МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ

РОССИЙСКОЙ ФЕДЕРАЦИИ

ФГАОУ ВО «СЕВЕРО–КАВКАЗСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

ИНСТИТУТ ЦИФРОВОГО РАЗВИТИЯ

МЕЖИНСТИТУТСКАЯ БАЗОВАЯ КАФЕДРА

**КУРСОВОЙ ПРОЕКТ**

по дисциплине «Междисциплинарный проектный практикум»

на тему:

«Разработка приложения, реализующего минимизацию энтропии»

**Выполнил:**

Турклиев Владимир Назирович

студент 2 курса

группы ПИЖ–б–о–21–1

направление подготовки

09.03.04 Программная инженерия

направленность (профиль)

«Разработка и сопровождение

программного обеспечения»

очной формы обучения

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

(подпись)

**Руководитель проекта:**

Свистунов И.В., доцент

межинститутской базовой кафедры

Проект допущен к защите \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(подпись руководителя) (дата)

Проект выполнен и

защищен с оценкой \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Дата защиты \_\_\_\_\_\_\_\_\_\_\_\_\_

|  |  |  |
| --- | --- | --- |
| Члены комиссии: |  |  |
| доцент кафедры инфокоммуникаций канд. физ.-мат. наук |  | Братченко Н.Ю. |
| доцент кафедры инфокоммуникаций канд. техн. наук | (подпись) | Воронкин Р.А. |
| доцент кафедры инфокоммуникаций | (подпись) | Свистунов И.В. |

(подпись)

Ставрополь, 2023 г.

|  |  |
| --- | --- |
|  | УТВЕРЖДАЮ  Зав. кафедрой инфокоммуникаций  докт. техн. наук, профессор  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Г.И. Линец  «\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_2023 г. |

|  |  |
| --- | --- |
| Институт | цифрового развития |
| Кафедра | инфокоммуникаций |
| Направление подготовки | 09.03.04 Программная инженерия |
| Профиль подготовки | Разработка и сопровождение программного обеспечения |

**ЗАДАНИЕ**

**на курсовой проект**

|  |  |
| --- | --- |
| студента | Турклиева Владимира Назировича |
| по дисциплине | Междисциплинарный проектный практикум |
| **1. Тема проекта:** разработка приложения, реализующего метод минимизации энтропии путем отбора информативных признаков. | |
| **2. Цель:** Разработка приложения, реализующего минимизацию энтропии путем отбора информативных признаков. | |
| **3. Задачи:**  1) Провести теоретическое описание и анализ метода минимизации энтропии.  2) Программная реализация отбора наиболее информативных признаков. | |
| **4. Перечень подлежащих разработке вопросов:**  **а) по теоретической части:**  1 Корреляционный анализ  1.1 Основные сведения  1.2 Что такое энтропия  1.3 Как работает метод  1.4 Какими бывают методы минимизации энтропии  1.4.1 Перекрестная энтропия  1.4.2 Информационная энтропия  1.4.3 Метод максимальной энтропии  1.4.4 Метод максимума правдоподобия  1.4.5 Метод условной энтропии  1.4.6 Метод отбора признаков  2. Разработка программного продукта  2.1 Требования к проекту  2.2 Разработка проекта  2.3 Подход к разработке  2.3.1 Выбор языка программирования  2.4 Области применения  **б) по практической части:**  3 Разработка программного приложения  3.1 Результаты проводимых расчетов и анализ  3.2 Описание функций программы | |
| **5. Исходные данные:**  а) по литературным источникам:  - использовать приведенное в рекомендуемых источниках описание различных алгоритмов поиска в неупорядоченных последовательностях.  - использовать известные в науке результаты исследований для сравнительной оценки полученных результатов и подтверждения адекватности реализуемого алгоритма.  б) по вариантам, разработанным преподавателем:  - самостоятельно с учетом степени подготовленности выбрать язык программирования для реализации алгоритма;  - использовать известные из учебного курса методические рекомендации для исследования алгоритмов поиска.  **6. Список рекомендуемой литературы:**   1. Шелонцев В.А., Шелонцева Л.Н. Непараметрические методы статистики: Учебное пособие. – 3-е изд. – Омск: Полиграфический центр КАН, 2016. – 60 с. 2. М. Холлендер, Д.А. Вулф, пер. с англ. Д. С. Шмерлинга. Непараметрические методы статистики. – Москва: Финансы и статистика, 1983. – 518с. 3. Горяинова Е.Р., Панков А.Р., Платонов Е.Н. Прикладные методы анализа статистических данных. - М.: Издательский дом Высшей школы экономики, 2012. - 310 с. 4. Лаврищева Е. М., Петрухин В.А. Методы и средства инженерии программного обеспечения: Учебное пособие. – Москва: Московский физико-технический институт (государственный университет), 2007. – 415 с. 5. Гайдышев И. Анализ и обработка данных. – Санкт-Петербург: Питер, 2001. – 752 с. 6. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика, Основы моделирования и первичная обработка данных. – Москва: Финансы и статистика, 1983. – 471 с. 7. tkinter — Python interface to Tcl/Tk // Python Documentation URL: https://docs.python.org/3/library/tkinter.html (дата обращения: 20.03.2023). 8. io — Core tools for working with streams // Python Documentation URL: https://docs.python.org/3/library/io.html (дата обращения: 20.03.2023). 9. pandas documentation // Pandas URL: https://pandas.pydata.org/docs/ (дата обращения: 20.03.2023). | |

