Сравнение алгоритмов Machine Learning в решении задач распознавания изображений

В. Д. Лука, email: lucavd120201@gmail.com

Санкт-Петербургский государственный экономический университет[1]

Аннотация. *В данной работе сравниваются четыре алгоритма Machine Learning в решении задач распознавания черно-белых изображений цифр с использованием уменьшения количества признаков в наборе данных и без.*

Ключевые слова: *Machine Learning, задачи классификации, Decision Tree, Logistic Regression, SVC, KNN, tSNE, PCA, распознавание рукописных чисел.*

Введение

Машинное обучение – это раздел науки об искусственном интеллекте, который изучает способы решения задач путем поиска закономерностей в наборе исходных данных при помощи алгоритмов и моделей, построенных на вычислительных и статистических методах. Главной особенностью данного подхода к решению прикладных задач является отсутствие явной итоговой инструкции, выполняемой разработанной математической моделью. Кроме того, необходимость применения такой формы искусственного интеллекта возникает преимущественно при обработке ресурсами компьютера больших массивов данных, с которым не может эффективно справляться исследователь с более простыми аналитическими инструментами.

Таким образом, использование алгоритмов Машинного обучения, как инструментов в решении прикладной задачи, предполагает наличие 3 основных компонентов: *компьютерная среда*, организующая и реализующая весь процесс, *исходный набор данных* значительного объема, *математическая модель*, выполняющая поиск закономерностей в данных с помощью вычислительных и статистических методов для вывода требуемого результата, а также преобладание *принципа черного ящика* во время применения модели.

Благодаря алгоритмам Машинного обучения крупные компании, накопившие, сохраняющие и продолжающие сбор большого массива данных, способны извлекать из них пользу, оптимизируя собственные бизнес-процессы, внедряя умные модели и решая сложные задачи. Важно отметить, что сфера применения Машинного обучения

[1] Лука В. Д., 2023

разнообразна и широка. Бизнес может совершенствовать процессы производства товаров, предоставления финансовых услуг, розничной торговли, здравоохранения, маркетинга и рекламы. Кроме того, описываемая область искусственного интеллекта применима и в государственном управлении.

Алгоритмы Машинного обучения в информационном пространстве обычно делят на две стандартные группы относительно способа обучения модели и требуемого результата:

- Машинное обучение с учителем
- Машинное обучение без учителя

Первая группа алгоритмов используется в задачах, при которых имеется набор данных, состоящий из определенного числа признаков и конечных результатов, относящихся к данным признакам. Целью в данной ситуации является оценка корреляции признаков и соответствующих результатов и построение математической модели, понимающей закономерности в данных, для последующего распознавания нового набора признаков, у которых не определена конечная категория или значение. Для оценки корректности прогнозных значений применяются различные метрики, сопоставляющие их с истинными результатами, что, например, можно делать через среднюю ошибку. Классическими типами задач, которые решаются данной группой алгоритмов являются задачи регрессии, прогнозирования, классификации и ранжирования.

Вторая группа алгоритмов применяется в задачах, условия которых предполагаются исключительно наличие входных признаков. То есть целью в данной ситуации будет определение значимых связей внутри исходного набора данных. Данный случай также предполагает поиск закономерности, однако уже в зависимостях между объектами. В такой задаче нет возможности предварительно однозначно определить ответы модели, поэтому исследователь опирается в конечном принятии решения только на специфику и ожидаемые результаты конкретной задачи, из-за чего точность прогноза становится относительной. Данные алгоритмы используются в задачах кластеризации, уменьшения размерности набора данных, что применимо для компьютерного зрения, фильтрации выбросов и обнаружения аномалий, заполнении пропущенных значений, поиска ассоциативных правил, например для определения значения слов в тексте или потенциально нужного товара клиенту.

Стоит отметить, что в случае обучения с учителем, как становится понятно, исследователю требуется набор данных с уже готовыми результатами, что часто становится невозможно из-за размера заранее

неразмеченных объектов, однако данной проблемы нет в обучении без учителя, что является некой компенсацией за невозможность делать объективно точный прогноз.

