**Министерство науки и высшего образования Российской Федерации**

федеральное государственное автономное образовательное учреждение   
высшего образования

«Национальный исследовательский Томский политехнический университет»

|  |  |
| --- | --- |
| Школа / филиал | ИЯТШ ТПУ |
| Обеспечивающее подразделение | ОЭФ |
| Направление подготовки / специальность | 01.03.02  Прикладная математика и информатика |
| Образовательная программа (направленность (профиль)) | Математические и программные средства исследования операций в экономике |

**ОТЧЕТ ПО УЧЕБНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ СТУДЕНТА**

**на тему:**

Методы отбора признаков для моделей машинного обучения

Выполнил

Студент группы 0В91                            Змитрович Н.Л.  
  
  
  
Проверил  
доцент, кандидат физико-математических наук Семенов М. Е.

Томск 2021

**Содержание**

[Введение 3](#_Toc91076985)

[Цель работы 3](#_Toc91076986)

[Основные задачи 3](#_Toc91076987)

[Ключевые слова 3](#_Toc91076988)

[Ключевые слова на английском языке 4](#_Toc91076989)

[План по развитию работы 4](#_Toc91076990)

[Аналитический обзор литературных источников 4](#_Toc91076991)

[Описание методологии CRISP-DM 5](#_Toc91076992)

[Методы выявления и обработки ошибок в данных 6](#_Toc91076993)

[Методы отбора признаков для задачи много-классовой классификации 7](#_Toc91076994)

[Разработка алгоритмов, используемых при решении задачи 9](#_Toc91076995)

[Анализ данных 9](#_Toc91076996)

[Подготовка данных 10](#_Toc91076997)

[Анализ полученных результатов 11](#_Toc91076998)

[Заключение 14](#_Toc91076999)

[Список литературы 15](#_Toc91077000)

# Введение

С развитием технологий люди начали собирать и накапливаться все больше самых разных данных во всех областях, начиная от погоды в определенных регионах и заканчивая отслеживанием активности пользователей какого-либо приложения. Такая информация может помочь решить многие задачи, для отслеживания текущего состояния некоторой системы, её анализа и составления будущего прогноза. Для решения таких задач обычно используют методы статистики или машинного обучения. Такие методы имеют множество различных инструментов для решения конкретных задач, например, классификации или регрессии.

Но для получения результата необходимо, чтобы данные были достаточно репрезентативными, а для этого они требуют хорошей обработки и подготовки, для дальнейшего использования их в моделях машинного обучения. Потому что в данных, например, могут быть аномальные или пропущенные значения, выбросы или в принципе в данных будут какие-то лишние параметры, которые только усложнят построение модели. Также данные чаще всего не представлять собой набор готовых параметров для создания модели, а являются скорее набором таблиц, из которых эти нужно вычленить такие параметры, которые будут достаточно хорошо характеризовать систему.

Для решения таких проблем с обработкой данных существуют различные методы отбора признаков для моделей машинного обучения.

# Цель работы

Программно реализовать ключевые методы отбора признаков для обучения модели машинного обучения.

# Основные задачи

1. Изучить методологию по исследованию данных CRISP-DM
2. Предложить методы выявления и обработки ошибок в данных
3. Привести критерии для отбора ключевых методов отбора признаков для задачи много-классовой классификации.

# Ключевые слова

Датасет, отбор признаков, обучение модели, обработка данных, машинное обучение

# Ключевые слова на английском языке

Dataset, feature selection, model training, data processing, machine learning.

# План по развитию работы

Для достижения поставленной цели необходимо ознакомиться с литературой. Изучить методы выбора признаков, способов их предварительной обработки, научиться решать проблему пропущенных значений, корреляции признаков. Научиться определять признаки с низкой и нулевой важностью. Ознакомиться с методами фильтрации признаков, нормализации данных. Узнать для чего необходима регуляризация данных, нормализация и научиться применять данные методы для выделения наиболее важных признаков в датасете.

После изучения литературы необходимо будет провести отбор признаков для обучения модели машинного обучения для конкретного датасета.

Среда выполнения работы GitHub, Python.

