**МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ**

**Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский Нижегородский государственный университет им. Н.И. Лобачевского»**

Радиофизический факультет

Кафедра «Безопасности информационных систем»

**Отчет по учебной практике**

Выполнил: студент группы 0419ДСИ8Г

Михеев М.А

Проверил: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Ф.И.О.,

Нижний Новгород

2022

***Цель:***

Учебная практика проходила на базе ННГУ, кафедра «Безопасность информационных систем» с 02.02.2023 по 31.05.2023 г. В ходе практики для достижения поставленной цели решались следующие задачи:

1. Обзор группы методов анализа сетевого трафика ;

3. Разбор метода: деревья принятия решений.

Прохождение практики состояло из нескольких этапов. Первый организационный этап заключался в ознакомлении с программой практики, постановке цели и задач практики, составление плана учебной практики, получение заданий от руководителя.

Следующий — подготовительный этап — подразумевал подбор материала по данной теме .

Контрольный этап заключался в составлении отчета, оформление рабочего графика оформления практики, совместном обсуждении с руководителем практики результатов работы и в подготовке отчёта по учебной практике.

**Содержание**

* Введение
* Использование машинного обучения для классификации трафика
* Методы машинного обучения, используемые, для классификации трафика
* Деревья принятия решений
  + Структура дерева принятия решения
  + Задачи, которые решает дерево принятия решения
  + Процесс построения дерева принятия решения
  + Критерий остановки алгоритма
  + Преимущества и недостатки
* Деревья принятия решений: на практике
* Выводы

**Введение**

Задачу идентификации характеристик системы можно рассматривать как сопряженную по отношению к задаче управления системой. Нельзя управлять системой, если она не идентифицирована либо заранее, либо в процессе управления. Например, мы не можем управлять самолетом, пока не познакомились со всеми его управляющими устройствами. Процесс ознакомления с самолетом – это и есть процесс идентификации.

Задача классификации сетевого трафика может быть сформулирована как получение некоторых характеристик сетевого трафика с определением класса, к которому данный вид трафика относится. Классификация трафика важна для мониторинга безопасности, повышения качества обслуживания и тд.

Анализ сети – это процесс захвата трафика и его просмотра для определения наличия в нем проблем. Анализ сетевого трафика приобретает все большую актуальность в связи с развитием сетевых технологий(соответственно растет объем данных, передаваемых по сети).

**Использование машинного обучения для классификации трафика**

**Машинное обучение (Machine Learning)** — это раздел искусственного интеллекта, изучающий математические, статистические и вычислительные методы для разработки алгоритмов, способных обучаться и решать задачи не прямым способом, а на основе поиска закономерностей в разнообразных входных данных. Решение вычисляется не по точно заданной формуле, а по установленной зависимости результатов от конкретного набора признаков и их значений.

**На практике** машинное обучение широко применяется в широком спектре исследовательских и прикладных задач:

* прогнозирование событий и ситуационный анализ, например, отток клиентов в ритейле или заблаговременное предсказание поломок промышленного оборудования;
* распознавание образов (изображений, лиц, голоса и т.д.);
* классификация образцов, в частности, рентгеновских снимков для постановки диагноза;
* выявление мошеннических операций (антифрод-системы в банках и cybersecurity)

Рост скорости передачи данных, распространение шифрования, тенденции изменения сетевого трафика, рост необходимости увеличения обработки информации – все это потребовало появление новых способов классификации. Методы машинного обучения позволяют во многом упростить работу с созданием наборов различающих характеристик классов(собрать проще, чем проанализировать вручную).

**Методы машинного обучения, используемые, для классификации трафика**

Среди них можно выделить:

1. Наивный байесовский классификатор;
2. Метод опорных векторов;
3. Метод k-ближайших соседей;
4. Деревья принятия решений( с разными алгоритмами построения дерева: CART,C4.5,C5.0);
5. Методы бэггинга(случайный лес);
6. Методы бустинга(Adaboost,XGBoost);
7. Нейронные сети;

Далее будем рассматривать подробно метод «Деревья принятия решений( с разными алгоритмами построения дерева: CART,C4.5,C5.0)».

**Деревья принятия решений**

Дерево решений — метод автоматического анализа больших массивов данных. Дерево решений — эффективный инструмент интеллектуального анализа данных и предсказательной аналитики. Он помогает в решении задач по классификации и регрессии.

Правила генерируются за счет обобщения множества отдельных наблюдений, описывающих предметную область. Поэтому их называют индуктивными правилами, а сам процесс обучения — индукцией деревьев решений.

Структура дерева представляет собой «листья» и «ветки». На рёбрах («ветках») дерева решения записаны признаки, от которых зависит целевая функция, в «листьях» записаны значения целевой функции, а в остальных узлах — признаки, по которым различаются случаи. Чтобы классифицировать новый случай, надо спуститься по дереву до листа и выдать соответствующее значение.