Кроме того, существуют и другие популярные группы алгоритмов Машинного обучения, которыми являются, например, обучение с подкреплением и нейронные сети, однако в данной статье они не будут использоваться на практике.

1. Постановка цели и задач

Цель исследования. Автор в рамках этой статьи пытается экспериментальным путем сравнить модели для классификации черно-белых изображений среди 4 выбранных алгоритмов Машинного обучения, с использованием понижения числа используемых признаков и нет.

Задачи. Для того, чтобы выполнить поставленную цель автор ставит следующие задачи.

1. Последовательно выполнить настройку гиперпараметров обучаемых моделей на наборе данных изображений цифр от нуля до девяти и сформировать таблицу результатов работы каждого алгоритма и произвести сравнение.
2. Для каждой модели произвести уменьшение количества признаков и подвести итоги для набора данных изображений чисел.
3. Сформулировать общий вывод по проделанной работе о сравнении моделей для классификации черно-белых изображений.

Используемые метрики. Оценка точности моделей будет происходить с трех сторон. Во-первых, с помощью accuracy будет рассчитана доля верных предсказаний относительно количества изображений в наборе. Во-вторых, применяя precision будет получена доля верных предсказаний относительно суммы истинно верных и ошибочно предсказанных правильными для каждого класса в выборке. В-третьих, с помощью recall будет рассчитана доля верных предсказаний относительно суммы истинно верных и пропущенных правильных для каждой группы картинок в наборе данных. Также будут зафиксированы такие параметры, как количество используемых признаков и время обучения модели.

Используемые алгоритмы. Логистическая регрессия (Logistic Regression), машина опорных векторов (SVC вид Support Vector Machine), метод ближайших соседей (K-nearest neighbors, KNN) дерево решений (Random Forest Classifier, RFC), метод главных компонент

(Principal Component Analysis, PCA), стохастическое вложение соседей с t-распределением (t-distributed Stochastic Neighbor Embedding, t-SNE).

2. Сравнение результатов предсказания моделей до уменьшения размерности данных

Рассмотрим особенности набора данных с изображениями цифр. С одной стороны, он состоит из 1797 объектов, которые образуют десять групп рисунков цифр от нуля до девяти в среднем по 160 штук. То есть классы являются сбалансированными и явного преобладания одного над другим не наблюдается. С другой стороны, каждый объект состоит из 64 числовых признаков, отвечающих за степень темноты данного пикселя в итоговом изображении. Совокупность признаков образует рисунок размером восемь на восемь пикселей. Для каждого объект определено его значение в виде цифры.

Для тренировки моделей была выбрана среда Python. Исходный набор данных был разделен методом train_test_split на тренировочную выборку и тестовую в соотношении 1 к 4, то есть тестовая часть составила 20% от всех данных. Далее автор последовательно подбирал оптимальные гиперпараметры для каждой модели.

Для KNN подбиралось эффективное для предсказания число соседей от 2 до 10 и выбирался способ расчета расстояния между ними между 'minkowski' и 'manhattan'. Оптимальными оказались гиперпараметры 'minkowski' и число соседей равное трем.

Для Логистической Регрессии изменялись максимальное число итераций в диапазоне от 5 до 51 с шагом 5 и обратная силе регуляризации (параметр C) среди коэффициентов 0.005, 0.025, 0.035 и 0.050. Лучшими оказались C равное 0.005 и макс. число итераций равное 30.

Для Дерева Решений подбирались параметры 'max_depth' в диапазоне от 3 до 21 и 'min_samples_split' в диапазоне от 2 до 7, из которых были выбраны для максимальной глубины – 11 и минимальному числу объектов для разделения – 4.

Для SVC определялись параметры C среди 5, 10, 15, 20, 25 и 30, а также gamma среди 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.1. В результате оптимальными оказались C равное 10 и Гамма равная 0.0005.

