***ФГБОУ ВО «Кубанский государственный аграрный университет имени И.Т. Трубилина»,***

***Российская Федерация***

**175 Информационные системы и технологии. Практическая 4. работа 3.0.3 2020-11-02**

**Заголовок:** АСК-анализ в системе Эйдос: Лабораторная работа 3.03 и методика научного исследования

**Резюме текста:**

**1. Введение и организационные моменты:**
Проводится практическое занятие №4 для магистерской группы ИТ-1941 по дисциплине "Интеллектуальные системы и технологии" 2 ноября 2020 года. Занятие посвящено полному прохождению лабораторной работы 3.03 и демонстрации типовой методики научного исследования в системе Эйдос с применением автоматизированного системно-когнитивного анализа (АСК-анализа). Студентам предлагается запустить систему Эйдос на своих компьютерах.

**2. Подготовка к работе в системе Эйдос:**
Перед началом работы рекомендуется установить систему Эйдос (демонстрируется установка на диск C) и удалить предыдущие приложения для чистоты эксперимента. Демонстрируется загрузка данных для лабораторной работы 3.03 через диспетчер приложений (режим 1.3), выбор типа работы и установка переключателя на нужную ЛР (3.03). Система предупреждает об удалении существующих данных в папке исходных данных.

**3. Работа с исходными данными:**
Объясняется интерфейс для импорта данных из внешних источников (API), в частности, из Excel-файлов. Демонстрируется структура типового Excel-файла: строки – наблюдения, колонки – шкалы (классификационные и описательные). Шкалы могут быть числовыми или текстовыми. Подчеркивается, что наличие нечислового символа в числовой шкале переводит ее в текстовый формат. Объясняется важность корректного выделения области данных (Ctrl+Shift+End) и проблемы, возникающие при наличии "мусора" за пределами фактических данных (пустые строки/столбцы), что может привести к неверной интерпретации шкал системой. Рекомендуется нормализовать данные, скопировав их на новый лист.

**4. Формализация предметной области:**
После импорта данных происходит их преобразование во внутренний формат (.dbf). Система подсчитывает количество шкал разных типов и число градаций. Этот этап (создание классификационных/описательных шкал и градаций, кодирование исходных данных и получение обучающей выборки) называется формализацией предметной области. Демонстрируется режим 6.4, где описаны этапы системно-когнитивного анализа. Суть АСК-анализа: выявление причинно-следственных связей и преобразование данных в информацию, а затем в знания для решения задач управления и принятия решений.

**5. Синтез и верификация моделей:**
Синтез и верификация моделей выполняются в режиме 3.5. Обсуждаются новые возможности системы: расчеты на графическом процессоре (GPU), что может ускорять процесс до 4000 раз, и возможность раздельного выполнения синтеза и верификации. Это полезно для больших наборов данных, где полная верификация может занять много времени (дни, недели). Предлагается технология: выполнить быстрый синтез на всех данных, а верификацию провести на небольшой случайной выборке (например, 100 объектов) для оценки времени и экстраполяции на полный объем данных или на доступное время. Система также позволяет ограничивать объем файла результатов распознавания (до 2 ГБ) и удалять наименее достоверные результаты (с низким уровнем сходства), что рассматривается как удаление шума. Демонстрируется запуск синтеза и верификации для ЛР 3.03.

**6. Оценка достоверности моделей:**
Результаты синтеза и верификации включают созданные модели (матрицы абсолютных и относительных частот, информативности, ХИ-квадрат, Рои и др.) и оценку их достоверности. Классическая мера достоверности Ван Ризбергена (F-мера) основана на подсчете истинно положительных (TP), истинно отрицательных (TN), ложноположительных (FP, ошибка I рода) и ложноотрицательных (FN, ошибка II рода) решений при идентификации. Объясняется суть этих ошибок и их опасность (примеры с ПВО). Представлены формулы для полноты, точности и F-меры.

**7. Недостатки классической F-меры и ее обобщения:**
Указываются недостатки F-меры Ван Ризбергена: проблема мультиклассовости (объект может относиться к нескольким классам), проблема нечеткости (мера бинарна – истина/ложь, не учитывает степень уверенности системы) и зависимость от объема выборки (сумматоры TP, TN, FP, FN линейно растут с объемом). Предложены авторские обобщения F-меры: нечеткое мультиклассовое и инвариантное относительно объема данных. Инвариантная мера более устойчива на малых выборках и дает более адекватную оценку достоверности, часто выше классической, так как штрафует за ошибки пропорционально степени неуверенности системы.

**8. Взаимосвязь задач анализа данных:**
Задачи распознавания, классификации, идентификации и диагностики рассматриваются как синонимы. Задача идентификации и прогнозирования математически схожи, но различаются временной интерпретацией данных (прошлое/настоящее для идентификации, прошлое/будущее для прогнозирования). Задача принятия решений может решаться через прогнозирование или как обратная задача прогнозирования (поиск факторов для достижения целевого состояния). Исследование предметной области включает решение всех этих задач.

**9. Заключение и анонс:**
На следующем занятии планируется рассмотреть решение задач идентификации, прогнозирования, принятия решений и исследования предметной области на основе созданных моделей. Студентам предлагается задать вопросы.

**Детальная расшифровка текста:**

**1. Введение и организационные моменты:**

Здравствуйте, ребята.

Микрофоном отвечайте. Здравствуйте, здравствуйте.

Микрофоном я сказал отвечайте. Сергей. Идёт же запись, под запись идёт всё это.

Да.

Сегодня у нас 2 ноября 2020 года. Вторая пара, которая идёт 9:45-11:15. Практическое занятие номер четыре, на котором мы должны полностью пройти лабораторную работу 3.03. И я должен вам показать типовую методику проведения научного исследования в системе Эйдос с применением автоматизированного системно-когнитивного анализа в системе Эйдос.

Дисциплина: Интеллектуальные системы и технологии в магистерской группе ИТ-1941. Занятие ведёт профессор Луценко Евгений Вениаминович.

Ребята, у нас сейчас практическое занятие, поэтому очень желательно, чтобы вы запустили на своих компьютерах систему Эйдос. Значит, в чате, пожалуйста, поставьте плюсики, кто это может сделать. И минусики те, кто не может.

Борис, почему вы не можете это сделать?

А, понятно. На остановочник, и работаешь, и учишься. Понятно.

Ну а остальные ребята? Вот двое ответили, а сейчас у нас на совещании семь человек. Александр может, молодец.

Ну ладно. Значит, желательно, чтобы вы повторяли то, что я делаю, чтоб вы не только, так сказать, знания приобретали, но и навыки.

Ребята, а вот остальные почему ни плюсиков не поставили, ни минусиков? А? Три чека поставили. А четыре не поставили, ничего. Почему?

Ну понятно с вами всё.

**2. Подготовка к работе в системе Эйдос:**

Я устанавливаю систему Эйдос на диске C. И сейчас начнём работу.

Запускаем систему. И сразу же удаляем все существовавшие до этого приложения, ну и хотя их там и не было, но просто я, чтобы вы привыкли это делать, показываю вам таким образом.

Потом заходим в диспетчер приложений, режим 1.3. Нажимаем на кнопочку, на которой жирным шрифтом написано "Добавить лабораторную работу". Выбираем в верхней части формы экранной третьего типа лабораторную работу и ставим переключатель на лабораторную работу 3.03.

Нажимаем О'кей. Появляется сообщение, что сейчас будут удалены все данные из папки, предназначенной для исходных данных. Если там что-то было ценное, то нужно будет заранее его, конечно, удалить.