Исходные данные Kaggle: https://www.kaggle.com/uciml/iris

# Аналитический обзор литературных источников

Основной задачей машинного обучения, обычно является построение модели. Такие модели обычно представляют собой некоторое отображение, которое сопоставляет входному набору данных, результирующий набор данных. Модель может получать на вход таблицу из кейсов, при этом в каждом кейсе может быть несколько значений (атрибутов). Такими кейсами могут быть, например, данные о клиентах компании и соответственно атрибутами этих кейсов будут некоторые характеристики клиентов: возраст, средний чек, информация о покупках и так далее. На основе таких данных модель делает свой прогноз, например, совершит ли клиент ещё одну покупку, если поставлена задача классификации.

Так как построение модели машинного - достаточно комплексная задача, которая предусматривает множество этапов, начиная от постановки задачи, создания обучающей выборки и заканчивая выбором метрики, которую необходимо оптимизировать, для получения максимальной практической выгоды и организации работы следует придерживаться определенного плана – методологии.

Существует множество различных методологий анализа данных. Самой популярной и часто используемой на данный момент из опроса на аналитическом портале kdnuggets.org [1] является методология CRISP-DM. Эта методология в первую очередь направленна на решение аналитических задач для бизнеса.

## Описание методологии CRISP-DM

Методология CRISP-DM выделяет шесть основных этапов при решении задач машинного обучения: бизнес аналитика (business understanding), анализ данных (data understanding), подготовка данных (data preparation), создание модели (modeling), оценка результатов (evaluation), внедрение (deployment) [2,3].

Первый этап – бизнес аналитика направлен на постановку задачи и выбора пути дальнейшего решения задачи. Сначала требуется определить то, зачем вообще нужна будет модель машинного обучения, какие результаты мы хотим получить и существуют ли уже готовые решения. После чего необходимо оценить текущие возможности, например, наличие необходимых вычислительных ресурсов, а также нужно оценить риски, например, возможные проблемы со сбором данных или необходимость дополнительных финансовых вложений. Далее нужно определить то, как будем оценивать результат. Для этого нужно выбрать метрику, которая будет характеризовать качество полученных данных, а также нужно определить некоторый порог, превзойдя который модель будет считаться успешной. И последним пунктом первого этапа является составление подробного плана проекта.

Следующим этапом является анализ данных. В первую очередь данные нужно собрать и объединить в общую таблицу. Если собственных данных слишком мало, их можно найти из прочих источников, или попробовать собрать. После сбора данные нужно описать. Найти ключевые параметры для всей таблицы и каждого атрибута в отдельности. На этом этапе можно начинать исследование данных, нужно построить графики и составить первичное представление о том какие атрибуты могут быть полезны для решения задачи. И на последнем этапе анализа данных нужно проверить целостность и качество исходной таблице. Например, проверить наличие пропущенных значений и опечаток. Также данные могли быть собраны из разных источников поэтому кодировка атрибутов может различаться, то есть в разных кейсах один атрибут может быть записан в различных единицах измерения или с помощью различных наборов символов.

На этапе подготовки данных сначала проводится отбор. То есть нужно определить какие данные не стоит включать в обучающую выборку. Такой отбор нужно проводить, учитывая специфику используемой модели, например, если будет использоваться логистическая регрессия, значит параметры, которые коррелируют с другими параметрами стоит исключить. Также если в атрибутах или конкретных кейсах много пропущенных значений, их тоже стоит исключить. После чего на основе отбора проводится отчистка данных. Следующим пунктом является генерация новых данных. Для улучшения качества модели можно сгенерировать новые данные. Например, применить агрегацию к атрибутам, найти их отношения или средние значения. Так если в данных есть атрибут – «траты за все время» и атрибут – «количество всех покупок», то можно создать новый атрибут – «средний чек». Также на этом этапе производится нормализация данных и заполнение пропущенных значений при необходимости. Завершающими этапами являются – интеграция данных, то есть объединение данных из разных источников и форматирование данных, то есть приведение таблицы к формату, который будет необходим модели для обучения.