Затем был выполнен итоговый тест для каждой натренированной модели и зафиксированы результаты измерений по заявленным метрикам. Таким образом, часть из метрик была сформулирована в виде таблица ниже, где P – это precision, а R – это recall, которая далее будет описана.

Таблица 1

Сравнение точности прогноза моделей до уменьшения размерности данных

	KNN		**LR**		**DT**		**SVM**	
Цифра	*P*	*R*	*P*	*R*	*P*	*R*	*P*	*R*
0	1	1	1	1	0,94	0,88	1	1
1	0,97	1	0,96	0,96	0,85	0,79	1	1
2	1	1	0,94	1	0,81	0,76	1	1
3	0,97	1	1	0,97	0,77	0,88	1	0,97
4	0,98	1	1	0,98	0,77	0,89	1	1
5	0,98	0,98	0,91	0,91	0,88	0,91	0,98	0,98
6	0,97	1	0,97	0,97	0,94	0,94	0,97	1
7	1	0,97	1	0,97	0,88	0,85	0,97	0,97
8	1	0,97	0,97	0,97	0,84	0,7	1	1
9	0,97	0,93	0,93	0,95	0,87	0,82	0,97	0,97

По данной таблице видно, что все модели, кроме Дерева Решений отлично справились с задачей классификации, даже не прибегая к уменьшению количества признаков, о чем говорит число единиц. Кроме того, попарно сравнивая метрики у каждой модели можно установить, что для KNN в рамках данной задачи характерна лучшая точность по recall, для Logistic Regression – это precision, для Decision Tree– это также precision, а для SVM количество случаев, где преобладал recall или precision, одинаково. Из чего можно предположить, что модели Машинного обучения в задачах распознавания изображений без использования уменьшения размерности склонны к лучшему прогнозированию каждого класса понемногу, чем нескольких особо отличительных.

Далее набор данных был исследован алгоритмами PCA и tSNE на лучший способ уменьшения размерности. Для этого каждым из них автор понижал количество признаков объектов и сравнивал визуализации полученных результатов в 2D и 3D. Оказалось, что оптимальным числом размерности пространства параметров является два, о чем свидетельствует изображения ниже.