**7. Контрольные сроки представления отдельных разделов курсового проекта:**

25 % - Первый раздел « » \_\_\_\_\_\_\_\_\_\_\_\_ 2023 г.

50 % - Второй раздел « » \_\_\_\_\_\_\_\_\_\_\_\_ 2023 г.

75 % - Третий раздел « » \_\_\_\_\_\_\_\_\_\_\_\_ 2023 г.

100 % - Оформленная работа, прошедшая нормоконтроль « » \_\_\_\_\_\_\_\_\_2023 г.

**8. Срок защиты студентом курсового проекта** « » 2023г.

|  |  |  |
| --- | --- | --- |
| Руководитель курсового проекта |  |  |
| доцент межинститутской базовой кафедры |  |  |
|  |  |  |
| Задание принял(а) к исполнению студент(ка) очной формы обучения | | |
| 2 курса ПИЖ-б-о-21-1 группы \_\_\_\_\_\_\_\_\_\_\_\_ В.Н.Турклиев.  СОДЕРЖАНИЕ  [ВВЕДЕНИЕ 5](#_Toc136004568)  [1. ТЕОРЕТИЧЕСКАЯ ЧАСТЬ 6](#_Toc136004569)  [1.1 Основные сведения о методе 6](#_Toc136004570)  [1.2 Что такое энтропия 8](#_Toc136004571)  [1.3 Как работает метод 9](#_Toc136004572)  [1.4 Какими бывают методы минимизации энтропии 10](#_Toc136004573)  [1.4.1 Перекрестная энтропия 10](#_Toc136004574)  [1.4.2 Информационная энтропия 11](#_Toc136004575)  [1.4.3 Метод максимальной энтропии 11](#_Toc136004576)  [1.4.4 Метод максимума правдоподобия 11](#_Toc136004577)  [1.4.5 Метод условной энтропии 12](#_Toc136004578)  [1.4.6 Метод отбора признаков 13](#_Toc136004579)  [Выводы 13](#_Toc136004580)  [2. РАЗРАБОТКА ПРОГРАММНОГО ПРОДУКТА 15](#_Toc136004581)  [2.1 Требования к проекту 15](#_Toc136004582)  [2.2 Разработка проекта 16](#_Toc136004583)  [2.3 Подход к разработке 17](#_Toc136004584)  [2.4 Области применения 18](#_Toc136004585)  [Выводы: 20](#_Toc136004586)  [3 ПРАКТИЧЕСКАЯ ЧАСТЬ 21](#_Toc136004587)  [3.1 Результаты проводимых расчетов и анализ 21](#_Toc136004588)  [3.2 Описание функций программы 22](#_Toc136004589)  [Выводы 23](#_Toc136004590)  [ЗАКЛЮЧЕНИЕ: 24](#_Toc136004591)  [СПИСОК ИСПОЛЬЗУЕМОЙ ЛИТЕРАТУРЫ 25](#_Toc136004592)  [ПРИЛОЖЕНИЕ А 26](#_Toc136004593) | | |

# ВВЕДЕНИЕ

В современном информационном обществе объемы данных растут с каждым днем, и умение извлекать полезную информацию из этих данных становится все более важным. Одним из ключевых шагов в анализе данных является отбор наиболее информативных признаков, которые содержат наибольшую информацию о целевой переменной или являются наиболее релевантными для задачи. Это позволяет улучшить качество моделей машинного обучения, снизить размерность данных, избавиться от шума и улучшить интерпретируемость результатов.

Одним из методов, используемых для отбора наиболее информативных признаков, является метод минимизации энтропии. Этот метод основан на понятии энтропии, которая измеряет неопределенность или разнородность данных. Чем меньше энтропия, тем более информативными являются признаки.

В данной работе рассмотрен метод минимизации энтропии и его применение для отбора наиболее информативных признаков. Мы представим реализацию этого метода на языке программирования Python и рассмотрим его область применения. Также будут представлены другие методы минимизации энтропии, такие как метод условной энтропии и метод максимума правдоподобия, и проведено сравнение их преимуществ и ограничений.