**3. Работа с исходными данными:**

Появляется экранная форма стандартного программного интерфейса универсального системы Эйдос API (Automatic Program Interface), позволяющая вводить данные из внешних источников данных. В этом интерфейсе мы можем вводить числовые и текстовые данные из экселевских файлов.

Показываю вам экранную форму, то есть это исходную таблицу экселевскую, в которой строчки - это наблюдения, а колонки - это способы группировки, классификационные шкалы, и свойства и их значения - описательные шкалы. Все шкалы, классификационные и описательные, могут быть как числовыми, так и текстовыми. Если в шкале, которая вроде как вот является числовой, появится хоть один символ, не связанный с записью чисел, то есть не цифра, не запятая, не плюс и минус, а ещё что-нибудь появится, буква Г, например, там или Д там, или ещё что-нибудь, то тогда, или точка просто появится даже, то это уже будет считаться текстовой шкалой, потому что текстовый формат более общий, чем числовой. То есть все числа являются текстами, но не все тексты являются числами. И если появляется текст, то тип данных переводится в неизменный формат, более универсальный, то есть текстовый.

Значит, мы здесь видим, что у нас классификационные шкалы, вот на жёлтом фоне, это вторая и третья шкала. Вверху мы здесь формулы видим. А описательные шкалы с четвёртой по двенадцатую. Вот формула, позволяющая удобно это узнать.

Область данных. Ставим в клеточку А1 курсор, нажимаем клавиши Ctrl и Shift одновременно и нажимаем End. У нас появляется область данных, выделенная блоком, соответствующая фактическому расположению данных. Если этот блок получается шире, ну, например, я вот здесь вот возьму сейчас символ введу, а потом удалю. Потом нажму Ctrl+Home, при этом у нас курсор переместится в начало. А потом Ctrl+Shift+End, то у нас выделится более широкая область, чем область данных, видите? Система Эйдос эту область будет анализировать. Поскольку там есть пустая строка в конце, то это значит что? Что числовая шкала будет как текстовая рассматриваться, потому что там не число в клеточке. А колоночка М будет рассматриваться как колонка данных, ну, описательная шкала, в которой нет наименований и в которой нет вариабельности значений. Об этом будут выданы соответствующие сообщения.

Значит, если у нас такая ситуация появляется, то надо выделить область данных, чтобы нормализовать эту ситуацию, чтобы не было этой проблемы, создать новую страницу, вставить туда из буфера обмена информацию, а ту страницу, которая была, удалить тот лист. Тогда у нас будет идеальная абсолютно область данных соответствовать тому, что мы под этим подразумеваем, тому, что фактически этим является.

Вот у нас экранная форма, в которой мы задаём тип данных старый Excel 2003 года. Колонки классификационные со второй по третью, описательные с четвёртой по двенадцатую. Все вот эти вот параметры здесь заданы по умолчанию те, которые там были заданы, когда эта работа загружалась. Вот. И нули и пробелы считать отсутствием данных. Нажимаем О'кей.

**4. Формализация предметной области:**

Происходит преобразование экселевской таблицы, которую я только что показывал, во внутренний формат данных, который используется в системе, это язык, это стандарт данных ДБС, ДБФ. И подсчитывается, какое количество каких типов шкал разных типов данных.

Мы здесь видим, что у нас классификационных шкалы две обнаружено с суммарным числом градаций 14. Градациями в текстовых шкалах являются уникальные значения в клеточках, в ячейках. Числовых шкал классификационных не обнаружено. А в описательных шкалах одна обнаружена числовая шкала описательная, где 12 градаций, и восемь текстовых, где 38 градаций. Здесь вот мы можем изменить, вместо 12 поставить пять, например, и тогда надо пересчитать. Вот здесь мы видим размерность матрицы модели. Тогда изменится размерность матрицы модели изменится. Но сейчас давайте всё оставим, как оно было по умолчанию. И выйдем на создание модели.

Что произошло? Созданы классификационные и описательные шкалы и градации, и с их помощью закодированы исходные данные, в результате чего получилась обучающая выборка. Затем всё переиндексировано все базы.

Значит, ребят, вот эта вот операция создания описательных, классификационных и описательных шкал градаций и кодирование исходных данных с их помощью и получение обучающей выборки - эта операция называется формализация предметной области. Здесь вот для студентов рекомендуется, что делать дальше. Сейчас мы это и сделаем. Файл мы уже посмотрели исходных данных, а потом посмотрим сами классификационные описательные шкалы градации, обучающую выборку и будем потом двигаться дальше.

Значит, там упоминается режим 6.4. Режим 6.4 - это режим, в котором всё вот это описано, то, что сейчас я вам рассказываю. Система, по большому счёту, что делает? Выявляет, то есть анализирует данные, находит там причинно-следственные связи и таким образом преобразует данные в информацию. И потом эта информация используется для решения задач управления, принятия решений и преобразуется в знания таким образом. То есть система преобразует данные в информацию, а её в знания. Об этом была предыдущая лекция.

Значит, первый этап вот этот сереньким фоном - системно-когнитивного анализа состоит в том, что мы определяем, что у нас будет рассматриваться в качестве причин, что в качестве их последствий, что в качестве факторов, что в качестве результатов их действия, что в качестве свойств и их значений, а что в качестве обобщающих категорий, к которым относятся объект моделирования.

Значит, вот это этап, который сейчас мы только что выполнили автоматизированно - формализация предметной области. То есть были проанализированы исходные данные, созданы классификационные описательные шкалы и градации. Здесь даже прямо вверху написаны названия баз данных. И потом они были использованы для кодирования исходных данных, и получилась в результате обучающая выборка.

Система Эйдос имеет развитые программные интерфейсы, позволяющие вводить из таблиц числовые текстовые данные, из файлов текстовые данные и из файлов графические данные. Значит, это одна из её, одно из её достоинств системы Эйдос, причём такое довольно существенное достоинство, что она позволяет, то есть имеет развитые средства ввода данных из различных типов внешних источников данных.

**5. Синтез и верификация моделей:**

Теперь следующий этап - это синтез и верификация моделей. Давайте мы этот этап и выполним. Он выполняется в режиме 3.5.

Значит, здесь есть очень интересные моменты, которые я вам отмечу. И многие из них отсутствовали, когда вы 5 лет назад изучали ту версию системы Эйдос, когда была на тот момент, то тогда этих многих из этих моментов и не было.

Во-первых, появилась возможность расчёта, расчётов на графическом процессоре видеокарты. Иногда это ускоряет до 4.000 раз расчёты на некоторых моделях. Значит, ну и ЦПУ тоже осталось.

Теперь, есть возможность только синтез модели осуществить, потом только верификацию. Что это, для чего это сделано? Значит, если в модели, допустим, миллион записей в обучающей выборке, то полная синтез и верификация может затянуться надолго, ну на неделю, например, может, может на месяц. Вот, смотря сколько там у нас шкал, сколько признаков, сколько классов и сколько признаков, то есть значение факторов или значение свойств.

Вот. И тогда такой режим я разработал ещё, когда не было графического процессора, не использовался графический процессор для расчётов. Мы можем осуществить только синтез модели. Синтез модели даже на очень больших выборках, таких вот как миллион объектов обучающей выборки, и даже я делал и 4 млн модели считал. Система позволяет где-то до 12 млн примерно исследовать обучающие выборки. Примерно, в зависимости от того, сколько у нас классов и сколько признаков.

Ну, э-э, значит, э-э, но это длится сутками такие расчёты. Вот. То есть это уже довольно-таки большие объёмы данных, которые, в общем, обрабатываются довольно долго. И вот синтез модели, даже на очень больших выборках, происходит, ну, там за разумное время, там 15-20 минут, там час, например, ну то есть какие-то разумные сроки. Причём сами эти сроки, они здесь прогнозируются в системе, и она описывает, что осталось ещё 27 минут там, допустим. То есть она это прогнозирует и пишет об этом.

