***ФГБОУ ВО «Кубанский государственный аграрный университет имени***

***И.Т. Трубилина»,***

***Российская Федерация***

***Колесников Роман Юрьевич, ПИ2102***

***roman563412@gmail.com***

**Практическое занятие №10. По дисциплине "Теория информации, данные и знания".**

**Заголовок**

Практическое занятие: Очистка данных, моделирование и кластеризация в системе Эйдос (данные клиентов и обзоров)

**Резюме**

Практическое онлайн-занятие №10 по дисциплине "Теория информации, знания" для группы ИТ-2003 состоялось 24 ноября 2020 года под руководством профессора Луценко Е.В. В начале занятия профессор предложил студентам выбрать формат: помощь в решении их задач или опрос по пройденному материалу и основным понятиям. Студенты выбрали помощь с задачами.

Основная часть занятия была посвящена анализу данных, предоставленных студенткой Анной (предположительно, данные о клиентах торгового центра). Процесс включал:

Подготовку данных: Загрузка данных, проверка корректности диапазона, копирование в новый лист Excel, удаление некорректных данных. Идентификация типов шкал (возраст как классификационная, остальные – описательные). Проверка числовых столбцов на наличие нечисловых значений.

Моделирование (Модель 1): Построение модели в системе Эйдос с использованием возраста (разделенного на 20 интервалов) как классификационного признака для предсказания пола, годового дохода и оценки расходов.

Фильтрация шума: Применение методики отсечения наименее достоверных данных (оставлено 10% наиболее достоверных) для повышения качества модели (Модель Инф1).

Кластеризация: Проведение кластерного анализа на основе улучшенной модели. Результаты показали схожесть самых молодых и самых старых клиентов по выбранным параметрам, но в целом были признаны неочевидными.

Моделирование (Модель 2): Построение альтернативной модели, где каждый ID клиента рассматривался как отдельный класс. Этот подход занял больше времени, но показал высокую достоверность после фильтрации шума (Модель Инф3). Были выявлены характерные признаки для групп с высоким и низким рейтингом.

Далее обсуждался анализ данных студента Марка (большой объем, ~300 МБ, предположительно, данные фондового рынка или обзоров товаров). Были затронуты вопросы:

Конвертации данных большого объема (проблемы с онлайн-конвертерами и встроенным конвертером Эйдос из-за запятых).

Анализа текстовых данных (обзоров) путем выделения слов как признаков.

Необходимость определения целевого признака (например, рейтинга пользователя).

В конце занятия профессор упомянул о разработанном им методе сценарного анализа для динамических данных, сославшись на свои публикации (в т.ч. в Библиотеке Конгресса США и ResearchGate), и предложил поделиться обработанными файлами с Анной и Марком.

Детальная расшифровка текста

1. Введение и организационные моменты

Добрый вечер, ребята.

Какие-то вы все знакомые.

Что-то вы такие знакомые, я не знаю.

Так.

Сегодня 24 ноября 2020 года.

Шестая пара. 17:20–18:50.

Практическое занятие номер 10 с группой ИТ-2003 по дисциплине "Теория информации, знания".

Занятие ведёт профессор Луценко Евгений Вениаминович.

И мы можем продолжить обсуждение тех вопросов, которые у нас были.

Люся, там нормально всё?

Да.

Это Никита позвонил.

Угу.

2. Варианты проведения занятия

Ну, какие у нас варианты? У нас вариант такой: либо смотрите, как вы реализуете какие-то свои задачи, могу вам помочь, либо я вас опрашиваю по тому, как вы решаете задачи, какие основные понятия у вас в этой области сформированы и так далее.

Давайте выбирайте, что вы выбираете.

Предлагайте свои варианты.

Там в чате написали.

Угу. Сейчас, сейчас. Мне часто писала, мне тут зав кафедры. Зав кафедры написала.

3. Анализ данных Анны

Так. Ага. Анна.

Так, сейчас тогда я сделаю экран.

Анна, Анна. Ну что-то не вижу я.

На может быть на ту почту направили? lutsenko@mail.kubsu.ru. Вот эту?

Да. Ну что-то нету. И в спаме нету, нигде нету.

Ну я слышал, что Gmail так иногда делает. Я не видел, чтобы у меня он не работал, но говорят, что вот посылают, а он там через час там, через 2 часа показывает.

А вот, сейчас Ане.

Ну и что здесь у нас за клиенты?

Что они от нас хотят?

3.1. Подготовка данных

Смотрите, ребята, область данных шире, чем сами данные. Видите, она захватывает ещё одну колоночку. Поэтому я выделяю область данных фактическую, копирую и создаю новый лист, а туда вставляю. А старый удаляю, потому что он некорректный. Лист.