Целью данной работы является предоставить практический инструмент для отбора наиболее информативных признаков и дать обзор областей, в которых этот метод может быть применен. Результаты данного исследования могут быть полезными для специалистов в области анализа данных, машинного обучения, статистики и других дисциплин, где требуется извлечение информации из больших объемов данных.

# 1. ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

## 1.1 Основные сведения о методе

Метод минимизации энтропии - это статистический метод, который используется для решения задач, связанных с прогнозированием и принятием решений на основе ограниченной информации.

В приложении этот метод может использоваться для прогнозирования вероятностей возникновения событий на основе имеющихся данных. Например, если мы имеем набор данных о погоде в определенном районе за несколько лет, мы можем использовать метод минимизации энтропии, чтобы прогнозировать вероятность того, что в следующий день будет дождь или солнечно.

Применение метода минимизации энтропии может также помочь при принятии решений в условиях неопределенности. Например, если у нас есть некоторое количество инвестиционных возможностей, мы можем использовать метод минимизации энтропии, чтобы принять решение о том, какую из них выбрать, основываясь на вероятности получения максимальной прибыли.

Следует отличать методы снижения размерности от методов сжатия информации, например архивирования, предназначенных для максимального уменьшения объема образов при их хранении и транспортировании по сетям передачи данных. Методы снижения размерности призваны найти компромисс между максимальным уменьшением объема, занимаемого информацией об образе, и сохранением его способности быть идентифицированным (узнанным) и классифицированным. причем само сжатие (снижение размерности) производится, прежде всего, с целью облегчения решения указанных двух вычислительных задач. Например, в задаче идентификации личности по отпечатку пальца на основе графического образа отпечатка может формироваться последовательность длиной до 2500 бит, хотя исходное изображение имеет гораздо больший объем.

Метод минимизации энтропии также может использоваться для построения моделей машинного обучения, таких как решающие деревья и нейронные сети. В этих случаях метод минимизации энтропии используется для определения наилучшего разбиения данных на подгруппы, что помогает улучшить качество модели.

В целом, метод минимизации энтропии является мощным инструментом для работы с неопределенностью и прогнозирования вероятностей в различных областях приложений.

Метод минимизации энтропии основывается на теории информации и является одним из статистических методов анализа данных. Суть метода заключается в том, чтобы найти вероятностное распределение, которое наиболее соответствует имеющейся информации, но не содержит дополнительной неопределенности.

В контексте прогнозирования, энтропия может быть использована для измерения того, насколько много информации нам нужно для того, чтобы сделать верный прогноз.

Метод минимизации энтропии заключается в выборе вероятностного распределения, которое минимизирует энтропию. Другими словами, мы выбираем вероятностное распределение, которое максимизирует неопределенность наших прогнозов настолько, насколько это необходимо для учета имеющейся информации, но не больше.

В простейшем случае, метод минимизации энтропии может быть использован для прогнозирования бинарных событий (например, солнечно или дождливо). В этом случае мы ищем вероятностное распределение, которое максимизирует энтропию, при условии, что сумма вероятностей двух событий равна единице. Например, если у нас есть 5 дней, когда было солнечно и 5 дней, когда было дождливо, то мы можем выбрать вероятностное распределение 0,5 и 0,5, которое максимизирует энтропию.

В более сложных случаях, метод минимизации энтропии может быть использован для прогнозирования более сложных событий с помощью статистических моделей, таких как регрессионный анализ и машинное обучение.

## 1.2 Что такое энтропия

В теории информации и статистике, энтропия - это мера неопределенности или "беспорядка" в системе. Она позволяет оценить количество информации, необходимое для описания или предсказания состояния, или события.

Энтропия определяется как средняя степень неопределенности в вероятностном распределении. Чем выше энтропия, тем более неопределенной является система, а значит, тем больше информации требуется для ее описания или предсказания.

Формально, для дискретного вероятностного распределения событий {x1, x2, ..., xn}, энтропия H(X) вычисляется следующим образом:

H(X) = -Σ p(xi) \* log2(p(xi))

где p(xi) - вероятность события xi.

Энтропия измеряется в битах или натуральных единицах информации. Она достигает минимального значения, равного нулю, когда все события в распределении имеют вероятность 1 (то есть система полностью определена), и достигает максимального значения, когда все события равновероятны (то есть система максимально неопределена).

Энтропия имеет множество применений, включая прогнозирование, компрессию данных, принятие решений и моделирование статистических систем. Она является ключевым понятием в теории информации и играет важную роль в обработке и анализе данных.

## 1.3 Как работает метод

Метод минимизации энтропии работает путем нахождения оптимального вероятностного распределения, которое наиболее соответствует имеющейся информации, при условии минимальной энтропии.