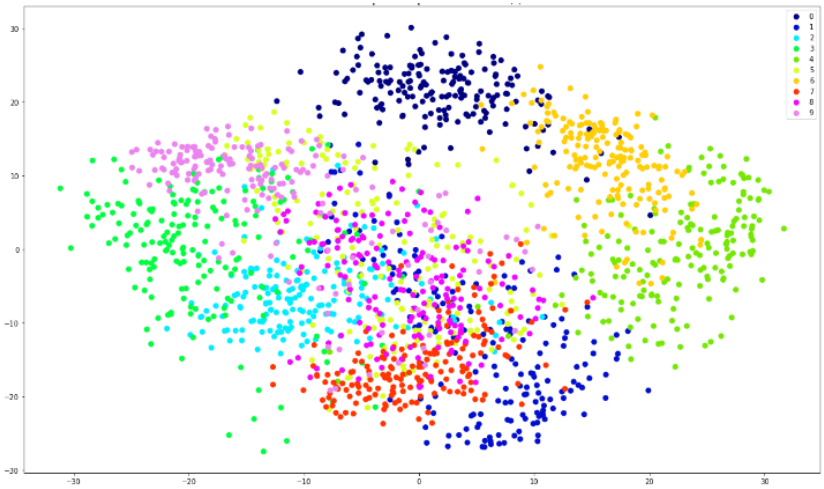

Рис. 1. Уменьшение размерности методом PCA

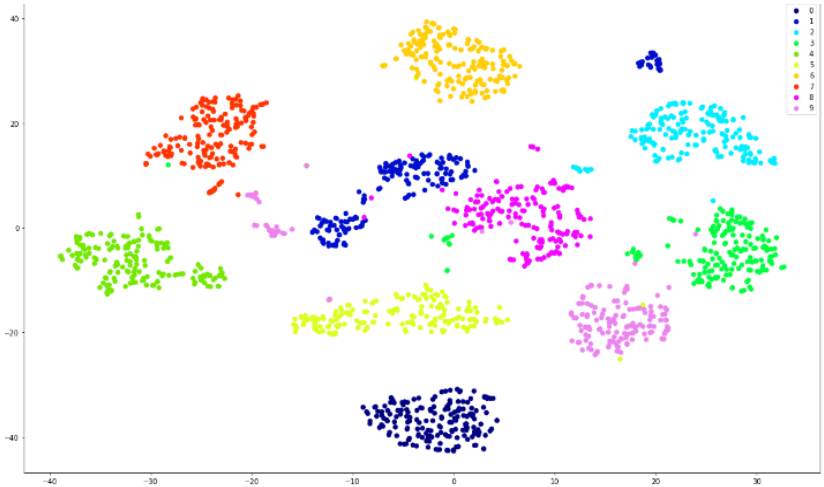

Рис. 2. Уменьшение размерности методом t-SNE

Также видно, что нелинейный tSNE значительно удачнее разбил цифры на кучки, чем его линейный конкурент PCA, поэтому он и был использован далее для создания данных для обучения новых моделей.

По предыдущим данным о выборе оптимальных гиперпараметров для модели KNN оказалось, что число соседей теперь лучше выбирать 2.

Для Логистической Регрессии диапазон выбора оптимального максимальной числа итераций остался прежним и лучшим значением стало 45, а вот коэффициент C уже выбирался среди 0.001, 0.002, 0.004, 0.005 и первый вариант был наиболее эффективным.

Для Дерева Решений в тех же диапазонах поиска лучших гиперпараметров глубина была выбрана равной девяти, а число разделений понизилось до двух.

Для SVC определялись параметры C среди 0.05, 0.1, 0.2, 0.4, а также gamma среди 0.025, 0.05, 0.2, 0.4. В результате оптимальными оказались C равное 0.2 и Гамма равная 0.05. Новый результаты по recall и precision отражены в таблице ниже.

Таблица 2

Сравнение точности прогноза моделей после уменьшения размерности данных

	KNN		**LR**		**DT**		**SVM**	
Цифра	*P*	*R*	*P*	*R*	*P*	*R*	*P*	*R*
0	1	1	1	1	1	1	1	1
1	0,97	1	0,77	0,82	0,97	1	0,97	1
2	1	1	0,87	1	1	1	1	1
3	0,97	1	0,92	1	1	1	0,97	1
4	0,98	1	0,96	1	0,98	1	0,98	1
5	0,96	0,98	0,98	0,98	0,98	0,98	0,98	0,98
6	0,97	1	0,97	1	0,97	1	0,97	1
7	1	0,97	0,92	1	1	0,97	1	0,97
8	1	0,97	1	0,77	1	0,97	1	0,97
9	0,97	0,9	0,97	0,75	0,97	0,95	0,97	0,93

По новым данным видно, что существенных потерь точности для алгоритмов в прогнозе цифры нет, и можно даже заметить, как резко выросла точность для Дерева Решений, опираясь на количество единиц. Кроме того, стоит отметить и смещение роли доминирующей метрики в сторону recall, что означает более резкое разделение данных. То есть некоторые цифры теперь точно можно определить и модели не отдают предпочтение каждому классу понемногу, из-за чего общая точность либо выросла, как у Дерева Решений, либо незначительно снизилась. Это кажется логичным, учитывая, что автор уменьшил размерность признаков с 64 до 2. То есть новые параметры объектов позволяют их

резче отделять друг от друга, игнорируя выбросы, что с одной стороны эффективно, когда нет очень похожих друг на друга классов, и наоборот, как в случае с изображением девятки и тройки.

Теперь рассмотрим сравнение общих зафиксированных параметров по моделям до уменьшения размерности и после, что отражено в таблице ниже.

