Санкт-Петербургский политехнический университет Петра Великого

Институт машиностроения, материалов и транспорта

**Высшая школа автоматизации и робототехники**

Курсовая работа

**«Изучение и применение методов машинного обучения»**

по дисциплине

«Математические методы интеллектуальных технологий»

Выполнил

студент гр. 3341506/00401 А.Е. Комаров

Руководитель А.В. Бахшиев

«\_\_\_» \_\_\_\_\_\_\_\_\_\_ 2020 г.

Санкт-Петербург

2020

содержание

[ВВЕДЕНИЕ 3](#_Toc59565686)

[1. Цели и задачи 4](#_Toc59565687)

[1.1 Цели 4](#_Toc59565688)

[1.2 Задачи 4](#_Toc59565689)

[3. Анализ и обработка базы 5](#_Toc59565690)

[3.1 Выбор базы 5](#_Toc59565691)

[3.2 Анализ базы 5](#_Toc59565692)

[3.3 Определения набора признаков для обучения 7](#_Toc59565693)

[3.5 Подготовка целевых признаков 10](#_Toc59565694)

[3.5 Построения диаграмм рассеяния 11](#_Toc59565695)

[3.6 Постановка задачи машинного обучения 12](#_Toc59565696)

[4. Применение алгоритмов машинного обучения 13](#_Toc59565697)

[4.1 Выбор применяемых алгоритмов 13](#_Toc59565698)

[4.2 Применение методов машинного обучения 13](#_Toc59565699)

[Заключение 15](#_Toc59565700)

# ВВЕДЕНИЕ

В современном мире алгоритмы машинного обучения окружают нас повсюду. Термин «машинное обучение» часто используют как синоним к искусственного интеллекта, однако машинное обучение — это лишь малая часть этой сферы. Хорошим примером работы алгоритмов машинного обучения является фильтрация спам-сообщений на электронной почте. Алгоритмы обучаются и сами себя программируют для лучшей классификации сообщений и фильтрации их. Также отличным примером является реклама на сайтах. В этом случае алгоритмы также анализируют последние и наиболее часто посещаемые сайты пользователя, запоминают тематику, интересующую пользователя, и на основе этих данных предоставляют соответствующую рекламу.

В последние десятилетия машинное обучение получило резкий толчок благодаря развитию науки, увеличению вычислительных мощностей. Обычно, когда машина выполняет какую-то задачу, все ее действия выполняются по определенному программистом-разработчиком алгоритму. Машина, использующая алгоритм машинного обучения, угадывает все сама, делает вывод на основе входных данных, и чем больше этих данных, тем точнее результат ее работы.

## 1. Цели и задачи

## 1.1 Цели

В рамках настоящей курсовой работы рассматриваются следующие цели:

* изучение алгоритмов машинного обучение,
* изучение методов анализа баз данных для применения на них алгоритмов машинного обучения,
* применение вышеупомянутых алгоритмов на произвольном наборе данных.

## 1.2 Задачи

При выполнении работы затрагиваются следующие задачи:

* выбор произвольной базы данных,
* анализ выбранной базы и подготовка ее для применения алгоритмов машинного обучения,
* постановка задачи машинного обучения,
* применение нескольких алгоритмов машинного обучения на оригинальных, нормализованных и стандартизованных данных,
* анализ полученных результатов.

## 3. Анализ и обработка базы

## 3.1 Выбор базы

В качестве базы данных взята база с сайта www.kaggle.com, основанной на компьютерной игре Civilization 5. Суть датасета в общих чертах заключается в наборе признаков для каждого персонажа игры. Признаками являются черты характера персонажа, его желаний и намерений, чаще всего в вероятностном виде, например, насколько вероятно, что персона объявит войну. В качестве записей датасета выступают все персонажи этой игры, их около 40. Изначально по этой базе планируется предсказывать поведение персонажа на основе признаков его описывающих.

## 3.2 Анализ базы

В таблице 1 представлены все признаки, характеризующие персонажей. Важно заметить, что все признаки являются численными и изменяются в диапазоне от 1 до 10.