В основе метода лежит концепция максимальной энтропии. Идея состоит в том, что в отсутствие дополнительной информации мы должны предполагать, что все события равновероятны. Таким образом, максимальная энтропия достигается, когда все возможные исходы равновероятны.

Однако, когда у нас есть дополнительная информация о системе, мы можем использовать эту информацию для более точного прогнозирования или принятия решений. Метод минимизации энтропии позволяет найти вероятностное распределение, которое учитывает имеющуюся информацию, но при этом минимизирует дополнительную неопределенность.

Процесс работы метода минимизации энтропии обычно включает следующие шаги:

Определение целевой функции: Целевая функция определяет меру энтропии и зависит от вероятностей событий. В общем случае, цель состоит в том, чтобы минимизировать энтропию, что может быть достигнуто путем определения оптимальных значений вероятностей.

Ограничения: Вероятностные распределения должны удовлетворять определенным ограничениям. Например, сумма вероятностей должна быть равна 1 или могут быть заданы другие ограничения в зависимости от конкретной задачи.

Поиск оптимального распределения: С помощью оптимизационных методов, таких как градиентный спуск или методы нелинейного программирования, итерационно находится оптимальное вероятностное распределение, которое минимизирует энтропию при соблюдении ограничений.

Валидация и применение: После нахождения оптимального распределения, его можно использовать для прогнозирования вероятностей, принятия решений или построения моделей, в зависимости от конкретной задачи.

## 1.4 Какими бывают методы минимизации энтропии

### 1.4.1 Перекрестная энтропия

В теории информации перекрёстная энтропия между двумя распределениями вероятностей измеряет среднее число бит, необходимых для опознания события из набора возможностей, если используемая схема кодирования базируется на заданном распределении вероятностей q , вместо «истинного» распределения p

Существует много ситуаций, когда необходимо измерить перекрестную энтропию, но распределение  p  неизвестно.

Примером является моделирование языка, когда модель создается на основе обучающего набора Т, а затем ее перекрестная энтропия измеряется на тестовом наборе, чтобы оценить, насколько точна модель в прогнозировании тестовых данных. В этом примере p  - истинное распределение слов в любом корпусе и q - распределение слов, предсказанное моделью. Поскольку истинное распределение неизвестно, перекрестная энтропия не может быть вычислена напрямую.

### 1.4.2 Информационная энтропия

Мера неопределённости некоторой системы, в частности, непредсказуемость появления какого-либо символа первичного алфавита. В последнем случае при отсутствии информационных потерь энтропия численно равна количеству информации на символ передаваемого сообщения.

Например, в последовательности букв, составляющих какое-либо предложение на русском языке, разные буквы появляются с разной частотностью, поэтому неопределённость появления для некоторых букв меньше, чем для других.

### 1.4.3 Метод максимальной энтропии

Метод максимальной энтропии позволяет оценить спектр ограниченной во времени последовательности по конечному числу выборок сигнала и соответственно значений корре-ляционной функции. Основная идея метода МЭ состоит в выборе такого спектра, который бы соответствовал наиболее непредсказуемому временному ряду, чья корреляционная функция совпадает с из-вестными коэффициентами корреляции. Он широко применяется в статистическом моделировании и машинном обучении.

### 1.4.4 Метод максимума правдоподобия

Этот метод используется для оценки параметров распределения данных, которое наиболее правдоподобно, исходя из наблюдаемых данных. Он может быть применен для нахождения вероятностного распределения с минимальной энтропией.

Метод максимума правдоподобия широко применяется в статистике и машинном обучении для оценки параметров вероятностных моделей, таких как линейная регрессия, логистическая регрессия, смесь распределений и другие. Он обладает хорошими статистическими свойствами и может быть применен к различным типам данных.

### 1.4.5 Метод условной энтропии

Этот метод используется для минимизации условной энтропии между двумя переменными при заданных условиях. Он позволяет определить зависимости и взаимосвязи между переменными и выбрать наиболее информативные признаки.

Условная энтропия определяет количество информации, необходимое для описания переменной, при условии, что известны значения другой переменной или группы переменных. Минимизация условной энтропии позволяет идентифицировать наиболее информативные признаки или комбинации признаков, которые обладают наибольшей способностью разделить или классифицировать данные.

Преимущества метода условной энтропии включают возможность выявления взаимосвязей и зависимостей между переменными, а также выделение наиболее информативных признаков, которые могут быть использованы для построения моделей и принятия решений.

Для оценки условной энтропии могут использоваться различные алгоритмы, такие как деревья решений, методы информационного прироста, алгоритмы максимальной правдоподобия и другие. Выбор конкретного алгоритма зависит от характеристик данных, типа переменных и требуемых результатов анализа.