Таблица 3

Сравнение показателей обучения и тестирования по моделям с использованием уменьшения размерности и без

Этап	**Модель**	**Accuracy**	**Акцент**	**Время обуч.**	**Кол-во призн.**
Без уменьшения размерности	KNN	0.98	recall	0.00226s	64
	LR	0.97	precision	0.09673s	64
	DT	0.85	precision	0.01833s	64
	SVM	0.99	нет	0.04998s	64
С уменьшением размерности	KNN	0.98	recall	0.00103s	2
	LR	0.94	recall	0.06083s	2
	DT	0.99	нет	0.00304s	2
	SVM	0.98	recall	0.06315s	2

Опираясь на полученные данные, можно сделать следующие выводы. Уменьшение размерности позволило сократить число признаков с 64 до 2 и существенно ускорить процесс обучения для трех из четырех моделей без ощутимой потери точности. В случае, когда размерность входных данных составляет два, все модели отлично справились с распознаванием изображений, однако в случае с SVC и KNN, данный тренд сохраняется и для 64 признаков, хотя и с потерей времени обучения у модели на методе ближайших соседей. Стоит также сказать, что SVC оказался самым эффективным и сбалансированным по всем метрикам для данного типа задач.

Заключение

В качестве общих рекомендаций по выбору модели для классификации черно-белых изображений можно заключить следующее. Если исследователь хочет быстро провести процесс выбора оптимальной по гиперпараметрам модели ему стоит попробовать алгоритмы KNN и DT причем скорость будет увеличена, если количество признаков будет снижено. Также стоит помнить, что Логистическая Регрессия требует сравнительно много времени для обучения, что может затянуть выбор оптимальных гиперпараметров

Если исследователь по условиям задачи может пренебречь точность распознавания некоторых классов, то он может использовать алгоритмы по уменьшению размерности исходных данных, тем самым увеличив резкость разделения классов между собой, что позволит модели точнее определять особо отличительные группы объекты. Также исследователь должен помнить, что уменьшение количества признаков исходного набора данных может как увеличить точность распознавания, так и понизить ее в зависимости от алгоритма и степени уменьшения размерности. Кроме того, можно отметить, что алгоритмы Логистической регрессии и Дерева Решений более размыто делят схожие объектов между собой, в то время как модели, использующие метод ближайших соседей, наоборот делает это резче. Более того, исследователь обязательно должен попробовать SVC для получения максимально сбалансированного результата в части точности и времени обучения.

Список литературы

1. Галиахметов, Д. Г. Сравнение алгоритмов классификации применительно к задаче обнаружения вредоносных доменных имен / Д. Г. Галиахметов // Математические методы в технике и технологиях - ММТТ. – 2019. – Т. 12-1. – С. 190-194.
2. Э. А. Челышев, Сравнение методов классификации русскоязычных новостных текстов с использованием алгоритмов машинного обучения / Э. А. Челышев, Ш. А. Оцоков, М. В. Раскатова, П. Щеголев // Вестник кибернетики. – 2022. – № 1(45). – С. 63-71.
3. Агапитов, Д. В. Сравнение эффективности алгоритмов традиционного машинного обучения и нейронных сетей в задаче классификации / Д. В. Агапитов, Я. А. Колташев, К. И. Брагин // Научный альманах Центрального Черноземья. – 2022. – № 1-1. – С. 5-14.
4. Документация. sklearn SVC. [Электронный ресурс]. – Режим доступа: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
5. Документация. sklearn KNN [Электронный ресурс]. – Режим доступа: https://scikitlearn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
6. Документация. sklearn.tree.DecisionTreeClassifier [Электронный ресурс]. – Режим доступа: https://scikitlearn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

7. Статья о Машинном обучении на Amazon [Электронный ресурс]. – Режим доступа: https://aws.amazon.com/ru/what-is/machine-learning/

8. Статья о Машинном обучении на BigDataSchool [Электронный ресурс]. – Режим доступа: https://www.bigdataschool.ru/wiki/machine-learning

9. Статья о метриках Машинного обучения на Pythonru.com [Электронный ресурс]. – Режим доступа: https://pythonru.com/baza-znanij/metriki-accuracy-precision-i-recall