А вот верификация - это процесс значительно более трудоёмкий в числительном отношении, потому что каждый объект обучающей выборки копируется в распознаваемую, ну, она вся целиком копируется. А потом цикл по объектам организуется распознаваемой выборки. Вот. Потом для кажд-, каждый объект сравнивается со всеми классами, то есть внутри организуется цикл по классам. А чтобы сравнить объект с классом, для этого организуется цикл по признакам. Вот представьте себе, что если признаков, скажем, 11.000, а классов, скажем, 1.500, а объектов там миллион, ну это, конечно, огромное количество вычислений. Вот. И оно может стать, скажем так, неприемлемо длительно, может этот процесс осуществляться.

Поэтому предлагается такая технология. Сначала только синтез модели осуществляется, я повторяю, это происходит в разумные сроки всегда, даже при очень больших выборках. А потом только верификацию проводим. Но как мы это делаем? Мы копируем какое-то число объектов обучающей выборки в распознаваемую и на них проводим верификацию. Ну, допустим, берём 100 объектов. *[Телефонный звонок, пауза в объяснении]*

Так, ребята, это звонил проректор, прошу простить меня. Не могу же я ему сказать, что я занятие веду. Вот. Так, на работе.

Так, на чём мы остановились? Вот что. Значит, мы можем здесь задать не миллион, у нас же там миллион было обучающей выборки, или 700.000, например. Берём 100 задаём там, или 1.000, например, объектов. Случайным образом они выбираются, очень хороший генератор случайных событий здесь я разработал. Могу рассказать о нём, на основе генератора фон Неймана, но с внутренним источником энтропии. Вот, очень хорошо работает, быстро и очень хорошие даёт результат, никогда не повторяюсь, без циклов. То есть у него нет вообще циклов у этого оператора, генератора случайных событий.

Выбирается 100 объектов выборки. Осуществляется с точностью до сотых долей секунды измерение времени исполнения верификации и строится пропорция, ребята. Слушайте внимательно, какая пропорция строится? Что если у нас 100 объектов выборки выполняла, верификация на 100 объектов выборки осуществлялась... Верификация - это проверка достоверности. Скажем, там 7 минут, то на, а у нас есть 8 часов, вот, 6 8 48, 480 минут, будет X объектов выборки. Вот этот X мы вычисляем, ставим сюда, и у нас система за 8 часов осуществляет верификацию модели на каком-то подмножестве, на каком-то подмножестве обучающей выборки. То есть очень хорошо, удобно сделано для того, чтобы можно было измерить достоверность модели, даже тогда, когда какие-то там заоблачные размерности её и обучающая выборка. То есть это идёт речь о том, что в реальном разумное время это не может быть осуществлено. А вот такой подход - это вполне позволяет сделать.

Теперь, вот здесь вот у нас очень интересный есть показатель. Смотрите, у нас, значит, здесь написано, смотрите, что данная модель составляет 6 десятитысячных процента от того, что, в принципе, может обрабатывать система Эйдос. А у неё есть ограничение 2 ГБ на файл результатов распознавания. И вот мы здесь вот можем задать, если она нам напишет, что там 120% и не может быть выполнен расчёт при таких размерностях модели, с таким числом объектов распознаваемой выборки, тогда она нам рекомендует вот здесь поставить не 100, а какой-то процент, причём она его рекомендует. Ну можно поставить ещё меньше, ну, скажем, 5%, к примеру, можно поставить. Тогда что будет сделано? Будут удалены те результаты распознавания, которые имели наименьшее по модулю уровень сходства. Значит, очень является хорошо обоснованным такой вывод, что это те результаты распознавания, которые, скорее всего, являются ложными. Ну, то есть ложные будут удалены результаты распознавания, а истинные оставлены.

Ну, сейчас мы ничего здесь менять не будем. Вот. Это, кстати, имеет большой смысл, ну, потому что шум удаляется, проще выражаясь так. Вот. То есть какие-то объекты, которые, в которых нарушены закономерности, основные, базовые, которые выявлены. Ну и дальше мы просто переходим на синтез и верификацию моделей. Это проходит на данной модели очень быстро, это маленькая модель, учебная. Вот. Мы видим результат прогнозирования, что закончится в 11:13, она так и закончилась, 14 секунд отошло.

**6. Оценка достоверности моделей:**

Что сейчас было сделано, ребят? Сначала были модели созданы, а потом проверены на достоверность путём решения задачи идентификации. Возникает вопрос: а можно ли было проверять достоверность модели не путём решения задачи идентификации? Вот я сейчас этот блок рассказываю, ромбик. А, скажем, путём решения задачи принятия решений или задачи прогнозирования, например? Можно. Можно проверять достоверность модели путём решения задачи прогнозирования. Но тогда придётся ждать, когда осуществится или не осуществится прогноз, и потом мы тогда сможем оценить достоверность модели. Можно это делать на ретроспективных данных, тоже путём решения задачи прогнозирования. Тогда ждать не придётся ничего. То есть мы сразу будем знать уже, какой результат. Но задача идентификации и прогнозирования, они очень тесно связаны между собой. Это я попозже расскажу. По сути дела, это одна и та же задача математически.

Теперь, решение задачи принятия решений. Можно ли её использовать для проверки достоверности? Можно. Но тогда придётся ждать, когда реализуется это решение или даст свои результаты. То есть вот мы, допустим, выработали какое-то управляющее решение, оказали соответствующее воздействие на объект моделирования и управления в данном случае уже. И потом подождали немножко и посмотрели: перешёл он в заданное целевое состояние или нет? Если перешёл, значит, э-э, модель адекватна, отражает его характеристики, его поведение этого объекта управления под действием факторов. Если нет, то значит нет.

Ну, конечно, всё это не очень удобно. Наиболее удобным вариантом является определение оценки достоверности модели путём решения задачи идентификации. Так и сделано в системе Эйдос. То есть обучающая выборка копируется в распознаваемую, и потом проводятся решения задачи идентификации. И подсчитывается, сколько у нас каких решений получилось, истинных и ложных, и определяется достоверность модели.

Идентифицировать можно не только обучающую выборку, а какую-то новую выборку, которую обычно называют тестовой. И эксперты, которые оценивают достоверность модели, они знают результаты, к каким классам относятся объекты тестовой выборки, к каким обобщающим категориям. А те, кто работает в системе и эту выборку используют как тестовую, распознаваемую, этого не знает. Им выдаются результаты, и эксперты оценивают на этой основе достоверность модели. Так сделано на сайте Kaggle, ребята. На сайте Kaggle именно так вот и сделано, реализовано, как я сейчас сказал.

Но тогда получается вот что. Тогда, значит, если такие можно сделать выводы, значит, и предположения и выводы. Значит, если на основе обучающей выборки выявляются некоторые закономерности моделируемой предметной области, и эти закономерности, они действуют в некоторой более широкой генеральной совокупности, которую представляет обучающая выборка по отношению, к которой оно репрезентативно. И вот если тестовая выборка относится к этой генеральной совокупности, тогда будут хорошие результаты распознавания этой тестовой выборки. А если она выходит за пределы генеральной совокупности, которая представлена обучающей выборкой репрезентативно, то тогда получатся результаты не очень, может быть, хорошие. Вот. И если они получаются не очень хорошие, то может означать две вещи: либо что модель так себе, средней паршивости там, грубо говоря, то есть не особо достоверная, на три с плюсом там, например, или там три с минусом. А может означать, что эта вот обучающая, тестовая выборка, извините, на основе которой проверяется достоверность модели, что она не относится к той генеральной совокупности, по отношению к которой обучающая выборка, на которой создана модель, является репрезентативной. То есть если она выходит за пределы генеральной совокупности, тогда система, в принципе, и не должна э-э, идентифицировать хорошо эти объекты тестовой выборки.