### 1.4.6 Метод отбора признаков

Этот метод используется для отбора наиболее информативных признаков путем сравнения их вклада в энтропию или информационный прирост. Он позволяет уменьшить размерность данных и удалить ненужные или коррелирующие признаки, что способствует более эффективному моделированию и классификации данных.

Процесс отбора признаков включает следующие шаги:

А) Оценка важности признаков: Для каждого признака вычисляется его важность или информативность в контексте задачи анализа данных. Это может быть выполнено с использованием различных метрик, таких как информационный выигрыш, коэффициент корреляции, коэффициент дисперсии и другие.

Б) Ранжирование признаков: После оценки важности признаков они ранжируются по убыванию их информативности. Признаки с наибольшей информативностью считаются наиболее значимыми и выбираются для включения в итоговый набор признаков.

В) Выбор оптимального набора признаков: В зависимости от конкретной задачи и требований, можно выбрать определенное количество наиболее информативных признаков или определить порог важности, выше которого признаки будут включены в итоговый набор.

## Выводы

Минимизация энтропии - это важный метод в анализе данных, который позволяет идентифицировать наиболее информативные признаки или комбинации признаков. Он основан на концепции энтропии, которая измеряет неопределенность или разнородность данных. Чем ниже энтропия, тем более информативными считаются признаки.

Существуют различные методы минимизации энтропии. Один из них - метод отбора наиболее информативных признаков, который использует машинное обучение и рекурсивное исключение признаков. Другой метод - метод условной энтропии, который анализирует зависимости между переменными и выбирает признаки с наименьшей условной энтропией.

Кроме того, был упомянут метод максимума правдоподобия, который используется для оценки параметров вероятностных моделей на основе имеющихся данных. Он позволяет выбрать значения параметров модели, при которых вероятность наблюдаемых данных будет максимальной.

Методы минимизации энтропии имеют широкое применение в анализе данных, машинном обучении, статистике и других областях. Они помогают выделить наиболее информативные признаки, идентифицировать зависимости между переменными и построить модели, которые могут использоваться для прогнозирования, классификации и принятия решений.

В целом, методы минимизации энтропии представляют собой мощный инструмент в анализе данных, который помогает извлекать информацию и понимать структуру данных. Они играют важную роль в процессе принятия решений и оптимизации систем в различных областях и приложениях.

# 2. РАЗРАБОТКА ПРОГРАММНОГО ПРОДУКТА

## 2.1 Требования к проекту

Программа должна обеспечивать реализацию метода минимизации энтропии путем отбора наиболее информативных признаков, на основе введенных посредством файла исходных данных определенного формата и предоставлять конкретный результат в файле и графическом виде.

Программа должна обеспечивать возможность выполнения следующих

функций:

1. Загрузка данных: Открывать и считывать файлы формата .csv.
2. Визуализация данных: программа должна предоставлять

графическое представление данных.

1. Отбор признаков и вычисление энтропии.
2. Экспорт результатов в выходной файл (при наличии).

Выполнение программы должно происходить без сбоев, для этого необходимо обеспечить прохождение поэтапного анализа программы. К техническим требованиям относятся такие решения как использование современных инструментов разработки, таких как интегрированные среды разработки (IDE) или текстовые редакторы.

Совместимость с различными операционными системами.

Работа с различными форматами данных, такими как CSV, Excel, JSON и другими.

Реализация модульных тестов для проверки корректности работы алгоритма.

## 2.2 Разработка проекта

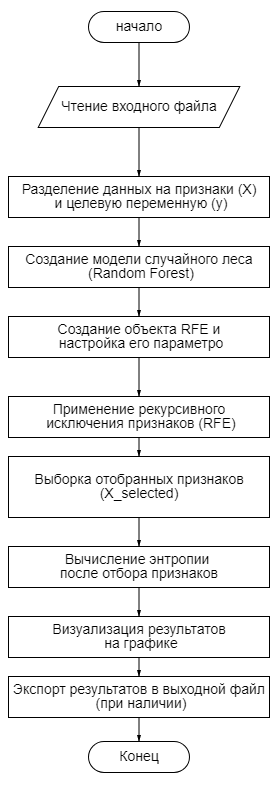


Рисунок 2.1 – Алгоритм работы программы.

## 2.3 Подход к разработке

Цель работы – разработка приложения, реализующего метод минимизации энтропии. Для этого необходимо:

1)изучить предметную область и узнать больше о методе;

2)составить схему алгоритма;

3)оформить программную документацию;

4)написать программный код на языке программирования Python;

5)протестировать приложение.