Четвертый этап – моделирование. На этом этапе сначала определяется алгоритм и архитектура модели, обычно выбор строится на основе специфики входных данных и необходимого результата. После данные разбиваются на тренировочные, тестовые и валидационные. Далее происходит обучение модели на тестовых данных и подбор оптимальных гиперпараметров на валидации. После чего считаются итоговые метрики и оцениваются результаты работы модели на тестовых данных.

На следующем этапе происходит оценка модели и её результатов. Оценивается то, на сколько качественной получилась модель и на сколько полученные результаты будут полезны для решения поставленной бизнес задачи. После делается окончательный вывод о целесообразности использования построенной модели. Если модель устраивает заказчика, то стоит переходить к следующему этапу, а если модель не смогла дать нужный результат, стоит перейти к предыдущим этапам и попробовать построить другую модель.

На последнем этапе происходит внедрение модели. Создается подробный план развертывания (применения) модели. Если решение требует поддержки в будущем, то производится настройка мониторинга модели. И последним этапом является написание отчета по результатам моделирования.

## Методы выявления и обработки ошибок в данных

Обучающие данные являются основой для построения модели, и какой бы сложной не была модель, если данные на которых она обучалась были достаточно «плохими», она не сможет показать хороший результат.

Для того, чтобы понять какие данные являются «плохими», необходимо определить какими качествами должны обладать чистые данные. Обычно данные характеризую по девяти пунктам [4]:

1. Доступность – возможность использовать данные в свободном доступе и подходящем формате.
2. Точность – то на сколько данные правильно описывают истинные значения
3. Взаимосвязанность – если не будет связи между данными и целевой переменной, модель всегда будет давать случайных результат.
4. Полнота – отсутствие пропусков или опечаток
5. Непротиворечивость – во всех источниках одинаковые атрибуты одних и тех-же кейсов должны быть равными.
6. Однозначность – каждое поле должно быть определенно однозначно
7. Релевантность – данные должны зависеть от характера анализа
8. Надежность – совокупность точности и полноты
9. Своевременность – данные должны быть актуальны для проводимого исследования

На практике при наличии проблем в данных, аналитик может решить лишь часть из них. Так, например, повысить точность или решить проблему своевременности обычно не представляется возможным. Но при этом некоторые проблемы решить возможно.

Для этого при анализе данных всегда нужно строить распределения, для обнаружения аномальных значений. Например, при наличии атрибута возраста, его значение не может быть меньше нуля. Также стоит проверять наличие пустых или пропущенных значений, иногда ячейки могут быть просто пустыми, а иногда заполненными аномальными значениями.

Если данные были получены и с конкатенированы из разных источников, то может возникнуть проблема с однозначностью, например, один атрибут может быть записан в разных единицах измерениях, либо набор символов, которыми обозначались значения будет различным в разных таблицах. Такие проблемы тоже стоит учитывать и проверять данные на однозначность.

Также, когда есть несколько таблиц со смежными значениями, например, имеются различные таблицы с данными о пользователях с некоторыми одинаковыми атрибутами, значит и значения в разных таблицах у одних и тех же пользователей не должны различаться и противоречить.

Релевантность также можно улучшить, исключая из выборки значения, не влияющие на целевую переменную. Исключать атрибуты можно эмпирически, например, уникальный индекс пользователя может ухудшить модель прогнозирующую активность этого пользователя. Также можно проверять корреляцию между параметрами и целевой переменной или выдвигать гипотезы о зависимости или независимости.

## Методы отбора признаков для задачи много-классовой классификации

В задача много-классовой классификации данные являются размеченными, то есть в явном виде присутствует целевая переменная характеризующая класс объекта, то есть обучение такой модели будет с «учителем». Большинство алгоритмов обучения модели машинного обучения с «учителем» могут самостоятельно оценивать вклад атрибутов в формировании целевой переменной, например, в нейронной сети веса малозначимых параметров могут стать нулевыми, а решающее дерево, которое чаще всего используется в задачах классификации при небольшой глубине может вообще не учесть параметры плохо разделяющие данные на классы. Но несмотря на это исключив из выборки не релевантные параметры, мы можем значительно ускорить и улучшить процесс обучения, так-как такие параметры являются шумом и могут сильно повлиять на переобучение модели. В основном методы отбора атрибутов различают по критерию качества пространства признаков, который использует метод [5].