Вот. Ну что, мы можем взять, допустим, сделать анализ по возрасту и по определить, какой годовой доход и оценка расходов соответствует какому возрасту и какому полу, например. Или просто по полу. Ну по полу будет неинтересно, по возрасту. По возрасту можно сделать. Можно определить клиенты какого возраста, э-э, какой у них пол, э-э, годовой доход и расход. И понять, что вот в таком возрасте такое, в таком такое. Вот, по-моему, так надо делать.

То есть пол это у нас будет классификационная шкала. Нет, возраст, возраст. Классификационная шкала.

Вот. Теперь что смотрим ещё? Значит, нужно проверить, нет ли там не чисел в этих колонках. Как это можно сделать? Я знаю самый простой способ такой: берём последнюю колонку и смотрим: B2, D101. То есть всё правильно. Вот. Берём предыдущую, то есть следующую колонку. Тоже смотрим. Э, не, а, ну да, это шкала не числовая. Ёлки-палки. Вот. Смотрим на эту колонку. D2, всё правильно. И смотрим на эту. Если там где-нибудь есть э какие-то символы, то надо их заменять на пробелы эти символы.

Вот. Ну вот так вот. Значит, смотрим, у нас вторая колонка классификационная, третья, четвёртая, пятая - описательные.

Так, классификационная, третья, четвёртая, пятая - описательные.

Удаляем все приложения.

Вот. Так, колонка классификационная, третья, четвёртая, пятая - описательные.

Исходные данные. Файл исходных данных, который я поправил немножко. Угу. Он ещё и в старом Excel.

Вот, я вам посылаю, чтобы у вас были файлы с параметрами.

3.2. Моделирование (Модель 1) и фильтрация шума

Ну, явно надо возрастов побольше взять, чтобы видеть интервальных значений. Сейчас посмотрим, может ещё больше возьмём. Вот если когда 10 интервальных значений возраста, то у нас получилось по 20 примеров на возрастной диапазон. Я думаю, что можно было бы и побольше взять. Как-то, ну 20, например.

Ну так вроде получилось разумно.

Смотрим ещё на модель сейчас. Что там у нас с моделью будет.

То есть у нас по каждому возрастному диапазону 10 примеров в нашей выборке. Идём на создание модели.

И смотрим. 20 у нас градаций возраста. Вот, и у нас ещё пол, годовой доход и оценка расходов. И обучающая выборка. 200 примеров.

И давай посмотрим. Как это называется задача?

Ну вот это сообщение выдаётся, потому что это многовато, конечно.

Ну остаётся такой вариант, э-э, запасной у нас. Вот здесь вот мы можем написать, оставить 10% наиболее достоверной информации. Значит, поскольку система э-э измеряет уровень уверенности в своём решении, и этот уровень уверенности она очень адекватно измеряет, то есть действительно это соответствует доле истинных и ложных решений. То есть чем выше уровень достоверности, тем выше доля истинных решений. То можно, пользуясь этим критерием, удалить заведомо ложные решения вот из э-э модели. И тогда получится, что мы как бы удалили шум, вот так получается. Выделим наиболее существенное, наиболее разумные результаты. Вот, с высоким уровнем сходства. И тогда получается картина гораздо, гораздо лучше. И так можно вытащить любую модель, подавив шум. Это очень эффективный алгоритм подавления шума. К тому же он ускоряет расчёты, потому что уменьшается размерность, ну, размер баз данных уменьшается, раз из них шум удалён.

Вот. Поэтому смотрим. Модель Инф1. Ну стало гораздо лучше, я скажу. Гораздо лучше.

Вот, возьмём вот эту.

Ну, в принципе, этого достаточно, чтобы можно было уверенно говорить о том, что задача эта у нас получилась, мы можем её анализировать и описывать. Вот именно с такими параметрами, ребята.

3.3. Кластеризация

Ну что, здесь у нас за клиенты? Что они от нас хотят? Кластеризация? Кластеризация, так кластеризация.

Значит, давай сразу 4 2 2 1. Матрица сходства. И потом... Это в модели Инф1. Это то, что нам надо.

Вот.

Ну тут есть интересные особенности.

Модель Инф1. Проводим кластеризацию.

Ну и тут что мы видим? Что, как это ни странно, самые молодые похожи на самых старых по вот этим своим параметрам. И... Ну тут, в принципе, это удивительно, а дальше мы здесь что видим? Э-э, 12, здесь 18, 9, 12, 13. А здесь 7, 2, 9, 10. Ну, в общем, неочевидный результат кластеризации, странный какой-то.

3.4. Моделирование (Модель 2)

Ещё есть другой вариант. Сейчас вот я сообразил. Он займёт больше времени для расчёта.

