МИНОБРНАУКИ РОССИИ

Федеральное государственное автономное образовательное
учреждение высшего образования
«Южный федеральный университет»

Институт математики, механики
и компьютерных наук им. И. И. Воровича

Кафедра информатики и вычислительного эксперимента

Тупикин Олег Витальевич

ИСПОЛЬЗОВАНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ РЕШЕНИЯ ЗАДАЧ КЛАССИФИКАЦИИ

КУРСОВАЯ РАБОТА
по направлению подготовки
02.03.02— Фундаментальная информатика и информационные технологии

Научный руководитель –
ст. преп. Ячменева Наталья Николаевна

100 _____
оценка (рейтинг)

_____
подпись руководителя

Ростов-на-Дону – 2024

Задание на курсовую работу
студента Тупикина О. В.

Направление подготовки: фундаментальная информатика и информационные технологии.

Студент: О. В. Тупикин

Научный руководитель: ст. преп. Н.Н. Ячменева

Год защиты: 2024

Тема работы: ИСПОЛЬЗОВАНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ РЕШЕНИЯ ЗАДАЧ КЛАССИФИКАЦИИ

Цель работы: Используя возможности методов машинного обучения реализовать задачу классификации объектов в любой предметной области.

Задачи работы:

- Определить предметную область для задачи классификации. Подобрать набор примеров для обучения.
- Изучить возможности библиотек машинного обучения.
- Выполнить анализ и предобработку данных.
- Реализовать систему классификации с использованием методов машинного обучения.
- Выполнить поиск оптимальной модели.

Научный руководитель Н.Н. Ячменева

Студент О. В. Тупикин

25 ноября 2023 г

Отзыв на курсовую работу Тупикина О. В. на тему

«ИСПОЛЬЗОВАНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

ДЛЯ РЕШЕНИЯ ЗАДАЧ КЛАССИФИКАЦИИ»

Задачи классификации объектов часто решаются с помощью методов машинного обучения. Работа Олега посвящена этому вопросу на примере задачи классификации стекла на основе его химического состава. Студент применил библиотеки машинного обучения в решении задачи, использовал алгоритмы, нейронные сети и возможности сервиса Google Colaboratory и др. В работе выполнены анализ и предобработка данных обучающей выборки. Представлены результаты работы различных классификаторов, выполнен их сравнительный анализ.

В учебном году О. В. Тупикин работал систематически и добросовестно, проявил самостоятельность и активность, выполнил все задания. Считаю, что заслужил оценку «отлично» (100 баллов).

Научный руководитель,

Ст.преп. ИММиКН Н.Н. Ячменева

АНТИПЛАГИАТ
ОБНАРУЖЕНИЕ ЗАИМСТВОВАНИЙ

СПРАВКА

о результатах проверки текстового документа
на наличие заимствований

Южный федеральный университет

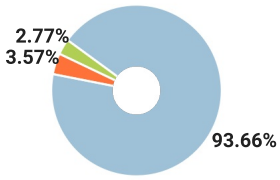
ПРОВЕРКА ВЫПОЛНЕНА В СИСТЕМЕ АНТИПЛАГИАТ.ВУЗ

Автор работы: Тупикин Олег Витальевич
Самоцитирование рассчитано для: Тупикин Олег Витальевич
Название работы: Тупикин_Олег_ФИИТ_2024
Тип работы: Курсовая работа
Подразделение: ИММиКН им. Воровича

РЕЗУЛЬТАТЫ

СОВПАДЕНИЯ	3.57%
ОРИГИНАЛЬНОСТЬ	93.66%
ЦИТИРОВАНИЯ	2.77%
САМОЦИТИРОВАНИЯ	0%

ДАТА ПОСЛЕДНЕЙ ПРОВЕРКИ: 04.06.2024



Структура документа: Проверенные разделы: библиография с.18, содержание с.2, основная часть с.3-17