Кстати, никто не знает на самом деле границ генеральной совокупности, что сюда входит. И поэтому такой способ вот оценки достоверности модели является как раз способом определения границ генеральной совокупности. А генеральная совокупность, она может иметь смысл пространственный и временной. В пространстве это может означать переход в другую организацию какую-то, там, другое место, другую страну. И там использование этой модели. А другой город, например, там, ну то есть это территориальные такие аспекты. А временной аспект генеральной совокупности называется, отражается понятием эргодичность. Что это значит? Это периоды времени, в течение которых закономерности предметной области существенно не меняются. Вот это период эргодичности. То есть если обучающая выборка относится к началу периода эргодичности или середине, то, а тестовая выборка относится к этому же периоду эргодичности, то есть когда действуют те же самые закономерности предметной области, которые действовали на момент создания обучающей выборки и моделей на её основе, то будет хороший получаться результат идентификации, прогнозирования. Если же тестовая выборка выходит за пределы периода эргодичности, то тогда её использование, то есть тестирование модели с её помощью, даст не очень хорошие результаты обычно.

Ну что я вам сейчас могу привести вам пример, что я имею в виду? Вот, допустим, Советский Союз, мы создали модели, которые прогнозируют развитие, успешное развитие Советского Союза, так тогда выражались именно таким способом. Прогнозируют успешное развитие Советского Союза. И вот, значит, то есть другого варианта, то есть даже книжки так назывались. То есть даже никому в голову не приходило, что может быть что-то другое там прогнозироваться может. И вот, э-э, Советский Союз распался, и наступил период послесоветский, который примерно с девяносто первого по девяносто восьмой год длился. То есть тогда сохранялась, в общем-то, та система, которая была в Советском Союзе. То есть если взять тестовую выборку, обучающую взять, допустим, восемьдесят седьмой год, а тестовую взять девяносто второй год, то будет, в общем-то, получаться неплохие результаты прогнозирования. Вот. Но если взять, э-э, тестовую выборку с девяносто, девяносто девятый год, после дефолта, то результаты будут совершенно отвратительные прогнозирования. То есть уже советская система рухнула практически там во многом экономическая, и стала формироваться новая система, но она находилась в процессе становления. Если мы, допустим, в девяносто втором году сделаем модель, возьмём обучающую выборку, сделаем модель, э-э, и попробуем её применить в девяносто третьем году, то она тоже даст очень плохие результаты. Почему? Потому что очень высокая динамичность предметной области. Вот период с девяносто восьмого года по 2003 - это практически революционные изменения в экономике происходили, перестройка её совершенно на другие принципы, в соответствии совершенно с другими принципами. Это изменения характера революционных, то есть качественные изменения. В теории процессов, э-э, из той же области, где у нас термин эргодичность, там же применяется термин точка бифуркации. То есть точка бифуркации - это смена одного периода эргодичности другим. То есть это, по сути дела, революционная ситуация. То есть период эргодичности - это период эволюционного развития, а точка бифуркации - это революционное качественное изменение самих закономерностей в предметной области.

Вот. Так вот, где-то примерно к 2003 году, э-э, стабилизировалась экономика уже нового типа, и потом дальше она развивалась уже на новых принципах. Если мы будем вот так вот, э-э, тестировать, э-э, модели, созданные в определённый период, ну, скажем, в девяносто втором году, и, э-э, мы увидим, что достоверность их, э-э, падает, и особенно резко она упадёт в 2003 году, когда уже будет новая совершенно экономическая система. А вот если мы сделаем модель в 2004 году, то она где-то и 2007, 2008 будет работать, и даже 2008, чуть-чуть достоверность понизится. Я просто почему так говорю про это уверенно? Потому что я это делал, проводил такие исследования. Когда вот так называемый мировой кризис был, то чуть-чуть понизилась достоверность модели, но очень несущественно, там на несколько процентов, там 3%, там 5%. Ну, то есть изменения действительно были в мировой экономике, но не такие масштабные, как вот в девяносто восьмом году, к примеру, или в семнадцатом году. Аналогичный вариант был, когда произошла революция и произошло экономика стала перестраиваться на социалистические принципы. И был период, э-э, до НЭПа, когда э, был вообще бардак, грубо говоря, потом, значит, НЭП стабилизировал ситуацию, а потом НЭП прекратился и началась коллективизация и все вот эти вот э-э, такие колоссальные преобразования в обществе, я бы так сказал. Вот. И где-то к войне, примерно ко Второй мировой войне, у нас уже была экономика социалистическая построена. И потом она существовала где-то до э-э, восемьдесят четвёртого года в чистом виде, а потом э-э, что-то с восемьдесят четвёртого года по девяносто первый, что-то вроде НЭПа у нас было, когда была и социалистическая экономика, и э-э, коммунистическая партия, и одновременно при этом были и кооперативы какие-то, частные фирмы стали появляться. То есть какие-то стали появляться рыночные отношения. Это примерно то, что в НЭПе было. И примерно то, что в Китае э-э, изначально, и в Юго-, ещё было нечто похожее в Югославии.

Вот. Так что, что получаем мы в результате? Мы получаем, э-э, после синтеза модели, получаем, э-э, и верификации, сами эти модели. Как они выглядят? Строчки - это значения свойств. Ну здесь какое-то свойство, цвет, например, и через чёрточку его значение. А колоночки - это обобщённые категории, э-э, объектов, э-э, классификация объектов, то есть классы. И единички, то двоечки, троечки стоят в тех местах, ну, где какой-то признак, сколько-то раз встретился у объектов какого-то класса. И классы здесь разной степени общности. Есть более конкретные, более общие классы. Там, где более общие, там больше получается этих признаков.

И в самом низу у нас есть строчка, в которой написано, какое количество объектов обучающей выборки э было представлено по каждому из классов. А справа есть колоночка "сумма", и там суммарное количество встреч какого-то значения свойства, какого-то признака, как я называю коротко это, у всех объектов обучающей выборки, по всей выборке. Значит, ну каждый объект обучающей выборки у нас относится к двум классам, поэтому, там же две классификационных шкалы. Поэтому это число, оно как бы в два раза увеличено. И здесь вот у нас число объектов выборки тоже в два раза увеличено. То есть я для того, чтобы это объяснить на каком-то какой-то вербализации, то есть на каком-то словесном языке, предложил следующие термины: физический объект обучающей выборки и логический. Физический - это конкретная мышка, например, а логический - это, э, с одной стороны мышка, а с другой, обобщённый образ уже мышки компьютерной, а с другой стороны - это элемент компьютера. То есть, э, конкретный образ конкретной мышки физический, он представляет собой суперпозицию двух логических объектов - мышки и элемента компьютера. И вот так она этот и используется.

Значит, на основе такой вот матрицы абсолютных частот нельзя делать корректных выводов, э, о том, насколько характерен или не характерен тот или иной признак для объектов того или иного класса. Почему? Ну, допустим, вот я сейчас найду сейчас. Вот у нас есть признак: размер один под руку. Он встречается три раза у средств связи и три раза у элементов компьютера. Казалось бы, можно предположить, что этот признак одинаково характерен для этих категорий. Однако мы видим, что у нас три средства связи предъявлены в качестве объектов выборки обучающей, и семь элементов компьютеров. То есть ясно, что в процентном отношении, если смотреть, то здесь вот у нас, э-э, размер под руку для средств связи - это 100% объектов этого класса обладают этим признаком, а для элементов компьютеров около 43%.