2.3.1 Выбор языка программирования

Выбор языка программирования Python для реализации метода минимизации энтропии обосновывается несколькими причинами:

1. Простота и читаемость кода: Python имеет простой и понятный синтаксис, который делает код более читаемым и позволяет разработчикам быстро понять его логику. Это особенно важно в научных и исследовательских областях, где ясность и понятность кода являются приоритетными факторами.
2. Богатая экосистема библиотек: Python имеет обширную экосистему библиотек и инструментов для анализа данных, машинного обучения и статистики, таких как NumPy, Pandas, Scikit-learn, SciPy и другие. Эти библиотеки предоставляют готовые реализации методов и функций, упрощая разработку и эксперименты.
3. Популярность и активное сообщество: Python является одним из наиболее популярных языков программирования в области анализа данных и машинного обучения. Это означает, что существует большое и активное сообщество разработчиков, готовых поделиться знаниями, помочь с проблемами и предоставить обновления и улучшения.
4. Переносимость и доступность: Python является переносимым языком программирования, поддерживаемым на различных операционных системах. Это делает программы, написанные на Python, доступными для широкого круга пользователей. Кроме того, Python имеет свободную и открытую лицензию, что делает его доступным для использования и модификации.

Python представляет собой мощный инструмент для анализа данных и машинного обучения, и его выбор для реализации метода минимизации энтропии обусловлен его простотой, гибкостью, богатой экосистемой и широкой поддержкой сообщества разработчиков.

## 2.4 Области применения

Метод минимизации энтропии путем отбора наиболее информативных признаков находит широкое применение во многих областях и задачах анализа данных. Некоторые из основных областей применения включают:

Машинное обучение и классификация: Отбор наиболее информативных признаков позволяет улучшить качество моделей машинного обучения и классификации. Выбор оптимального подмножества признаков может снизить размерность данных, устранить шум и избыточность, а также сделать модель более интерпретируемой и эффективной.

Анализ данных и исследования: Отбор информативных признаков помогает выделить наиболее значимые факторы и понять важные закономерности в данных. Это может быть полезно для проведения статистических исследований, выявления связей между переменными и понимания влияния факторов на целевую переменную.

Биоинформатика и геномика: В анализе геномных данных, где размерность данных очень высока, метод отбора признаков позволяет идентифицировать гены или маркеры, которые наиболее сильно связаны с конкретным фенотипом или болезнью. Это помогает в понимании генетических механизмов, выявлении генетических маркеров и разработке диагностических или терапевтических методов.

Обработка естественного языка: В анализе текстовых данных метод отбора признаков может использоваться для выделения наиболее информативных слов или фраз, которые вносят наибольший вклад в понимание или классификацию текстов. Это может быть полезно для автоматической обработки текста, категоризации документов, анализа тональности и других задач анализа текста.

Изображения и компьютерное зрение: В области обработки изображений метод отбора признаков может использоваться для выделения наиболее дискриминативных или характерных признаков, которые помогают в распознавании образов, сегментации объектов, извлечении характеристик и других задачах компьютерного зрения.

Это лишь несколько примеров областей, где метод отбора наиболее информативных признаков может быть применен.

Медицина и биология: В медицинских и биологических исследованиях метод отбора признаков используется для идентификации наиболее значимых биомаркеров, генов или фенотипических признаков, связанных с различными заболеваниями. Это может помочь в диагностике, прогнозировании и разработке индивидуализированных лечебных подходов.

Финансы и экономика: В финансовых и экономических исследованиях метод отбора признаков может использоваться для выделения наиболее информативных факторов, влияющих на финансовые показатели, акции, рыночные тренды и экономические прогнозы. Это помогает в принятии обоснованных инвестиционных решений и анализе финансовых рисков.

Интернет-реклама и маркетинг: В области интернет-рекламы и маркетинга метод отбора признаков может использоваться для определения наиболее важных факторов, влияющих на эффективность рекламных кампаний, поведение потребителей и предпочтения аудитории. Это помогает в оптимизации рекламных стратегий и персонализации маркетинговых подходов.

Обработка естественного языка: В области обработки естественного языка метод отбора признаков используется для выделения наиболее значимых слов или фраз, которые вносят наибольший вклад в классификацию текстов, анализ тональности, извлечение информации и другие задачи. Это помогает в улучшении точности и эффективности алгоритмов обработки текста.

Компьютерное зрение и распознавание образов: В области компьютерного зрения метод отбора признаков применяется для выделения наиболее дискриминативных признаков изображений, которые помогают в распознавании объектов, классификации изображений, сегментации и других задачах обработки изображений.

## Выводы:

Метод может использоваться во многих других задачах, где требуется снизить размерность данных, повысить качество моделей и улучшить понимание данных. Его применения возможно в разных областях, он применим как в медицине, так и в космонавтике, его применения зачастую играет ключевую роль во многих информационных процессах, именно по этому данному методу уделяется отдельной место в информационной теории.