Модули поиска: ИПС Адилет; Перефразирования по Интернету (EN); Цитирование; Публикации eLIBRARY; Диссертации НББ; Перефразирования по коллекции издательства Wiley; Коллекция НБУ; Публикации eLIBRARY (переводы и перефразирования); Кольцо вузов (перефразирования); Шаблонные фразы; IEEE; Переводные заимствования по Интернету (EnRu); Медицина; Перефразирования по Интернету; СПС ГАРАНТ: аналитика; Перефразированные заимствования по коллекции Интернет в русском сегменте; Переводные заимствования по коллекции Интернет в русском сегменте; Перефразирования по коллекции IEEE; Кольцо вузов; Патенты СССР, РФ, СНГ; Издательство Wiley; Переводные заимствования по коллекции Интернет в английском сегменте; Переводные заимствования*; Библиография; Публикации РГБ; СМИ России и СНГ;

Работу проверил: Ячменева Наталья Николаевна

ФИО проверяющего

Дата подписи:

Подпись проверяющего

Чтобы убедиться в подлинности справки, используйте QR-код, который содержит ссылку на отчет.

Ответ на вопрос, является ли обнаруженное заимствование корректным, система оставляет на усмотрение проверяющего. Предоставленная информация не подлежит использованию в коммерческих целях.

Содержание

Введение

Развитие и использование искусственного интеллекта(ИИ) в наше время обретает всё большую популярность. Во многие отрасли, например — медицину и экономику, интегрируют узконаправленные модели ИИ для решения той или иной задачи. Большинство из них используют в своей реализации методы машинного обучения [1] для обучения и настройки моделей.

Выделяют шесть основных задач машинного обучения: регрессии, классификации, кластеризации, идентификации, прогнозирования и извлечения знаний. В данной исследовательской работе мы будем рассматривать задачи классификации, методы их решения и проведем сравнительный анализ эффективности методов на выбранной предметной области.

Под «классификацией» объектов понимают разделение множества объектов по принадлежности к тому или иному классу. Яркими примерами решения задач классификации могут послужить следующие процессы: определение спама электронной почты, наличие онкологического заболевания у пациента, обнаружение объектов на изображении (например, БПЛА в небе, животных в лесу) и т. д.

Постановка задачи

Целью исследовательской работы является изучение способов решения задач классификации с помощью применения различных методов машинного обучения к реальному набору данных и сравнительный анализ их эффективности, посредством применения высокоуровневого языка общего назначения Python с подключением библиотек обработки данных — NumPy и Pandas, библиотеки для машинного обучения Scikit-learn (sklearn) [2], а также библиотек для визуализации данных Matplotlib и Seaborn.

1. Анализ и обработка данных

Прежде, чем начать обработку данных, опишем набор данных, используемый в данной работе. Будем рассматривать датасет химического состава стекла [3], состоящий из 214 объектов наблюдений, разделённых на классы positive и negative. Особенностью данного набора является то, что количество объектов, принадлежащих одному из классов гораздо превосходит количество объектов, принадлежащих к другому.

1.1 Анализ набора данных

Набор данных, как было сказано ранее, состоит из 214 объектов. В качестве признаков представлены следующие данные числового типа: RI — коэффициент преломления, Na — натрий, Mg — магний, Al — алюминий, Si — силикон, K — калий, Ca — кальций, Ba — барий, Fe — железо.

В наборе отсутствуют пропущенные значения и дубликаты – следовательно их обработка не требуется.

На круговой диаграмме (см. рис. 1) заметно несбалансированное распределение классов. Из 214 элементов 185 принадлежит классу negative, и 29 — positive.

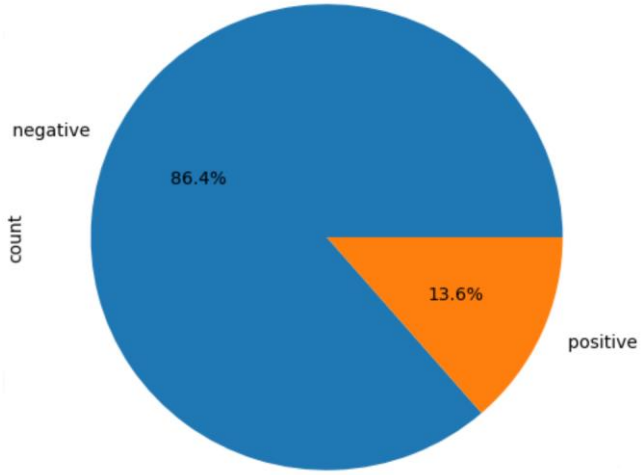


Рисунок 1. Круговая диаграмма распределения классов

Рассмотрим зависимости признаков между собой, включая классы. Для этого построим матрицу (тепловую карту) корреляции. (см. рис. 2).