Вот. Значит, ну ясно, что нужно перейти к относительным величинам, чтобы от этого модель не зависела от того, сколько объектов обучающей выборки представлено по той или иной категории. А для этого нужно перейти от абсолютных величин к относительным. Тогда мы увидим, что 100% объектов обучающей выборки обладают этим признаком размер под руку и 42,8 элементов компьютеров. То есть отсюда мы уже можем сделать вывод о том, насколько этот элемент, этот признак характерен для объектов той или иной категории. Ясно, что для, для средств связи он более характерен, чем для элементов компьютера.

А теперь посмотрим, ребята, по всей выборке. 30 объектов э-э, обучающей выборки логических этим признаком обладают. То есть это что означает? Что этот признак характерен для и, и для элементов компьютеров, и для средств связи. Но для средств связи он более характерен, чем для элементов компьютера.

Значит, такого рода рассуждения, они э уже могут осуществляться людьми на основе обоснованно, на основе матрицы условно безусловных процентных распределений, но только это люди могут это делать только на моделях очень маленькой размерности, когда вот можно обозревать глазами, грубо говоря, эти матрицы с помощью естественного интеллекта анализировать. Я решил так, что надо это дело автоматизировать сравнение вот этих условных и безусловных процентных распределений. И потом этот результат этого автоматизированного сравнения использовать для решения задач. То есть не непосредственно матриц абсолютных частот или условных безусловных процентных распределений использовать, а использовать уже результаты их анализа, а конкретно результат анализа матрицы условно безусловных процентных распределений. Ну и матриц абсолютных частот в некоторых случаях.

Значит, в системе Эйдос реализовано семь способов сравнения условных и безусловных процентных распределений. Эти способы сравнения я называю частными критериями, и они используются для расчёта различных моделей, э, которые я называю системно-когнитивные модели, и они являются моделями представления знаний. Почему знаний? Потому что на их основе можно решать задачи принятия решений и достижения цели. А я на предыдущей лекции вам рассказывал, что данные преобразуются в информацию, когда они осмысливаются. А процесс осмысления данных называется анализ, а смысл заключается в выявлении, знании причинно-следственных зависимостей в этих данных. Вот. А когда мы, э-э, зная эти причинно-следственные зависимости, принимаем решение для достижения цели, то тогда это фактически это значит, что э-э, э-э, знания, то есть информация преобразуется в знания.

И вот у нас есть несколько способов это сделать. Сравнение условных, безусловных процентных распределений. Если P(i|j) - это относительная частота встречи итого признака э-э, в у объектов житого класса, а P(i) - это и относительная, это безусловная относительная частота встречи этого признака по всей выборке, то тогда вот такое отношение позволяет нам посчитать, какое количество информации содержится в итом признаке о том, что объект с этим признаком относится к житому классу. Это мера Харкевича, так называемая.

Это если её представить, э-э, сами эти, э-э, условные, безусловные, э-э, относительные частоты представить в виде, э-э, через абсолютные частоты и через суммы по строке N(i), сумма по колонке N(j) и сумма по всей матрице N, то тогда это будет вот так выглядеть. Здесь я вам скажу очень интересную вещь, что есть два способа сравнения э-э, двух чисел - это вычитание и деление. Если мы из одного числа вычитаем другое, получается больше нуля, тогда значит, первое число больше. Если мы делим одно число на другое и получается у нас число больше единицы, значит, числитель больше. Вот. Значит, и соответственно, мы здесь и видим частные критерии. Допустим, вот критерий ХИ-квадрат - это разность между фактической частотой наблюдения этого признака в житой группе и теоретической частотой наблюдения этого признака в этой группе. И эта разница, это сравнение двух этих чисел путём вычитания. А мера Харкевича, ребята, это сравнение этих чисел путём деления. Если мы N(ij) разделим на теоретическую частоту, то это означает, что мы умножим её на N, N(ij), и разделим на N(i) и на N(j). Мы это и видим в выражении под логарифмом для меры Харкевича. То есть можно утверждать, и я это утверждаю, и очень давно уже, там уже лет 20, впервые я об этом написал в своих работах, что, по сути дела, количество информации по Харкевичу является, э-э, тесно связано с мерой ХИ-квадрат, и, по сути дела, имеет такой же точно смысл. Если мы возьмём, э-э, преобразуем э этот логарифм как разность логарифмов э-э, числителя и знаменателя, то мы можем тогда написать, что это будет логарифм N(ij) минус логарифм, э-э, ну, когда мы ставим минус, то это выражение под логарифмом переворачивается дробь. Э-э, получается минус логарифм N(i) \* N(j) / N. Фактически это то же самое, что ХИ-квадрат, только вместо абсолютных у нас, э-э, выражений, э-э, частоты абсолютной, э-э, фактической и теоретической, используется уже их логарифмы этих частот. Ну, поскольку логарифм является монотонной функцией аргумента, то есть при увеличении аргумента, э-э, само значение функции тоже растёт, э-э, и никогда не падает, а именно вот растёт, э-э, пропорционально, можно сказать так, что значению аргумента, то получается так, что количество информации и мера ХИ-квадрат ведут себя одинаково, сходно при увеличении, значит, этих частот.

Теперь мы посмотрим ещё на одну меру, которая называется коэффициент возврата инвестиций (ROI). Эта мера применяется в экономике, и тоже представляет собой результат сравнения фактической частоты и теоретической путём деления. Вот, но она нормирована к нулю, когда нет никаких зависимостей между итым признаком и принадлежностью объекта к классу житому. Она нормируется к нулю не путём того, что берётся логарифм, а просто вычитается единица. Вот. То есть нормировка осуществляется по-другому. Таким образом мы рассчитываем матрицы количества информации по Харкевичу. Вот. И здесь мы видим и положительные, и зна-, э-э, величины, и отрицательные. Она нормирована, э-э, к битам. Э-э, для этого служит вот этот коэффициент пси, который иногда спрашивают: "А что он делает, этот коэффициент?" Он нормирует к битам, э-э, в случае детерминистском случае. То есть если у нас каждому классу будет соответствовать один признак, и этот признак только этому классу будет соответствовать, то есть матрица модели будет у нас диагональная, ну так, по диагонали будут идти единички, матрица частот, вот, только по диагонали единички, то тогда вот матрица инф1 мы получим количество информации в признаки о классе, которое равно логарифму двоичному от числа классов. Вот этот коэффициент пси вот это и делает. Я его вывел в семьдесят девятом году формулу для него. Сейчас это называется коэффициент эмерджентности Хартли. Я его так назвал. И теперь многие думают, почему-то, что его предложил Хартли. Ну, предложил профессор Луценко, когда был ещё давно, задолго до того, как он стал профессором, в семьдесят девятом году я был старшим инженером на вычислительном центре мединститута, и никаким профессором не был. Вот. Но тогда я вывел эти формулы, и модель прошла рецензирование докторами наук, физмат наук, и сказали, что очень интересно, должно получиться неплохо. Ну вот потом я её развивал эту модель, совершенствовал, э-э, очень много там чего дополнил. Вот всё, что там есть, всё это делается на основе вот этих системно-когнитивных моделей.

И, значит, мы получаем что? Если какой-то, значит, здесь битах количество информации в этой модели инф1. Значит, э-э, если какой-то признак более характерен, то есть, то есть, вернее, встречается в какой-то группе, в классе чаще, чем в среднем, то получаем количество информации положительное. Если же он в этом классе встречается, но реже, чем в среднем, тогда получаем количество информации отрицательное. Ну, например, э-э, что это значит, во-первых, количество информации отрицательное? Это значит, что этот признак не характерен для объектов этой категории. То есть он встречается, но не характерен.