Таблица 1 – Все признаки из датасета

|  |  |
| --- | --- |
| Признак | Описание признака |
| Boldness | Насколько вероятно, что лидер планирует войну против игрока с большей военной мощью и откажется уступить торговым требованиям. |
| Chattiness | Насколько вероятно, что лидер насмехается над врагами и предупреждает их о неизбежных атаках. |
| Denounce Willingness | Насколько вероятно, что лидер осудит игрока, который злит его или ее. |
| Diplomatic Balance | ----Насколько вероятно, что лидер предложит и примет справедливые сделки. Он может также указать, как трудно лидер будет работать, чтобы остановить другие игрок от получения непреодолимых ведут. |
| Friendship Willingness | Насколько вероятно, что лидер объявит о дружбе с другим игроком. |
| Forgiveness | Насколько вероятно, что лидер забудет о прошлом преступлении, удалив связанный с ним дипломатический штраф. |
| Loyalty | Насколько вероятно, что лидер будет сотрудничать и соблюдать соглашения со своими союзниками. Это также может указывать на то, насколько вероятно лидер возобновить дружбу. |
| Meanness | Насколько вероятно, что лидер вступит в войну против игрока, которому другие игроки объявили войну. Это также может указывать на то, насколько вероятно, что лидер будет участвовать в совместных войнах со своими союзниками. |
| City-State Competitiveness | Насколько вероятно, что лидер рассердится на игрока, который дружит с одним из его или ее союзников-городов-государств. |
| Neediness | Насколько вероятно, что лидер попросит у своих союзников золото или роскошные ресурсы, давая им шанс заработать дипломатический бонус. |
| Victory Competitiveness | Насколько вероятно, что лидер рассердится на игрока, который преследует те же условия победы. |
| Warmonger Hatred | Насколько вероятно, что лидер рассердится на игрока за ведение войн и захват городов. |
| Wonder Competitiveness | Насколько вероятно, что лидер рассердится на игрока, который строит чудо, которое он или она хотел построить. |
| Air | Насколько сильно лидер сосредоточен на создании авиационных подразделений для атаки и защиты. |
| Airlift | Насколько вероятно, что лидер перебрасывает подразделения по воздуху между городами. |
| Aircraft Carrier | Насколько сильно лидер сосредоточен на производстве авиаперевозчиков. |
| Anti-Air | Насколько сильно лидер сосредотачивается на производстве юнитов, которые могут перехватывать вражеские воздушные юниты. |
| Archaeology | Насколько сильно лидер сосредоточен на подготовке археологов. |
| City Defense | Насколько сильно лидер сосредоточен на защите своих городов. |
| Culture | Насколько сильно лидер сосредоточен на увеличении продукции культуры своих городов. |
| Defense | Насколько сильно лидер сосредоточен на создании оборонительных подразделений. |
| Diplomacy | Насколько вероятно, что лидер добьется дипломатической победы. |
| Espionage | Насколько сильно лидер сосредоточен на создании полицейских частей и полицейских участков в своих городах и использовании шпионов. |
| Expansion | Насколько сильно лидер сосредотачивается на основании новых городов. |
| Gold | Насколько сильно лидер сосредоточен на увеличении добычи золота в своих городах. |
| Great People | Насколько сильно лидер фокусируется на увеличении поколения великих людей в своих городах. |
| Growth | Насколько сильно лидер сосредоточен на увеличении населения своих городов. |
| Happiness | Насколько сильно лидер сосредоточен на увеличении общего счастья своей империи. |
| Infrastructure | Насколько сильно лидер сосредоточен на соединении своих городов автомобильными и железными дорогами. |
| Land Trade | Насколько сильно лидер сосредоточен на производстве караванов. |
| Sea Trade | Насколько сильно лидер сосредоточен на производстве грузовых судов. |
| Trade Destination | Сколько внимания лидер уделяет максимальному получению золота и науки от торговых маршрутов с другими игроками. |
| Trade Origin | Сколько внимания лидер уделяет развивающимся городам, из которых проходят его торговые пути. |
| Military Training | Насколько сильно лидер сосредоточен на строительстве казарм, оружейных складов и военных академий в своих городах. |
| Mobile | Насколько сильно лидер сосредоточен на производстве конных и бронированных единиц для своей армии. |
| Naval | Насколько сильно лидер сосредоточен на создании военно-морских подразделений для атаки и защиты. |
| Naval Growth | Насколько сильно лидер сосредоточен на увеличении населения своих прибрежных городов. |
| Naval Recon | Насколько сильно лидер сосредоточен на создании военно-морских подразделений для исследования карты. |
| Naval Tile Improvement | Насколько сильно лидер сосредоточен на улучшении водных плиток на своей территории. |
| Build Nuke | Насколько сильно лидер сосредоточен на завершении Манхэттенского проекта и производстве ядерного оружия. |
| Offense | Указывает, насколько сильно лидер сосредотачивается на производстве юнитов для атаки. |
| Production | Насколько сильно лидер сосредоточен на увеличении объемов производства в своих городах. |
| Long Range | Насколько сильно лидер сосредоточен на производстве юнитов дальнего боя для своей армии. |
| Recon | Насколько сильно лидер сосредоточен на производстве разведчиков для исследования карты. |
| Religion | Насколько сильно лидер сосредоточен на увеличении количества веры в своих городах и распространении религии своей империи. |
| Science | Насколько сильно лидер сосредоточен на увеличении результатов науки в его или ее городах. |
| Spaceship | Насколько сильно лидер сосредоточен на завершении программы Аполлон и производстве деталей космического корабля. |
| Tile Improvement | Насколько сильно лидер сосредотачивается на улучшении плиток земли на своей территории. |
| Use Nuke | Насколько вероятно, что лидер применит ядерное оружие. |
| Water Connection | Насколько сильно лидер сосредоточен на создании гаваней в своих прибрежных городах. |
| Wonder | Насколько сильно лидер сосредоточен на созидании чудес. |
| Afraid | Насколько вероятно, что лидер будет бояться игрока с превосходящей военной мощью, прогрессом в победе или ядерным оружием |
| Deceptive | Насколько вероятно, что лидер будет проявлять дружелюбие по отношению к другому игроку, скрывая при этом свое действительное отношение. Даже обманчивый лидер с высоким рейтингом лояльности может осудить или объявить войну союзникам, к которым он или она на самом деле не дружит. |
| Friendly | Насколько вероятно, что лидер будет дружелюбен по отношению к другому игроку. |
| Guarded | Насколько вероятно, что лидер будет защищаться от другого игрока. |
| Hostile | Насколько вероятно, что лидер будет враждебно настроен по отношению к другому игроку. |
| Neutrality | Насколько вероятно, что лидер будет нейтральным по отношению к другому игроку. |
| War | Насколько вероятно, что лидер объявит войну игроку, которого он или она считает слабым. |
| CS Bully | Насколько вероятно, что лидер потребует дань от близлежащих городов-государств. |
| CS Conquest | Насколько вероятно, что лидер нападет на близлежащие города-государства. |
| CS Friendly | Насколько вероятно, что лидер подружится с соседними городами-государствами. |
| CS Ignore | Насколько вероятно, что лидер проигнорирует близлежащие города-государства. |
| CS Protect | Насколько вероятно, что лидер пообещает свою защиту близлежащим городам-государствам. |