Рисунок 2. Матрица корреляции

По данной матрице видно, что признаки не коррелируют между собой. Деление на классы происходит под сильным влиянием Mg и Ba, а так же небольшого влияния Al и K.

Сгруппируем объекты исследования по классам, и посмотрим на средние значения признаков для каждой группы. Результаты приведены на рисунке 3:

Class	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
negative	1.227291e+08	732635.348649	279190.657292	106578.925564	534583.520054	1504.336213	726268.780703	118.879144	0.043750
positive	1.230051e+08	800020.068966	47005.655172	190670.724138	566745.523793	25268.348241	637251.911379	41664.981026	0.035348

Рисунок 3. Результаты группировки по классам

Легко заметить зависимость принадлежности к группе от значений признаков Mg и Ba. Также заметен разброс значений столбцов. Проанализируем максимальные значения по каждому признаку:

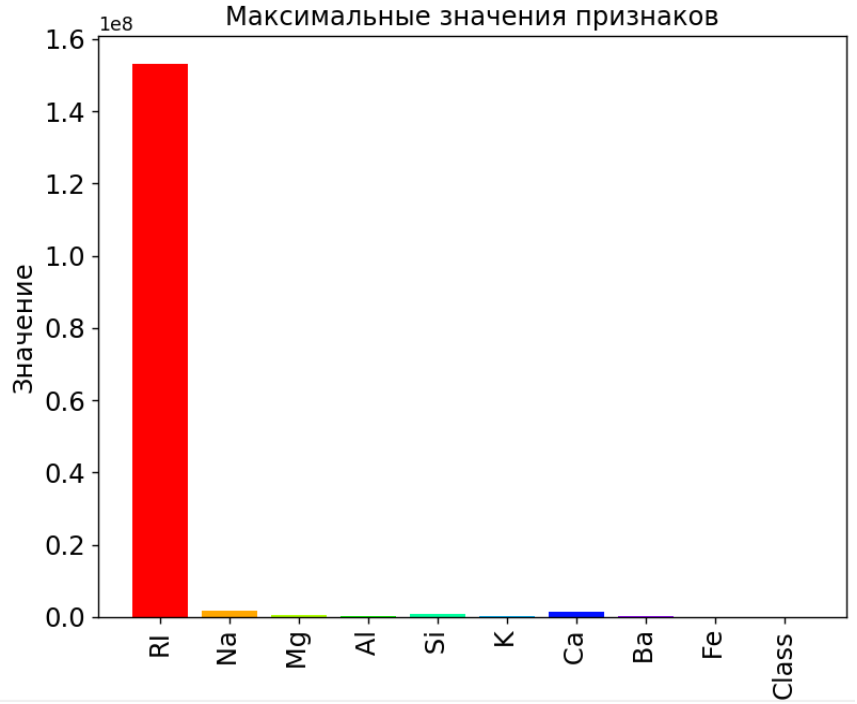


Рисунок 4. Максимальные значения признаков

В наборе содержатся аномально большие значения, что можно заметить по гистограмме на рисунке 4. Значения признака «коэфициент преломления» зашкаливают по сравнению с другими признаками.

1.2 Обработка данных

Прежде, чем начать работу с набором данных, следует обработать его. Обычно обработка состоит из:

1) обработки повторяющихся строк (дубликатов);
2) обработки пропущенных значений;
3) приведения всех типов столбцов к числовому типу;
4) нормализации/стандартизация значений;
5) обработки выбросов;

и другого при необходимости.

Так как при анализе мы определили, что в нашем наборе отсутствуют дубликаты и пропущенные значения, а также все признаки имеют числовой тип, то обработку данных случаев можно опустить, поэтому сразу перейдём к стандартизации значений.