Одними из таких методов отбора являются методы обертки. В качестве оценки качества признаков, такие методы используют предсказательную модель. То есть чем лучше получилась итоговая метрика на валидационных данных, тем лучше набор исходных признаков. Методы обертки обычно делятся на два типа:

* Прямой отбор – начинаем строить модели с наименьшим числом атрибутов, после чего добавляем новые и сравниваем результат
* Обратный отбор – начинаем строить модели сначала со всеми атрибутами, после чего исключаем те, что внесли наименьший вклад и сравниваем результат

Порядок добавления или исключения атрибутов зависит от вклада конкретного атрибута в результирующую переменную, этот вклад можно определить с помощью корреляции между атрибутом и целевой переменной. Главной проблемой таких методов заключается в большой вычислительной нагрузке, так как для каждого набора признаков нужно заново создавать и обучать модель. Также методы обертки отбирают оптимальный набор признаков для конкретного классификатора и соответственно при использовании другого классификатора или другой архитектуры модели, выбранный набор признаков не будет оптимальным.

Другим классом методов отбора признаков являются фильтры. Такие методы отбирают универсальный для всех классификаторов набор признаков. Обычно в качестве критериев качества такие методы используют информационные критерии, например, взаимную корреляцию или энтропию. Так как, показатель энтропии показывает степень неопределенности данных (формула 1), значит если атрибут достаточно информативный, то в итоге после разделения на классы на основе этого атрибута энтропия в данных должна уменьшиться. Эта разница в энтропии называется Information gain, и данный метод оценки информативности атрибутов используется в моделях решающих деревьев. Заметим, что если разделять выборку по атрибуту, то в итоге мы получим несколько подмножеств, поэтому нужно найти среднюю энтропию этих подмножеств по формуле 2, где – подмножество после разделения, – число элементов в этом подмножестве и – число элементов во всем множестве, а k – количество полученных подмножеств.

 (1)

 (2)

 (3)

Также если стоит задача много классовой классификации, то можно использовать элементы дискриминантного анализа. Для этого находятся матрицы корреляций между классами и внутри классов. Набор параметров в этом случае будет лучше тогда, когда рассеяние между классами будет больше, а рассеяние внутри классов будет меньше. Для определения качества набора атрибутов можно использовать критерии дискриминантного анализа, например, отношение суммы дисперсий между классами и суммы дисперсий внутри классов.

# Разработка алгоритмов, используемых при решении задачи

## Анализ данных

Для того, чтобы разработать алгоритм отбора параметров для начала нужно ознакомиться с данными. В источнике данные представлены в виде таблицы iris.csv. Таблица имеет пять атрибутов, одним из которых является целевая переменная Species.

Заметим, что целевая переменная имеет строковые значения, значит сначала их нужно перевести в числа. Для этого можно использовать метод из библиотеки Pandas – unique(), который вернет все уникальные значения, то есть группы. После определим для каждой группы свое уникальное числовое значение и произведем замену.

Построим графики распределений всех атрибутов и попробуем найти аномальные значения, исследуем данные на полноту, заполним пропущенные значения и избавимся от выбросов при их наличии. Для этого могут пригодиться такие методы как sum(), null(), mean(), meadian(). При нахождении выбросов будет необходимо найти межквартильное расстояние, для этого используем метод numpy – quantile().

Чтобы определить является ли кейс выбросом, нужно найти первый и третий квартиль каждого атрибута по всей выборке, после найти межквартильное расстояние a. Отложим отрезки равные 1.5\*a от первого и третьего квартиля (неравенство 5). Все значения, которые не попадут в такой интервал будут считаться выбросами. Для того, чтобы отследить наличие выбросов в выборке можно воспользоваться графиком box-plot из библиотеки matplotlib.

 (4)

 (5)

## Подготовка данных

Произведем отбор параметров для дальнейшего обучения модели. Воспользуемся фильтрами и найдем корреляционную таблицу, проверим какие атрибуты имеют наибольшую корреляцию с целевой переменной, а также определим есть ли сильно коррелирующие между собой атрибуты.