# 3 ПРАКТИЧЕСКАЯ ЧАСТЬ

## 3.1 Результаты проводимых расчетов и анализ

В качестве первого контрольного примера введем в программу набор значений, представленных на рисунке 2.



Рисунок 3.1 – Вводимые значения.

Данные значения представляют собой стандартную матрицу три на три, записанные в файл формата .сsv и выбранные в качестве входного файла.

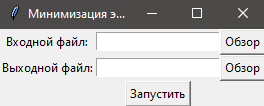


Рисунок 3.2 – Окно выбора входных и выходных данных.

После введения значений, программа представит нам график, показывающий зависимость энтропии от индекса признака после отбора наиболее информативных признаков. Таким образом мы узнаем какие признаки являются наиболее информативными после отбора.

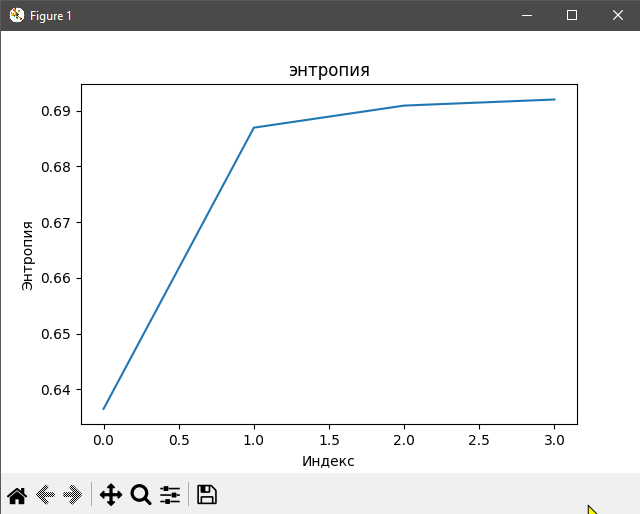


Рисунок 3.3 – График зависимости.



Рисунок 3.4 – Результат выводимый в файл.

## 3.2 Описание функций программы

def minimize\_entropy():

Минимизирует энтропию путем отбора наиболее информативных признаков.

:param input\_file: Путь к входному CSV-файлу.

:param output\_file: Путь к выходному файлу. Если не указано, результаты не будут экспортированы.

:param delimiter: Разделитель, используемый в CSV-файле.

def browse\_input\_file():

Открываем диалоговое окно для выбора входного CSV-файла.

def browse\_output\_file():

Открываем диалоговое окно для выбора выходного файла.

def run():

Запускаем функцию minimize\_entropy с выбранными входными и выходными файлами.

## Выводы

Осуществлена проверка результата, полученного разработанной программой и дано обоснование результатам работы программы В ходе разработки программного продукта мы определили базовый функционал программы, проверили работу программы на основе данного набора значений. Построили график, показывающий результат метода минимизации энтропии путем отбора информационных признаков.

# ЗАКЛЮЧЕНИЕ:

В данной работе мы рассмотрели метод отбора наиболее информативных признаков и его области применения. Метод отбора наиболее информативных признаков является важным инструментом в анализе данных и машинном обучении, позволяющим выделить наиболее значимые признаки, содержащие полезную информацию о целевой переменной или задаче.

Мы представили реализацию метода на языке программирования Python, что облегчает его использование и применение в различных задачах. Также мы рассмотрели различные методы минимизации энтропии, такие как метод условной энтропии и метод максимума правдоподобия, и провели сравнение их преимуществ и ограничений.

Области применения метода отбора наиболее информативных признаков охватывают медицину и биологию, финансы и экономику, интернет-рекламу и маркетинг, обработку естественного языка, компьютерное зрение и многие другие. Этот метод помогает улучшить качество моделей, снизить размерность данных, повысить интерпретируемость результатов и принимать обоснованные решения на основе данных.

Выводы из этой работы могут быть полезными для специалистов в области анализа данных, машинного обучения, статистики и других дисциплин, которым необходимо работать с большими объемами данных и извлекать полезную информацию. При правильном применении метода отбора наиболее информативных признаков можно достичь более точных моделей, более эффективных решений и лучшего понимания данных.