1.2.1 Стандартизация значений

На рисунке 4 заметен разброс значений признаков. Чтобы избежать сложностей с их использованием алгоритмами машинного обучения – преобразуем значения с помощью модуля preprocessing библиотеки sklearn:

```
# Преобразование текущего датафрейма в матрицу чисел
features = df.drop('Class', axis=1).to_numpy()

# Создание шкалировщика и стандартизация признаков
standart_scaler = preprocessing.StandardScaler()
standardized_features = standart_scaler.fit_transform(features)
```

Теперь наши признаки имеют среднее, равное 0, со стандартным отклонением 1.0, что упрощает их обработку.

1.2.2 Обнаружение выбросов

Выбросы — аномальные значения, которые могут негативно влиять на результаты работы алгоритмов. Чтобы обеспечить правильную работу методов машинного обучения и увеличить их точность, нам необходимо обнаружить эти выбросы и обработать.

Обнаружение выбросов реализуем на основе метода Z-оценки (стандартной оценки) [4]. В его основе лежит сопоставление данных с распределением, среднее значение которого равно 0, а стандартное отклонение 1. Смысл данного алгоритма состоит в следующем: после того, как мы центрировали и масштабировали данные – всё, что находится на расстоянии большем чем R от 0, следует считать выбросом.

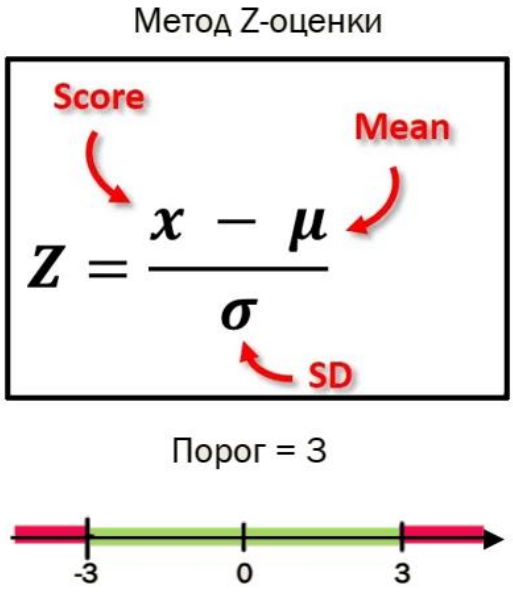


Рисунок 5. Метод Z-оценки

Центрирование и масштабирование данных мы обеспечили за счёт стандартизации признаков, поэтому теперь нужно применить наш метод к данным. Путём перебора порогового значения для метода, было решено взять значение равное 3. После применения метода получили следующее количество выбросов: «K» — 6 шт., «Ba» — 7 шт., «Fe» — 7 шт. Теперь необходимо обработать эти выбросы.

1.2.3 Обработка выбросов

Обработка выбросов достаточно неоднозначный процесс. Особенно, когда необходимо обработать значения, не имея представления о химических свойствах стекла, а значит не имея возможности самостоятельно предсказать, являются ли данные значения на самом деле выбросами. По этой причине было принято решение не ограничиваться одним методом обработки, а рассмотреть несколько из них.

Применим следующие методы:

- Удаление выбросов
- Замена выбросов на среднее значение
- Замена выбросов на среднее значение без учёта значений выбросов
- Замена на медианное значение
- Замена на максимальное/минимальное значение по признаку

Примечание: Замена на максимум/минимум определяется относительно значения выброса. Если выброс меньше минимального значения (без учёта выбросов) – заменяем его на минимум по признаку. Аналогично для выбросов, больших, чем максимум.

2. Применение методов классификации

Обучим наши модели на обработанном наборе данных с одинаковым разбиением на тренировочный и тестовый наборы (с целью лучшего сравнения алгоритмов). Рассмотрим метод логистической регрессии, наивный байесовский классификатор и метод случайного леса.