Приведу пример вам. Значит, вот представьте себе, что вот нам нужно отличить студентов от студенток. И у нас есть признак: длинные волосы. Значит, признак длинные волосы очень-очень часто встречается у девушек и очень редко у ребят. Вот. То есть у девушек он встречается намного чаще, чем в среднем, а у ребят намного реже, но встречается. Поэтому, если у кого-то длинные волосы, и мы не знаем, кто это, парень или девушка, то мы получаем очень большое количество положительное количество информации о том, что это девушка, и большое количество информации отрицательной о том, что это парень. То есть это информация о том, что это не парень. Вот, так бы сказал. Вот это смысл этого, э-э, минуса.

А вот, э-э, модель ХИ-квадрат, очень часто она, эта модель, является наилучшей по достоверности. Она чем интересна? Она хорошо сбалансирована. Сумма и среднее равно нулю и по строкам, и по колонкам. Это такая у неё интересная особенность этой модели. Здесь мы видим, кое-где вообще пробелы, а кое-где видим нули. Значит, с чем это связано? Значит, здесь расчёты ведутся семью знаками после запятой в системе Эйдос, в базах данных. А отображение ведётся здесь тремя знаками после запятой. То есть это что означает? Что там где-то в седьмом или шестом знаке, там есть не ноль. Это эти, это связано с точностью расчётов. То есть это погрешность расчётов семью знаками после запятой. Если считать там, скажем, с двадцать тремя знаками после запятой там, то тогда, э-э, будут все пустые клеточки.

**7. Недостатки классической F-меры и ее обобщения:**

Следующий вопрос. Мы модели посмотрели. Следующий вопрос такой: а какую они имеют достоверность эти модели? Это вот мы посчитали эти модели, посмотрели на них. И теперь переходим к вопросу о выборе наиболее достоверной модели. Это делается в режиме 3.4.

Значит, здесь мы видим несколько критериев достоверности, э-э, которые, э-э, хмм, основной - это F-мера Ван Ризбергена. И её обобщение этой меры, которые я предложил, э-э, несколько лет назад. Вообще, я их предложил очень давно, но в статьях об этом не писал. А вот несколько лет назад я написал статьи математические, где описал то, что я там предложил.

Ну сначала давайте посмотрим, что предложил Ван Ризберген. Он предложил считать достоверность модели по результатам решения задачи идентификации, оценивать. Что совершенно разумно. Я вам уже сказал, что можно это делать путём решения задачи идентификации, оценивать достоверность модели, или задачи принятия решений. Ну, задачи прогнозирования и задачи принятия решений. Ну, задача идентификации для этого больше всего подходит. Он это, видимо, прекрасно понимал Ван Ризберген, и именно таким путём и предложил рассчитывать, оценивать достоверность модели.

Значит, он высказался таким образом, что что должна вообще делать модель, когда она решает задачу идентификации? Она должна правильно относить объекты к тем классам, к которым они относятся, правильно не относить к тем классам, к которым они не относятся. И, значит, э, что значит правильно? Значит, она не должна ошибаться ни в первом, ни во втором случае. Но фактически, э, модели отклоняются в какой-то степени от идеальной. Это идеальная модель так должна делать. А реальные модели, они э относят объекты к классам, но иногда ошибаются. Не относят объекты к классам, и тоже иногда ошибаются.

Соответственно, он внёс, предложил ввести такие сумматоры, которые будут подсчитывать количество истинных и ложных, положительных и отрицательных решений. Что такое положительное решение, ребят? Это решение о принадлежности объекта к классу. То есть уровень сходства объекта идентифицируемого с обобщённым образом класса больше нуля. Отрицательное решение, соответственно, это решение о непринадлежности объекта к классу, когда мы видим, что уровень сходства меньше нуля.

И, соответственно, существуют истинные и ложные решения: True, False. True - это истинные, False - ложные. Соответственно, у нас получается четыре сумматора: True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN). Ну, можно без перевода понять, что они означают: истинно положительное, истинно отрицательное, ложноположительное, ложноотрицательное. Истинно положительное решение. Что это такое? Это решение о принадлежности объекта к классу, когда объект действительно относится к этому классу. Истинно отрицательное. Система не относит объект к классу, то есть уровень сходства меньше нуля. И объект действительно не относится к этому классу. Ложноположительное решение. Уровень сходства больше нуля, система относит объект к классу, а фактически он к нему не относится. Ложноположительное решение, ложное срабатывание. Это очень опасная, ребята, ситуация. Представьте себе, что средства ПВО должны охранять какие-то объекты очень большой важности, наземные от воздушных нападений и космических. И, значит, эти средства ПВО обнаруживают какой-то объект приближающийся воздушный. Вот, и идентифицируют его как э-э, объект противника, средства поражения. И включают системы ПВО и уничтожают этот объект. А это был пассажирский самолёт, не противника, просто авиалинии. Такая ситуация была над Крымом, когда, значит, украинские ПВО проводили учения в Крыму, и они сбили самолёт, который летел из Израиля в еврейскую автономную область, Ту-154. Ну и все, естественно, все погибли. Ракета С, комплекс С-200, средства ПВО советские ещё разработки. И аналогичный случай недавно был в Турции, когда взлетал лайнер с аэродрома пассажирский, на взлёте, и пролетел мимо, рядышком пролетел какой-то объект, ну, предполагаю, что это был израильский истребитель или что-то типа такого. То есть прямо рядом пролетел с этим лайнером. И средства ПВО, которые были сначала сориентированы на этот истребитель, перенаправились на воздушный лайнер этот пассажирский, потому что он имел гораздо большую отметку, и сшибли его, и все погибли. Был большой скандал, министр обороны извинялся там, э, ну, в общем, это было спровоцировано. И аналогичная ситуация была на базе российской Хмеймим в Сирии, когда самолёт заходил на посадку радиолокационной разведки Ил-18, какой-то там Б или что-то такое с каким-то индексом. Вот. И пролетел перед ним израильский истребитель, перед этим самолётом. А средства ПВО его вели сирийские, этот истребитель. И они перенаправились на, это известно всем, что на боль-, на цель с большей отметкой перенаправление происходит. Я считаю, что, может быть, это даже и неправильно. То есть, не-, возможно, что даже это и не нужно так было делать. Вот. Захватили цель, ведите её, зачем вы там перенаправляться на другие цели? Вот, ведите и ведите, и сбивайте её. Вот. А произошло перенаправление на самолёт радиолокационной разведки, и он был сбит средствами ПВО Сирии, дружественными, грубо говоря, силами. Вот. Из-за чего? Ну, я считаю, что из-за некомпетентности. Все эти случаи можно связать с некомпетентностью персонала средств ПВО и на базе Хмеймим, и на базе, и в Турции, и когда-то в Крыму произошло. Вот. Тоже всё это можно связать с некомпетентностью. Ну, хотя ясно, что была провокация. То есть это, так сказать, была проверка на степени компетентности. И проверка показала невысокую компетентность персонала. То есть, э-э, как вражеская цель был идентифицирован, э-э, гражданский объект, и он был уничтожен, или объект не вражеский был идентифицирован как вражеский, был уничтожен. Это грубая ошибка, называется, э-э, ложная идентификация, ложноположительное решение. Имеет очень такие нехорошие последствия эта ошибка. И в других областях тоже. Ну, я не буду дальше приводить примеры, но просто могу сказать, что во всех областях это приводит к очень нехорошим последствиям. И ложноотрицательные решения. Что это такое? Система не относит объект к какому-то классу, то есть уровень сходства меньше нуля, а фактически он к нему относится. Понимаете? Тоже ложные решения такого рода тоже могут очень иметь печальные последствия во всех областях, и в медицине, например, там. И вообще ложные решения очень опасные, я вам скажу.