После просмотра диапазонов значений для всех признаков, было выяснено, что признаки Airlift, Archaeology, Land Trade, Sea Trade, Trade Destination, Trade Origin имеют во всех записях значения 5, а значит являются полностью неинформативными и бесполезными.

В Civilization 5 существует пять возможных способов победить: культурная победа, военная (территориальная) победа, научная (космический корабль) победа, дипломатическая победа и победа по очкам. На основе анализа всех признаков было отобрано три признака для возможного в качестве целевых: Culture, Diplomacy, Spaceship. Каждый из этих признаков напрямую указывает желание двигаться по соответствующему пути к победе.

## 3.3 Определения набора признаков для обучения

Далее для определения признаков, от которых сильно зависят целевые признаки были выведены карты корреляций для трех целевых признаков. Из-за большого кол-ва признаков анализировать такую мелкоячеистую стеку очень тяжело, поэтому эта сетка была отсортирована и записана в таблицу Excel. К сожалению, максимальные абсолютные значения коэффициента корреляции находились около 0,4. Таким образом были отобраны наборы признаков, абсолютный коэффициент корреляции которых был более 0,2. На рисунке 1 представлены карты корреляции для трех целевых признаков и соответствующих им наборов признаков.

|  |
| --- |
|  |
| а) Culture |
|  |
| б) Diplomacy |
|  |
| в) Spaceship |
| Рисунок 1 – Карты корреляций |

Также на рисунке 2 представлены гистограммы распределения этих же наборов признаков. В данном случае гистограммы распределения оказались не информативны так как признаки можно считать изначально стандартизированными и в них абсолютно нет выбросов.

|  |
| --- |
|  |
| а) Culture |
|  |
| б) Diplomacy |
|  |
| в) Spaceship |
| Рисунок 2 – Гистограммы распределения |
|  |

## 3.5 Подготовка целевых признаков

Поскольку подразумевается задача классификации, а датасет маленький, разумным является разделить каждый целевой признак на три диапазона значений. Также для корректной работы алгоритмов классификации, необходимо максимально сбалансировать классы в целевых признаках. На рисунке 3 представлены целевые признаки после разбиения их на классы.

|  |  |  |
| --- | --- | --- |
|  |  |  |
| а) Culture | б) Diplomacy | в) Spaceship |
| Рисунок 3 | | |

К сожалению, из-за того, что значения в признаках целочисленные полностью сбалансировать их не удается.