Так-как целевая переменная является номинальным признаком (классом), значит можно найти энтропию каждого атрибута и сделать соответствующие выводы об информативности атрибутов. Для нахождения information gain можно воспользоваться библиотекой sklearn. Построим лес решающих деревьев, обучим на данных и вернем значение feature\_importances\_, для того, чтобы определить какие из атрибутов внесли наибольший вклад.

После чего можно сделать первые представления о том какие из параметров вносят больший вклад в целевую переменную и начать отбирать набор параметров. Для оценки качества пространства признаков, воспользуемся элементами дискриминантного анализа и найти рассеяние между классами и внутри классов, после чего вычислим критерий по формуле 6, где – матрица рассеяния между классами, а – среднее матриц рассеяния по каждому классу. Сначала найдем метрику для всей совокупности атрибутов, после чего будем исключать наименее значимые, до тех пор, пока критерий не будет наибольшим.

 (6)



**



**

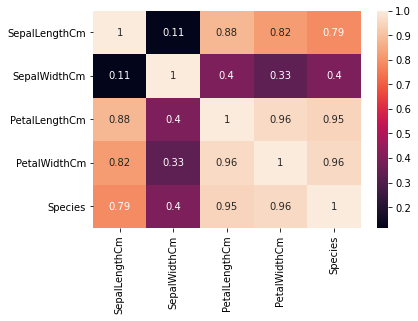
**Практическая часть**

Ссылка на github: <https://github.com/NLZer0/YIRS>

# Анализ полученных результатов

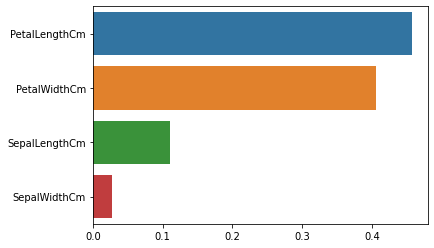
При обработки данных не нашел пропущенных значений, также заметил, что таблица отсортирована по целевой переменной. После исследования данных на выбросы, обнаружил, что параметр SepalWidthCm имеет несколько таких значений. Для того, чтобы избавиться от выбросов, исключил из выборки строки в которых значение параметра SepalWidthCm является выбросом.

Исследовал корреляцию между всеми параметрами и отобразил результат в виде графика heatmap (рис.1). Целевая переменная имеет наибольшую корреляцию с параметрами PetalWidthCm и PetalLengthCm. Но при этом оба этих параметра также сильно коррелируют друг с другом



(рис.1 График корреляционной матрицы атрибутов)

Обучил лес решений с максимальной глубиной дерева и отобразил на графике полученные Information Gain (рис.2). Совместный вклад параметров PetalWidthCm и PetalLengthCm превысил 86%, значит оба данных параметра достаточно значимы для классификации. Стоит заметить, что при построении случайного леса IG может изменять из-за наличия некоторой случайности при построении модели. Но при достаточно большом количестве решающих деревьев случайность сводится к минимуму и график информационного вклада атрибутов меняется незначительно.



(рис.2 График вклада атрибутов в суммарный IG)

В результате можем сделать вывод о том, что параметры PetalLengthCm и PetalWidthCm являются более значимыми, а чем SepalWidthCm – наименее значимый. С помощью статистики 6 буду оценивать и сравнивать пространства признаков. Для этого сначала найду статистику для исходного признакового пространства, после чего исключу из него признак SepalWidthCm и SepalWidthCm.