Вот. Значит, э-э, ну, мера рассчитана на основе этих, э-э, значений этих сумматоров. По Лизбергену рассчитывается полнота модели, вот, точность модели, и с их использованием рассчитывается F-мера.

**7. Недостатки классической F-меры и ее обобщения (продолжение):**

Всё это чудесно, как говорится. Но есть некоторые но. Какие именно?

Значит, сейчас я вам скажу. Вот статья моя, посвящённая, э-э, ну она называется, смотрите, "Инвариантное относительно объёмов данных, нечёткое мультиклассовое обобщение F-меры достоверности модели Ван Ризбергена". Ну, из названия статьи уже ясно, что, значит, сначала я рассматриваю в этой статье просто классическую меру Ван Ризбергена, э-э, потом традиционное описание в математических обозначениях. Дело в том, что вот эти вот сумматоры, которые мы сейчас рассматривали с вами, ТП, ТН, ФП, ФН (True Positive, True Negative...), это не математические обозначения. Такие иногда используются в экономике и программистами широко используются. То есть это, это название переменных, которые подсчитываются соответствующие суммы. А в математике приняты другие обозначения. То есть такие вот двойные ТП, например, это может означать две переменные: Т умножить на P, например, время умножить на давление, например, или на вес там, что-нибудь такое, понимаете?

Вот. Поэтому для того, чтобы дальше потом работать уже с этой мерой Ван Ризбергена, э-э, математически уже обобщать её математически, для этого нужно преобразовать её в математическую форму. Вот это вот я и сделал в разделе 1-2. А потом описал, какие существуют проблемы с мерой, с F-мерой Ван Ризбергена. Это первое - проблема мультиклассовости, второе - проблема нечёткости, и третье - зависимость этой F-меры от объёма данных. Ведь у нас же здесь, вот смотрите, сумматоры. Ясное дело, что когда мы проводим какую-либо идентификацию, любую абсолютно, ребята, абсолютно любую идентификацию, то мы получаем либо истинно положительное, либо истинно отрицательное, либо ложноположительное, либо ложноотрицательное решение. Это что значит? Что у нас в любом случае к одному из сумматоров прибавится единичка. Это значит, что сумма значений всех этих сумматоров э-э, является, э-э, равна объёму выборки, равна объёму распознаваемой выборки.

Вот. И, следовательно, при увеличении объёма выборки эти сумматоры будут линейно расти. Я когда про это рассказываю, рисую такую картинку обычно в последнее время. Вот. Это у нас система координат. Вот. Здесь у нас объём выборки. Вот. А здесь у нас количество решений.

И совершенно понятно, что количество решений пропорционально объёму выборки. Но решения бывают разные. Они бывают истинно отрицательные, истинно положительные, ложноотрицательные и ложно положительные.

Ребята, из моего опыта вытекает, он очень большой этот опыт решения различных задач, что счётчик, сумматор истинно положительных решений... а нет, истинно отрицательных решений, True Negative, растёт быстрее всего. То есть модели лучше, легче всего, э-э, совершают истинно отрицательные решения. То есть это такое решение, которое легче всего, то есть чаще всего является правильным. Истинно отрицательное решение. Их число растёт быстрее всего при увеличении объёма выборки. На втором месте находятся истинно положительные решения. Это когда модель хорошая. Здесь вот нужно сказать, когда модель хорошая, достоверная. Если модель плохая, тогда это может быть не истинно положительное решение, а другие, ложноположительные. Потом растут, э-э, э-э, ложноотрица-, то есть ложноположительные решения. Вот. И медленнее всего растут ложноотрицательные решения, False Negative.

Вот такая вот, э-э, картиночка. Значит, э-э, у нас такой чат, что я картиночку вам не могу послать в чат, но вы можете посмотреть, что при увеличении объёма выборки все эти сумматоры изменяют свои значения, но они могут, наклон может быть разный этих кривулек. Может быть побольше наклон, поменьше наклон, различные может быть наклон. Вот. Ну это вот я вам нарисовал примерно, как для хорошей модели это выглядит.

Поэтому возникает необходимость проверки, насколько модель, F-мера Ван Ризбергена зависит от объёма выборки. Я эту, эту проверку провёл и оказалось, что она зависит и очень нехорошо зависит.

Вот. Значит, соответственно, я предложил, э-э, обобщение меры Ван Ризбергена мультиклассовое. То есть, да, у Ван Ризбергена предполагается, что объект может относиться только к одному классу. А вот в системе Эйдос это не так, и в жизни тоже это не так. Объект может относиться к очень большому числу классов. Взять, допустим, дерево какое-нибудь. Это дерево относится к классу "дерево", во-первых, а во-вторых, относится к классу... *[Телефонный звонок, пауза]* ...к классу растений, к классу живых существ, к классу объектов природы и так далее.

Вот. Ну и, э-э, соответственно, предложил мультиклассовое обобщение меры Ван Ризбергена. Второе. Э-э, Ризберген, он очень логично, убедительно всё сделал, но, э-э, всё-таки как-то вот он парень такой крутой какой-то, прямо совсем прямо крутой. Вот смотрите, ребята, вам формулу-то видно, нет? Отвечайте мне микрофоном. Вам видно вам эту форму или нет? Да. Видно. Что вы так молчаливые какие-то совсем?

Вот смотрите, ребята, видите, вот тут есть ложноположительное решение: клавиатура два, то есть второй экземпляр клавиатуры, э-э, похож оказался на сумку и на аксессуар на 4%. Если нормировать, значит, вот это 100% к единице, тогда это 4 сотых. Это ложноположительное решение. Действительно, это неправильное решение. Правильные здесь птичками отмечены. Но она, система ошиблась. Но она ошиблась-то всего на 4%, то есть, э-э, Ризберген суммирует в данном случае к сумматору False Positive единичку. Я говорю: "Ну это как-то вообще, ну, жестоко просто, понимаете?" Дело в том, что система Эйдос - это ж не просто система идентификации, это система, которая ещё оценивает достоверность этой идентификации. То есть она принимает решение определённое о том, что объект относится к данному классу, и определяет степень его принадлежности к данному классу, степень принадлежности, как в нечёткой логике. И ещё она определяет степень уверенности своей, своей системы, в этом решении. То есть она относит его к этому классу, но она пишет, что я-то вообще-то очень сомневаюсь, потому что этот объект к этому классу относится. То есть уверенность очень низкая, 4%, ребята, из 100.

Я говорю, ну как-то суммировать единицу, когда система сама сообщает, что она очень сомневается, э-э, что степень уверенности всего 4% - это, по-моему, ну я бы сказал так, жестоко, понимаете, немножко. Это наказывает систему очень сильно и модель тем самым. Я предложил нечёткое обобщение меры Ван Ризбергена. У него там белое, чёрное, истина, ложь. И всё. А система Эйдос - нечёткая система. У неё она показывает степень принадлежности объекта к классу. Поэтому я предлагаю суммировать не единичку, э-э, к сумматору, э-э, соответствующему, вот в данном случае к сумматору False Positive, а уровень уверенности системы или степень принадлежности объекта к этому классу. Вот система сейчас считает, что он относится на 4 сотых к этому классу. Значит, надо суммировать не единичку к False Positive сумматору, а 4 сотых надо суммировать. Понятно, нет, ребят?

Слушайте, да, понятно, или нет, непонятно? Потому что я, честно сказать, вот когда заставляю вас эти плюсики ставить, то я ещё смотрю, кто из вас вообще реагирует на мои указания. И вижу, что реагирует половина примерно. А остальные, они там подключились, и не знаю, чем они там занимаются. Вот давайте сейчас, пожалуйста, все поставьте, кто сейчас меня слышит, все поставьте плюсики в чате. Быстренько давайте, раз, раз, раз, раз, раз. Вот Сергей даже два поставил плюсика. Это ты за кого? За себя, за того парня поставил?

