ухудшать общее качество ансамбля.
- HOG и SIFT извлекают разные типы признаков, и их комбинация может не давать синергетического эффекта.



Вывод

Из нелинейных ML-моделей лучше всего использовать извлечение признаков с помощью HOG, уменьшение размерности с помощью PCA и бустинг CatBoost. accuracy и macro-averaged f1 = 0.79

- Для ускорения можно заменить на LightGBM. accuracy и macro-averaged f1 = 0.78, но работает быстрее.