Какой вариант есть? Смотрите. Возраст тоже как признак использовать, добавить туда ещё одну колонку "покупатель" и провести кластеризацию всех покупателей: 1, 2, 3, 4. На основе каждого покупателя сделать класс. Класс сделать. Эту колонку сделать обычной, признаки. А здесь написать... Вот.

Тогда у нас что получится? Смотрите, что мы сейчас всех покупателей просто рассмотрим как отдельных, э-э, отдельные объекты, на основе которых мы сформируем образы класса. И потом их все друг с другом сравним. Единственное, это займёт больше времени, потому что размерность больше. Вторая, третья, четвёртая, пятая, шестая. То есть берём, делаем модель два. 2, 3, 2, 2. Третья по шестую колоночку, а всё остальное то же самое. Модель два.

Вот. Вот так. Значит, смотрим.

У нас получилось 200 классов и 10... Пусть так будет, наверное. Может, многовато, а может, меньше сделать, не знаю.

Смотрим. Количество классов у нас выступает рейтинг читательский. 10 градаций рейтинга. В качестве информации - это наименование. Наименование у нас разбилось на слова, которых аж 1047. Вот. Авторы тоже разбились на слова, которых уже 456 стало. А дальше идут какие-то предварительные просмотры, цена, год издания и ещё что-то такое, не знаю, эффективная колонка.

Вот. И синтез и верификация модели. Давай сделаем 10%, иначе там есть вероятность, что будет не очень. Ну посмотрим, может быть, сейчас посчитает.

Синтез и верификация модели у нас заняли 1 минуту 13 секунд. Ну это, ребята, в разумных пределах. Название отсюда перетаскиваем. Нижнее приложение - это то жёлтенькое, это текущее.

Вот. И смотрим, какие модели наиболее достоверные. Вот получилась такая у нас картина. Довольно высокая достоверность модели Инф3 по F-критерию. Вот. Это очень высокая достоверность, но я не знаю, как это будет выглядеть. Сейчас посмотрим. Даже неплохо. Да.

Странно, странно. Ну, 85% достоверные решения. Ладно, будем так считать. Ну главное - кластеризацию провести. Так же они от нас хотели. Дождаться кластеризации.

4. Анализ данных Марка (обсуждение)

Так, ну народу много. В общем, пытаемся решить задачу с Кагла. Вот эту.

Марк, у тебя слышишь, нет? Да, слышу. Здесь вот есть материал. Читать его в ResearchGate это довольно-таки несерьёзно, он здоровенный, там около 100 листов. Вот. Значит, есть смысл просто его скачать вот здесь вот и посмотреть. Ну я думаю, ты поймёшь там всё. Если что-то непонятно, просто дальше читай, да и всё, потом всё равно всё поймёшь смысл. Понял?

Ну тут идея в чём заключается? В том, чтобы, что система Эйдос имеет режимы, позволяющие ей как признак рассматривать динамику изменения значений в какой-то колоночке, а как результат - динамику изменения тоже колоночки в будущем. То есть анализируются все колоночки, все. И рассматривается как факторы сценарии, как вот способ, что ли, изменения значений за прошлый период, а как классы - сценарии изменения во всех колоночках в будущий период. И как что, на что влияет. Вот как вот менялось там в такой-то колоночке, вот так вот влияет на то, что будет в другой колоночке либо в этой же. Понял, да, о чём речь? Вот. Ну это сценарный подход называется. Сценарный АСК-анализ. Это очень мощный метод, очень интересный. Вот. И я его разработал в девяносто четвёртом году, а потом совершенствовал. Впервые я вот его здесь описал, этот подход.

Сейчас э-э ноябрьский номер находится в размещении. И в этом ноябрьском номере будет продолжение вот этой статьи. Здесь описана теория, а там будет описан численный пример, как э-э я это всё делал. А, вот ещё что, сейчас я найду. Сейчас я найду. Я сейчас вспомнил.

Вот здесь она есть целиком.

Марк, меня слышишь, нет? Да, слышу. Вот это вот 42 МБ мы не будем обрабатывать. Ну это возможно, но мы не сможем записать в облако это приложение, потому что там больше 10 МБ сразу же стираются. Понимаешь? Я ж писал на своём этом, в своей этой инструкции к заданию, что не должны у нас быть данные исходные больше 10 МБ. А лучше вообще, чтобы они в килобайтах измерялись. Сделаем это. Мы можем это приложение посчитать, но мы его не сможем в Эйдос-облаке разместить, понимаешь? Поэтому я не советую это делать. Ну если тебе интересно, можно, конечно, попробовать.