# СПИСОК ИСПОЛЬЗУЕМОЙ ЛИТЕРАТУРЫ

1. Никлаус Вирт «Методы». 1985 – 274 с. – ISBN 978-5-94074-584-6
2. Д.М. Златопольский «Программирование: типовые задачи, алгоритмы, методы». 2020 – 226 с. – 978-5-00101-789-9
3. Луридас Панос. «Библиотеки Python». 2018 – 608 с. – ISBN 978-5-04-089834-3
4. Дональд Кнут «Искусство программирования. Том 1.. 1998 – 722 с. – ISBN 978-5-8459-1984-7
5. Джон Клейнберг, Эва Тардос. «разработка и применение». 2016 – 800 с. – ISBN 978-5-496-01545-5
6. Панос Луридас «Алгоритмы. Самый краткий и понятный курс». 2020 – 195 с. – ISBN 978-5-04-167804-3
7. Генри С. Уоррен мл. «Трюки для программистов». 2-е издание. 2014 – 512 с. – ISBN 978-5-8459-1838-3
8. Джордж Хайнеман, Гари Поллис, Стэнли Селков «Алгоритмы. Справочник с примерами на C, C++, Java и Python». 2017 – 432 с. – ISBN 978-5-9908910-7-4
9. Сложность алгоритмов. Big O. Основы. [Электронный ресурс]. – URL: <https://bimlibik.github.io/posts/complexity-of-algorithms/> (дата обращения 03.05.2022)

Основные виды сортировок и примеры их реализации [Электронный ресурс]. – URL: <https://academy.yandex.ru/posts/osnovnye-vidy-sortirovok-i-primery-ikh-realizatsii> (дата обращения 03.05.2022)

1. Иванов, В. М. Интеллектуальные системы : учеб. пособие для вузов / В. М. Иванов ; под науч. ред. А. Н. Сесекина. — М. : Издательство Юрайт, 2017. — 91 с.

# ПРИЛОЖЕНИЕ А

Код программы:

import csv

import numpy as np

from scipy.stats import entropy

from sklearn.feature\_selection import RFE

from sklearn.ensemble import RandomForestClassifier

import matplotlib.pyplot as plt

from tkinter import Tk, Label, Entry, Button, filedialog

def minimize\_entropy(input\_file, output\_file=None, delimiter=','):

"""

Минимизирует энтропию путем отбора наиболее информативных признаков.

:param input\_file: Путь к входному CSV-файлу.

:param output\_file: Путь к выходному файлу. Если не указано, результаты не будут экспортированы.

:param delimiter: Разделитель, используемый в CSV-файле.

"""

# Читаем данные из CSV-файла

with open(input\_file, 'r') as f:

reader = csv.reader(f, delimiter=delimiter)

data = np.array([row for row in reader], dtype=float)

# Разделяем признаки и целевую переменную

X = data[:, :-1]

y = data[:, -1]

# Создаем модель машинного обучения для дальнейшего использования в программе

model = RandomForestClassifier()

# Создаем объект RFE и настраиваем его

rfe = RFE(estimator=model, n\_features\_to\_select=3) # Здесь можно указать желаемое количество признаков

# Применяем рекурсивное исключение признаков

X\_selected = rfe.fit\_transform(X, y)

# Вычисляем энтропию после отбора признаков

ent = entropy(X\_selected.T)

# Визуализируем результаты

plt.plot(ent)

plt.title('энтропия')

plt.xlabel('Индекс')

plt.ylabel('Энтропия')

plt.show()

# Экспортируем результаты в файл

if output\_file:

with open(output\_file, 'w') as f:

writer = csv.writer(f, delimiter=delimiter)

writer.writerow(ent)

# суета по первому окну

def browse\_input\_file():

"""

Открываем диалоговое окно для выбора входного CSV-файла.

"""

input\_file = filedialog.askopenfilename(filetypes=[('CSV Files', '\*.csv')])

input\_entry.delete(0, 'end')

input\_entry.insert(0, input\_file)

def browse\_output\_file():

"""

Открываем диалоговое окно для выбора выходного файла.

"""

output\_file = filedialog.asksaveasfilename(defaultextension='.csv', filetypes=[('CSV Files', '\*.csv')])

output\_entry.delete(0, 'end')

output\_entry.insert(0, output\_file)

def run():

""

Запускаем функцию minimize\_entropy с выбранными входными и выходными файлами.

"""

input\_file = input\_entry.get()

output\_file = output\_entry.get()

if not input\_file:

print('Пожалуйста, выберите входной файл.')

return

minimize\_entropy(input\_file, output\_file)

# Создаем GUI

root = Tk()

root.title('Минимизация энтропии')

# Входной файл

input\_label = Label(root, text='Входной файл:')

input\_label.grid(row=0, column=0)

input\_entry = Entry(root)

input\_entry.grid(row=0, column=1)

input\_button = Button(root, text='Обзор', command=browse\_input\_file)

input\_button.grid(row=0, column=2)

# Выходной файл

output\_label = Label(root, text='Выходной файл:')

output\_label.grid(row=1, column=0)

output\_entry = Entry(root)

output\_entry.grid(row=1, column=1)

output\_button = Button(root, text='Обзор', command=browse\_output\_file)

output\_button.grid(row=1, column=2)

# Кнопка запуска

run\_button = Button(root, text='Запустить', command=run)

run\_button.grid(row=2, column=1)

root.mainloop()