В основе метода логистической регрессии лежит прогнозирование путём сравнения с логистической кривой. Данный метод имеет несколько гиперпараметров, один из которых — параметр «C». Это параметр, который «сообщает» алгоритму, как именно обрабатывать данные в экстремальных случаях. Его будем подбирать с помощью поиска по сетке (алгоритмом GridSearchCV библиотеки sklearn), который будем использовать в большинстве рассматриваемых методов.

Алгоритм наивного байесовского классификатора основан на теореме Байеса. Его особенностью является то, что он подразумевает отсутствие зависимостей между признаками, и нечувствительный к выбросам. Данная модель не имеет гиперпараметров.

Метод случайного леса основан на построении деревьев решений, и их объединении в «лес». Он имеет следующие гиперпараметры: количество деревьев в лесу, максимальная глубина деревьев, минимальное количество элементов для разделения узла и т. д. Эти гиперпараметры также можно подобрать с помощью поиска по сетке.

После обучения моделей получили следующие результаты (см. рис. 6):

	Логистическая регрессия		Наивный Байес	Случайный лес	
	Без подбора параметров	С подбором параметров	Без подбора параметров	Без подбора параметров	С подбором параметров
1. Без обработки выбросов	95.3%	95.3%	90.7%	95.3%	93.0%
2. Удаление выбросов	97.4%	97.4%	97.4%	100%	100%
3. Замена на avg	86.0%	86.0%	97.7%	93.0%	93.0%
4. Замена на avg(без учета выбросов)	86.0%	86.0%	97.7%	95.3%	97.6%
5. Замена на медиану	86.0%	93.0%	93.0%	95.3%	93.0%
6. Замена на min/max	86.0%	95.3%	93.0%	95.3%	95.3%

Рисунок 6. Результаты методов классификации

По результатам обучения можно сделать вывод, что все методы работают с приемлемой точностью >85%. Следует выделить вариант обработки выбросов методом удаления, так как в этом случае мы получили наилучшие результаты в среднем по методам. Если рассмотреть классификаторы по-отдельности, то можно заметить, что на метод логистической регрессии сильно влияет обработка выбросов, при этом в худшую сторону. Чего нельзя сказать о наивном байесовском методе и методе случайного леса. После обработки выбросов данные методы показывают наилучшие результаты: 97.7% и 100% соответственно.

3. Применение boost’ингов

Бустинги — это мощные методы машинного обучения, в основе которых лежит построение ансамбля простых моделей. В их основе лежит «работа над ошибками» в обучении предыдущих моделей, а, следовательно, частая переобучаемость. Рассмотрим несколько из них: адаптивный, градиентный и категориальный (от Yandex).

Адаптивный бустинг (AdaBoost, Adaptive boosting) — метод, который на каждой последующей итерации рассматривает варианты решения проблем с неправильно классифицированными объектами.

Градиентный бустинг (gradient boosting) основан на минимизации функции потерь с помощью градиентного спуска.

Категориальный бустинг (CatBoost) — достаточно новый бустинг от компании Yandex, основанный на градиентном бустинге. Особенностью данного метода является автоматизированный подбор гиперпараметров, в сочетании с высокой скоростью работы.

Обучим модели, подобрав к ним лучшие параметры, и рассмотрим результаты работы (рисунок 7):

	Адаптивный бустинг		CatBoost	Градиентный бустинг	
	Без подбора параметров	С подбором параметров	Встроенный подбор пар.	Без подбора параметров	С подбором параметров
1. Без обработки выбросов	95.3%	97.6%	97.7%	95.4%	96.5%
2. Удаление выбросов	100%	97.4%	100%	97.4%	100%
3. Замена на avg	100%	97.7%	97.7%	95.3%	95.3%
4. Замена на avg(без учета выбросов)	97.7%	97.7%	97.7%	95.3%	95.3%
5. Замена на медиану	100%	97.7%	97.7%	95.3%	95.3%
6. Замена на min/max	100%	97.7%	97.7%	95.3%	97.7%

Рисунок 7. Результаты применения бустингов

По результатам работы алгоритмов заметен высокий уровень работы каждого (>95%), что превосходит по точности любой из рассмотренных ранее методов классификации. Лучшие результаты показали методы Ada-Boost и CatBoost. Стоит отметить, что подбор гиперпараметров в случае адаптивного бустинга в большинстве случаев показал результаты хуже, чем без него. С другой стороны, при подборе гиперпараметров результат стабилен, и держится на уровне ~97.5%, как и в случае с CatBoost.