Открытие, максимум, минимум, закрытие, объём, взвешенная цена. Ну метку её это как информацию о записи надо использовать, это не шкала. Открытие, максимум, минимум, закрытие, объём, взвешенная цена. Ну такая база, знаешь, она должна исследоваться в динамике, то есть это надо сценарный подход использовать. А сценарный подход описан в моей статье, я вам ещё не показывал этого ничего. Сейчас я статью свою найду, дам ссылочку.

Так, этот конвертер, замечательный конвертер этот, Конвертио, он больше 100 МБ не захочет конвертировать. 300 МБ весит. Ну, конечно, такие задачи очень интересно решать. Я тоже увлекаюсь этим, то есть мне вот интересно где-то на пределе возможностей попробовать. То, что невозможно. Попробую сделать то, что невозможно. Иногда получалось, ребят, сделать то, что невозможно. Сейчас такая теория господствует, что землетрясения полностью обусловлены внутренними факторами Земли и что космическая среда никакой роли не играет. А я провёл исследование с авторами, ну, очень большом объёме данных, начиная с конца XIX века, за 100, за 119 лет, короче говоря, данные. Вот. И там использовались, исследовались данные о землетрясениях, где, когда, что, сколько, как там, в общем, все эти характеристики, и использовалась информация о том, как Луна была расположена, Солнце относительно Земли и того места, где происходило землетрясение. И даже как были планеты расположены в Солнечной системе. И оказались, ребята, довольно явно выраженные прямые зависимости. То есть, обработав огромный объём данных, э-э, в системе Эйдос было выяснено, что существует зависимость процессов на Земле от космической среды. Сам факт вот этот, понимаете, что существует такая зависимость именно землетрясений. Вот. А также мы исследовали там и социальный статус, и экономические ситуации, то есть всё это зависит, оказывается, от космической среды. Вот. Зависимость э-э некоторых вещей вообще вообще жёсткая, а некоторых она помягче, не так сильно, но тоже зависит, есть зависимость. Объём данных таков, что для человека, если он будет читать всю жизнь с 8:00 часов до 17:00 с перерывом на обед, и с 18:00 до 60:00, то он ему нужно 30 жизней, чтобы прочитать этот объём данных. Поняли, к чему я клоню, да? То есть никакого сравнения с естественным интеллектом здесь быть не может, потому что человек не сможет даже прочитать эти данные, даже небольшую часть, не то, что там что-то проанализировать. То есть только эта технология это позволяет сделать. Других таких исследований нет, ребята, в мире, где вот такие объёмы данных были бы исследованы, выявлены такие зависимости. Больше таких, ну, не знаю, как это сказать попроще, чокнутых или идиотов, которым занимались тратой своего времени на такие задачи, больше не нашлось на Земле. Вот. Я об этом несколько книг написал, они размещены в Библиотеке Конгресса США эти книги. Ну я не целиком их писал сам, а с авторами писали свои разделы, я свои. Вот, Библиотека Конгресса США. И читаем, что здесь написано. Видно вам, ребят, нет? Library of Congress. Да-да. И поиск Луценко. Найдено раз Луценко, два, три. Это очень много, ребята. Вот одна работа девяносто девятого года, другая 2012, третья 2017. Это монографии. Это всё монографии, довольно толстенькие. Вот. Ну, я вам скажу так примерно, вот одна такая вот публикация монографии в Библиотеке Конгресса США примерно сопоставима с десятью публикациями Web of Science, вот так. Это или двадцатью, я не знаю. Ну, вообще-то таких вот, э-э, насколько я знаю, никого больше нету. Вот тут посмотреть, у кого ещё там монографии есть из университета. Посмотрите. Интересно просто. С Краснодара есть, кстати. Задавал как-то довольно много было из Краснодара. Вот по Краснодару нашло оно. 88 страниц. Довольно прилично.

5. Заключение

Так что осталось несколько минут до конца занятия. Ребят, какие вопросы у вас есть?

Вопросов нет.

Вопросов нет. И вот что я ещё хочу сказать, что вот когда эта дендрограмма досчитается, я надеюсь, она скоро досчитается, вот, там где 200 у нас этих, хотел сказать, пациентов супермаркета, клиентов супермаркета, то я тоже пошлю там Александре, я не помню кому там, ну в общем, вспомню, разберусь, пошлю эту дендрограмму. Александре, кажется, да? Да, Александре. Вот, пошлю. А нет, Ане, Ане. Ане пошлю. Александре мы всё сделали. Ане пошлю вот эти вот мильню, дендрограмму. А Марку пошлю CSV-файл этот, экселевский файл. Хорошо?

И на этом, ребята, до свидания, конец занятия. Всего самого хорошего. Надеюсь, вы что-нибудь новое узнали, полезное.