Так, раз, два, три, четыре. А сколько у нас участников? Восемь. То есть у нас сейчас на связи восемь участников собрания, а плюсика четыре. Ну, если считать, что Сергей два поставил, тогда пять можно считать. Но я думаю, что он просто там у него рука дрожит, наверное, он переживает, когда я спрашиваю. И он так вот дрогнул у него рука, и получилось два плюсика. Да, Сергей? Вот. То есть фактически у нас половина только участников отметили, что меня сейчас слышат.

Ну вот, видите, Сергей, молодец. Ну и я тоже молодец, потому что я догадался. Вот. А теперь смотрите, ребят, э-э, то, что я вам сказал, что эта мера зависит от объёма выборки. Тоже это требует своего решения. Я предложил соответствующее решение математическое. Вот это, как это выглядит. Всё это, значит, тут это. А потом численный пример привёл. Вот этот численный пример я хочу вам показать. Э-э, как зависят эти критерии достоверности от объёма выборки. Жёлтенький - это классическая мера F-мера Ван Ризбергена. Она вот так вот, э-э, хмм, сходится к какому-то пределу при объёме выборки где-то 2.500 примерно. Нечёткое мультиклассовое обобщение, вот это сиреневое, она чуть повыше идёт там, примерно на 1/10, если единицы нормировать максимальную достоверность. И тоже она где-то на 2.500 стабилизируется объёме выборки. А инвариантная относительно объёма выборки мера, э-э, обобщения, тоже она мультиклассовая, нечёткая, она где-то стабилизируется, когда около 500 у нас объектов, э-э, распознаваемой выборки. При этом, смотрите, пожалуйста, внимательно. Значит, э-э, классическая мера Ван Ризбергена крайне неустойчива при малом объёме выборки. Крайне неустойчива. То есть она сильно колеблется, понимаете? То есть она колеблется где-то на 2/10 примерно. Вот. И получается очень заниженная оценка достоверности. Вот верхний максимум, он является продолжением тренда вот этого вот, который потом дальше идёт уже, э-э, более-менее стабильно, неспадающая кривая. А минимумы они вообще выпадают, видите, куда? На 2/10 ниже. Да? Похоже на сходя-, на затухающие колебания. Мера мультиклассовая, нечёткая, она гораздо стабильнее ведёт себя. Просто гораздо стабильнее. И ещё тем более стабильно ведёт себя инвариантная относительно объёма выборки мера. Можно какие выводы сделать из этого? Что меру Ван Ризбергена при выборках меньше 500 вообще нельзя использовать. Она просто даёт какие-то результаты, похожие на случайные, колебания очень сильные, понимаете? Вот. При, значит, а вот эти меры, которые я предложил, э-э, инвариантную, нечёткую меру и, значит, вернее так, нечёткую мультиклассовую меру и ещё к тому же и инвариантную относительно объёма выборки, их можно использовать. Они себя очень стабильно ведут на малых объёмах выборки. Это значит, что система Эйдос хорошо работает на малых объёмах выборки, ребята. И это было экспериментально зафиксировано, когда её сравнивали с нейронными сетями, оказалось, что у неё в пять раз меньше нужен объём выборки для того, чтобы её обучить, чем нейронных сетях. То есть система Эйдос намного быстрее модель сходится у неё к разумному результату, чем в нейронных сетях. То есть она, скажем так, быстро обучаемая система. На небольших объёмах выборки она уже даёт разумные результаты, а нейронные сети, то надо в пять раз больше примерно объём выборки, чтобы они начали давать разумные результаты. Это, кстати, очень интересный результат, он получен в Перми, в Пермском университете. Они исследовали эту систему математически, сравнивали с другими. И оказалось, что она имеет интересные такие достоинства эта модель.

А я когда эту модель разрабатывал, я читал книжечки по статистике. В частности, у меня есть книжечка по, э-э, малым выборкам. Статистика объектов на малых выборках. Вот книжечка, ребята. Видно, нет? Вот, малая выборка. Книжечка советская, книжечка. Вы будете в шоке, ребята. Но эта книжечка семьдесят восьмого года. Видно, нет, что она семьдесят восьмого года?

Вот такие дела. Надо читать книжечки.

Вот. И мы видим, что модели, которые, то есть меры достоверности, которые я предложил, они дают более высокие оценки достоверности, чем классическая мера Ризбергена. Почему? Потому что ошибки, ребята, есть в моделях, есть ложные решения, и положительные, и отрицательные. Но есть очень замечательная закономерность: ошибочные решения всегда с очень низким уровнем сходства, а истинные решения всегда с высоким уровнем сходства. Поэтому получается, что, если суммировать, если нечёткое обобщение использовать меры F-меры Ван Ризбергена, то получается, что эти ошибки, они как бы меньше сказываются, а истинные решения, они сильно сказываются на достоверности, потому что у них высокий уровень сходства. Понимаете, о чём я говорю, нет? Получается, что они дают более адекватную оценку достоверности, чем мера F Ризбергена. И эта адекватная оценка, она выше, чем мера Ризбергена. У нечёткого мультиклассового обобщения она выше на 1/10 при единице максимум. А у инвариантной меры, э-э, нечёткой мультиклассовой инвариантной меры обобщения меры F Ризбергена, у неё, сейчас скажу тут, э-э, где-то примерно на 3/10 выше достоверность. Здесь 7/10, а здесь э 3 с половиной. То есть даже больше, чем на 3/10, 3 с половиной, 35 соток где-то. Ну, грубо говоря, на 4/10 выше достоверность. А это очень много, это половина почти что шкалы. Понимаете? То есть это уже не шутки, это уже серьёзно.

Вот.

**8. Взаимосвязь задач анализа данных (повторное упоминание и детализация):**

*Этот раздел был затронут ранее, здесь идет повторение и уточнение связей между задачами.*

**9. Заключение и анонс:**

Так. Что там дальше? Да, понятно. Да, примерно так и было. Вопросов нет.

Ну что ж, мне удобнее, когда вы микрофоном пользуетесь. А почему? Потому что я иногда показываю экран там, демонстрирую, да, когда рассказываю. И я просто не вижу этого чата. Мне надо переключаться, чтобы там увидеть, что вы там пишете.

Ну хорошо. Тогда сейчас небольшой такой анонс, который я уже на самом деле сделал. На следующем занятии, ребята, мы рассмотрим решение задач идентификации, прогнозирования, принятия решений и исследования моделируемой предметной области. То есть если вы до конца дослушаете следующее занятие, то вы будете полностью готовы провести автоматизированный системно-когнитивный анализ э в любой предметной области практически, но той, где исходные данные описаны в виде, приведены в виде экселевской таблицы или CSV файла, ну экселевской таблицы, э, и там могут быть и текстовые, и числовые данные. А потом на следующих работах мы рассмотрим, как обрабатываются тексты, интеллектуальная обработка текстов в системе Эйдос. Вот. И ещё у нас третья работа будет - обработка изображений, обобщённый спектральный анализ. То есть формирование обобщённых спектров классов, конкретных спектров, сравнение конкретных спектров с обобщёнными, сравнение обобщённых спектров друг с другом и так далее, и так далее. То есть, в общем, обработка изображений и автоматизированный системно-когнитивный спектральный анализ.

Вот такое дело. То есть я всё это вам покажу, расскажу. Потом начнём разрабатывать собственное приложение, описывать его, размещать его и .

Вот такие дела. Ну теперь, похоже, что действительно конец занятия у нас.

Всё самого-самого хорошего, ребят. До свидания.

До свидания.
До свидания.