## 3.5 Построения диаграмм рассеяния

На рисунке 3, 4 и 5 представлены диаграммы рассеяния для трех наборов признаков. Также для каждого набора признаков приведены диаграммы после операций стандартизации и нормализации. Стоит заметить, что диаграммы для исходных данных и стандартизованных отличаются не сильно, чего и требовалось ожидать, т.к. все признаки приведены к одному числовому диапазону. В этом случае вероятнее всего следует ожидать и близкие результаты предсказаний для оригинальных и стандартизованных данных. Также стоит заметить, что на взгляд очень тяжело провести четкие границы между группами признаков.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| Рисунок 3 - Диаграмма рассеяния для набора Culture | | |  |
|  |  |  |  |
| Рисунок 4 - Диаграмма рассеяния для набора Diplomacy | | |  |
|  |  |  |  |
| Рисунок 5 - Диаграмма рассеяния для набора Spaceship | | |  |

## 3.6 Постановка задачи машинного обучения

Для выбранной базы данных задача машинного обучения следующая:

предсказание методом классификации намерений персонажа игры двигаться к определенному виду победы.

## 4. Применение алгоритмов машинного обучения

## 4.1 Выбор применяемых алгоритмов

Для нахождения наиболее подходящего алгоритма классификации для поставленной задачи обучение будет происходить на пяти алгоритмах классификации:

* метод ближайших соседей,
* наивный Байсовский метод,
* дерево решений,
* метод опорных векторов,
* случайный лес.

Обучение и предсказание также будет производиться на трех типах данных: оригинальных, стандартизованных и нормализованных.

## 4.2 Применение методов машинного обучения

Поскольку методов классификации пять, а у каждого из них есть набор настраиваемых параметров, было принято решение применить функцию полного перебора параметров *GridSearchCV*. В данном случае время подбора набора параметров, дающий наилучший результат предсказания, полным перебором не так существенно. В случае если бы алгоритмов было больше, а комбинации параметров измерялись бы сотнями тысяч то имело бы смысл применять функцию рандомизированного поиска *RandomizedSearchCV.* В результате работы функции подбора параметров были получены файлы, в которых были записаны наилучшие результаты предсказаний для всех методов классификации и типов входных данных. Данные из файлов для наилучшей читаемости были преобразованы в таблицу 2.

Таблица 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Метод классификации | Тип базы | Culture | Diplomacy | Spaceship |
| KNeighborsClassifier() | Original | 57,9% | 67,0% | 57,1% |
| Standardization | 57,9% | 67,0% | 55,1% |
| Normalization | 57,9% | 69,8% | 55,1% |
| GaussianNB() | Original | 60,7% | 65,9% | 55,1% |
| Standardization | 63,0% | 63,5% | 57,9% |
| Normalization | 55,1% | 68,6% | 52,7% |
| DecisionTreeClassifier() | Original | 57,9% | 66,6% | 61,1% |
| Standardization | 57,9% | 66,6% | 61,1% |
| Normalization | 52,0% | 63,9% | 60,7% |
| SVC() | Original | 53,6% | 62,7% | 47,6% |
| Standardization | 48,4% | 59,1% | 47,6% |
| Normalization | 57,5% | 74,2% | 53,2% |
| RandomForestClassifier() | Original | 67,8% | 72,6% | 67,4% |
| Standardization | 65,5% | 74,2% | 67,8% |
| Normalization | 62,7% | 71,8% | 63,1% |

Из таблицы видно, что в среднем наилучший результат был достигнут в методе случайного леса при стандартизованных данных. Также стоит отметить, что самый сильно-коррелирующий набор признаков был для целевого признака *Diplomacy* и для него же был получен наивысший результат верных предсказаний намерений персонажа двигаться по этому пути к победе.

В общем процент верных угадываний для всех алгоритмов и целевых признаков относительно низкий. Обусловлено это малым наборов сильно-коррелирующих признаков и малому размеру датасета.

# Заключение

В рамках работы были изучены методы анализа и предварительной обратки баз данных, предоставляемых для алгоритмов машинного обучения. Найден произвольный датасет, для которого в дальнейшем была поставлена задача машинного обучения. Изучены и применены различные методы классификации.

В результате проделанной работы была решена задача классификации для определения намерений персонажа компьютерной игры Civilization 5 двигаться по определенному пути к победе. Наивысший значение процента верного предсказания достигнуто для алгоритма случайного леса и составляет 74,2% при предварительной стандартизации исходных данных. В общем же, можно заключить, что из-за отсутствия сильно-коррелирующих признаков результаты классификации полностью не удовлетворяют поставленную задачу.