Оценка качества пространства из всех признаков

Метрика равна: 7.86891403142311

Оценка качества пространства из признаков SepalLengthCm, PetalLengthCm, PetalWidthCm

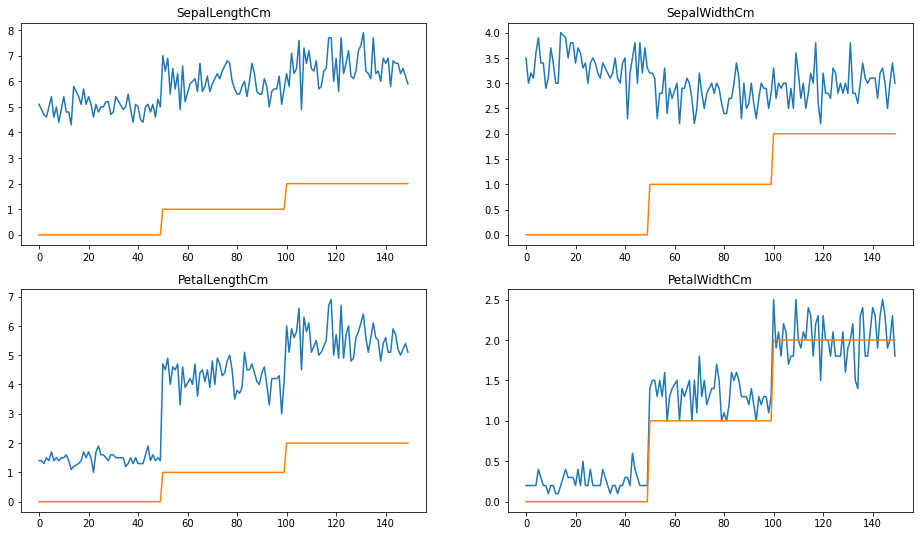
Метрика равна: 9.169016365770007

Оценка качества пространства из признаков PetalLengthCm, PetalWidthCm

Метрика равна: 16.53953669585818

Исключив из пространства признаков атрибуты SepalWidthCm и SepalLengthCm, удалось сильно увеличить отношения дисперсии между классов и внутри классов, значит и классифицировать объекты в таком пространстве будет проще.

Так-как исходная таблица отсортирована по классам объектов, построил совместные графики каждого атрибута и целевой переменной (рис.3). Таким образом можно графически заметить корреляцию между значением атрибута и классом объекта.



(рис.3 Совместные графики значений параметров и целевой переменной)

По графикам можем сделать аналогичные выводы о значимости параметров PetalLengthCm, PetalWidthCm.

Исходя их всех полученных результатов, можно сделать вывод о том, что признаковое пространство, состоящее из параметров PetalLengthCm, PetalWidthCm – является оптимальным. При этом, учитывая, что эти два параметра имеют большую взаимную корреляцию, можно сделать вывод, о том, что даже используя хотя бы один из двух параметров для классификации объектов, можно получить достаточно высокую точность.

# Заключение

В процессе работы над проектом изучил методику по исследованию данных CRISP-DM. Изучил и предложил различные методы для выявления и обработки ошибок в данных. Для исследования использовал известный набор данных - Ирисы Фишера. В рамках проекта из приведенных характеристик качества данных возможно было исследовать только полноту и однозначность. После исследования и проверки данных, избавился от выбросов.

Привел некоторые ключевые методы для отбора признаков в задаче много-классовой классификации. Для исследования значимости атрибутов использовал корреляционный анализ. Также нашел информационный вклад каждого из атрибутов с помощью модели random forest. Так-как в данных небольшое число параметров, а сама таблица отсортирована по классам, построил совместные графики значения атрибутов и целевой переменной. После чего сделал предположение о значимости параметров и проверил качество полученных признаковых пространств методами дискриминантного анализа.

# Список литературы

1. CRISP-DM, still the top methodology for analytics, data mining, or data science projects // Kdnuggets URL: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html (дата обращения: 13.10.2021).
2. CRISP-DM: проверенная методология для Data Scientist-ов // Habr URL: https://habr.com/ru/company/lanit/blog/328858/ (дата обращения: 13.10.2021).
3. IBM SPSS Modeler CRISP-DM Guide // IBM URL: ftp://ftp.software.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/CRISP\_DM.pdf (дата обращения: 13.10.2021).
4. Карл Андерсон Аналитическая культура. От сбора данных до бизнес-результатов. - Москва: Манн, Иванов и Фербер, 2017. - 420 с.
5. Обзор методов отбора признаков // Habr URL: https://habr.com/ru/company/jetinfosystems/blog/470622/ (дата обращения: 14.10.2021).