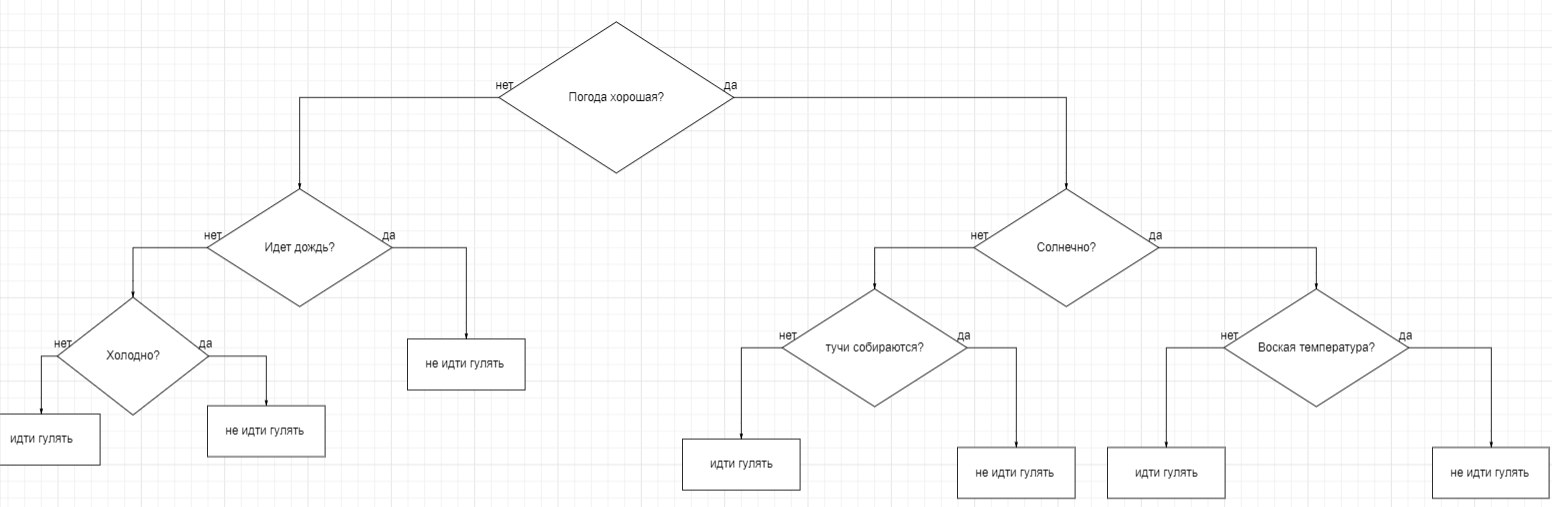
**Структура дерева принятия решений**

Дерево решений — метод представления решающих правил в определенной иерархии, включающей в себя элементы двух типов — узлов (node) и листьев (leaf). Узлы включают в себя решающие правила и производят проверку примеров на соответствие выбранного атрибута обучающего множества.

Простой случай: примеры попадают в узел, проходят проверку и разбиваются на два подмножества:

* первое — те, которые удовлетворяют установленное правило;
* второе — те, которые не удовлетворяют установленное правило.

Простой пример из жизни:



К каждому подмножеству снова применяется правило, процедура повторяется. Это продолжается, пока не будет достигнуто условие остановки алгоритма. Последний узел, когда не осуществляется проверка и разбиение, становится листом.

Лист определяет решение для каждого попавшего в него примера. Для дерева классификации — это класс, ассоциируемый с узлом, а для дерева регрессии — соответствующий листу модальный интервал целевой переменной. В листе содержится не правило, а подмножество объектов, удовлетворяющих всем правилам ветви, которая заканчивается этим листом.

Пример попадает в лист, если соответствует всем правилам на пути к нему. К каждому листу есть только один путь. Таким образом, пример может попасть только в один лист, что обеспечивает единственность решения.

**Задачи, которые решает дерево принятия решения**

Его применяют для поддержки процессов принятия управленческих решений, используемых в статистике, анализе данных и машинном обучении. Этот инструмент помогает решать следующие задачи:

**1)Классификация.** Отнесение объектов к одному из заранее известных классов. Целевая переменная должна иметь дискретные задачи.

**2)Регрессия (численное предсказание).** Предсказание числового значения независимой переменной для заданного входного вектора.

**3)Описание объектов.** Набор правил в дереве решений позволяет компактно описывать объекты. Поэтому вместо сложных структур, используемых для описания объектов, можно хранить деревья решений.

**Процесс построения дерева принятия решения**

Построение осуществляется в 4 этапа:

1. Выбрать атрибут для осуществления разбиения в данном узле.
2. Определить критерий остановки обучения.
3. Выбрать метод отсечения ветвей.
4. Оценить точность построенного дерева.