Градиентный бустинг оказался также достаточно точным, но местами заметны проблемы в классификации наблюдений, возможно, связанных с небольшим тренировочным набором данных.

4. Применение ансамблевого метода классификации Voting

Voting classifier (метод голосования) — ансамблевый метод, основанный на предшествующих результатах работы нескольких моделей. Голосование способствует уменьшению переобучения модели, а также улучшению общей производительности. Тем не менее, данный метод имеет и недостаток, которым является высокое время работы.

После применения данного метода к набору данных (с подбором гиперпараметров, но без обработки выбросов), получили следующий результат (в виде матрицы ошибок):

	Negative	Positive
Negative	34	0
Positive	2	7

Для объективной оценки метода – выведем остальные метрики:

CLASSIFICATION REPORT:

	precision	recall	f1-score	support
negative	0.94	1.00	0.97	34
positive	1.00	0.78	0.88	9
accuracy			0.95	43
macro avg	0.97	0.89	0.92	43
weighted avg	0.96	0.95	0.95	43

Рисунок 8. Отчёт по Voting classification

Алгоритм отработал с точностью 95.3%, что не является наилучшим результатом среди предыдущих моделей, но является средним значением по всем методам без обработки выбросов. Двум объектам из класса positive был ошибочно присвоен класс negative. В целом же, модель показывает хорошие результаты с точки зрения точности и F1-меры.

5. Дополнительные исследования

Нестабильность результатов наводит на мысль о возможности стабилизации за счёт увеличения количества признаков, оказывающих наибольшее влияние на целевой. После применения методов к набору данных, полученных за счёт увеличения количества признаков «Mg» и «Al» до количества, равному 5, действительно увеличилась точность некоторых моделей: случайный лес (97.7%) и Voting Classifier (97.7%). Но при этом ухудшилась работа других моделей: логистическая регрессия (90.7%), GradientBoost (93%). Это говорит о том, что для некоторых моделей данных признаков, или количества наблюдений – недостаточно. Исходя из этого возникла необходимость в повторном применении данного метода, но с увеличением количества признаков «Mg», «Al», «Ba» и «K».

Результаты работы на новом наборе данных, состоящем из 14 признаков улучшили результаты, по сравнению с предыдущей попыткой.

Логистическая регрессия улучшила точность до 97%. Остальные результаты не поменялись.

Данное дополнительное исследование показало, что четырёх из девяти признаков – достаточно для классификации с высокой точностью и отсутствием обработки выбросов.

Заключение

В данном исследовании мы реализовали решение задачи классификации на примере изучения датасета химического состава стекла, сравнили работу различных классификаторов на одинаковом наборе данных с несбалансированным распределением в условиях разной обработки выбросов путём использования библиотеки, ориентированной на машинное обучение scikit-learn, обработки данных библиотеками numpy и pandas, в сочетании с визуализацией данных библиотек matplotlib и seaborn. Проанализировали работу каждого из алгоритмов по-отдельности и особенности каждого из методов классификации. По результатам исследования явно видна сильная зависимость целевого признака от некоторых нецелевых значений, что позволяет моделям в условиях несбалансированного распределения обучаться с высокой точностью прогнозирования.

Полученный результат является достаточным для решения поставленной задачи применения к реальным данным методов машинного обучения, бустингов, методов подбора параметров, обнаружения и обработки выбросов.

Литература

1. *Элбон К.* Машинное обучение с использованием Python. Сборник рецептов: Пер. с англ. — СПб.: БХВ-Петербург, 2019. — 384 с.: ил

2. Scikit-learn. Machine learning in python. — URL: *https : / / scikit-learn . org*

3. Glass-imbalanced dataset — URL: *https : / / www . kaggle . com / datasets / ba-guspurnama / glass-imbalanced / code*

4. Метод Z-оценки — URL: *https : / / colingorrie . github . io / outlier-detection . html*