Основная задача при построении дерева решений — это последовательно и рекурсивно разбить обучающее множество на подмножества с применением решающих правил в узлах. Как долго надо разбивать? Этот процесс продолжается до того момента, пока все узлы в конце ветвей не станут листами.

Узел становится листом в двух случаях:

1)естественным образом — когда он содержит единственный объект или объект только одного класса;

2)после достижения заданного условия остановки алгоритм — например, минимально допустимое число примеров в узле или максимальная глубина дерева.

На сегодняшний день существует много алгоритмов обучения: ID3, CART, C4.5, C5.0, NewId, ITrule, CHAID, CN2 и другие. Наиболее популярными считаются:

**1)ID3 (Iterative Dichotomizer 3).** Алгоритм позволяет работать только с дискретной целевой переменной. Деревья решений, построенные на основе ID3, получаются квалифицирующими. Число потомков в узле неограниченно. Алгоритм не работает с пропущенными данными.

**2)C4.5.**«Продвинутая» версия ID3, дополненная возможностью работы с пропущенными значениями атрибутов. В 2008 году издание Spring Science провело исследование и выявило, что C4.5 — самый популярный алгоритм Data Mining.

**3)CART (Classification and Regression Tree).** Алгоритм решает задачи классификации и регрессии, так как позволяет использовать дискретную и непрерывную целевые переменные. CART строит деревья, в каждом узле которых только два потомка.

**Критерий остановки алгоритма**

Каков же критерий остановки алгоритма? Алгоритм обучения может работать до получения «чистых» подмножеств с примерами одного класса. В таком случае высока вероятность получить дерево, в котором для каждого примера будет создан отдельный лист. Такое дерево не получится применять на практике из-за «переобученности». Каждому примеру будет соответствовать свой уникальный путь в дереве. Получится набор правил, актуальный только для данного примера.

Переобучение в случае дерева решений имеет схожие с нейронными сетями последствия. Оно будет точно распознавать примеры из обучения, но не сможет работать с новыми данными. Еще один минус — структура переобученного дерева сложна и плохо поддается интерпретации. Поэтому специалисты решили останавливать алгоритм принудительно.

Для принудительной остановки используют несколько методов:

**1)Ранняя остановка.** Алгоритм останавливается после достижения заданного значения критерия (например, процентной доли правильно распознанных примеров). Преимущество метода — сокращение временных затрат на обучение. Главный недостаток — ранняя остановка негативно сказывается на точности дерева. Из-за этого многие специалисты советуют отдавать предпочтение отсечению ветвей.

**2)**О**граничение глубины дерева.**Алгоритм останавливается после достижения установленного числа разбиений в ветвях. Этот подход также негативно сказывается на точности дерева.

**3)**З**адание минимально допустимого числа примеров в узле.** Устанавливается ограничение на создание узлов с числом примером меньше заданного (например, 5). В таком случае не будут создаваться тривиальные разбиения и малозначимые правила.

**Преимущества и недостатки**

У метода «Дерево принятия решений» есть свои преимущества и ,конечно же, недостатки.

**Преимущества:**

1) Формируют четкие и понятные правила классификации. Деревья решений хорошо и быстро интерпретируются.

2)Способны генерировать правила в областях, где специалисту трудно формализовать свои знания.

3) Быстро обучаются и прогнозируют.

4)Легко визуализируются, то есть могут «интерпретироваться» не только как модель в целом, но и как прогноз для отдельного тестового субъекта (путь в дереве).

5) Не требуется много параметров модели.

**Недостатки:**

1) Деревья решений чувствительны к шумам во входных данных. Небольшие изменения обучающей выборки могут привести к глобальным корректировкам модели, что скажется на смене правил классификации и интерпретируемости модели.

2)Дерево решений делает константный прогноз для объектов, находящихся в признаковом пространстве вне параллелепипеда, который охватывает не все объекты обучающей выборки.

3) Возможно переобучение дерева решений, из-за чего приходится прибегать к методу «отсечения ветвей», установке минимального числа элементов в листьях дерева или максимальной глубины дерева.

**Деревья принятия решений: на практике**

Самым простым и доступным способом анализа сетевого трафика являются программы анализаторы. Программа-анализатор– это программа или программно-аппаратное устройство, которое применяется для захвата и последующего анализа захваченного трафика или отдельного сегмента сети. В процессе захватывания всех потоков, анализатор захватывает и записывает все пакеты, полученные из интернет-трафика. В случае подробного и информативного анализа происходит декодирование пакетов из зашифрованной формы представления, в читаемую.

Существует множество программ для реализации данного функционала. Одной из самых известных, гибких и удобных является Wireshark.