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

***Российская Федерация***

**274 Лабораторная работа №1 (3.03). По дисциплине "Интеллектуальные информационные системы". 2021-03-10**

## Создание и верификация моделей в системе Эйдос: Лабораторная работа 3.03

### Резюме

Лекция посвящена выполнению первой лабораторной работы (3.03) в интеллектуальной информационной системе Эйдос-Х++. Занятие проводит профессор Луценко Е.В. для группы заочного обучения ПИЗ 1822 10 марта 2021 года.

**1. Введение и установка лабораторной работы:**
Лекция начинается с обзора предыдущего занятия, где рассматривалась установка системы Эйдос и типы лабораторных работ. Основная тема – первая лабораторная работа (3.03), включающая установку и изучение ее компонентов. Процесс установки включает запуск системы и настройку параметров: указание диапазонов классификационных и описательных шкал градаций, типов исходных данных и интервалов для числовых шкал.

**2. Типы шкал и формализация данных:**
Объясняются различные типы шкал, используемые в системе: описательные, классификационные, текстовые (номинальные, порядковые) и числовые. В ходе установки лабораторной работы система автоматически формирует шкалы и градации на основе файла исходных данных, кодирует эти данные и создает базу событий и обучающую выборку. Этот процесс называется формализацией предметной области.

**3. Анализ исходных данных и структуры модели:**
Демонстрируется файл исходных данных (пример с предметами на столе студента), объясняется его структура: описательные шкалы (свойства объектов) и классификационные шкалы (обобщающие категории/классы). Подчеркивается важность корректности данных и возможность наличия ошибок при вводе. Первая колонка файла не является шкалой, а содержит описание объектов.

**4. Синтез и верификация моделей:**
Описывается основной рабочий процесс системы Эйдос: преобразование исходных данных в информацию (путем создания справочников и кодирования), а затем в знания (путем выявления причинно-следственных связей и построения моделей).

* **Синтез моделей (Режим 3.5):** Создаются 3 статистические модели (матрицы частот, распределений) и 7 системно-когнитивных моделей (модели знаний). Обсуждается возможность расчетов на ЦПУ или ГПУ (Nvidia), причем ГПУ значительно быстрее. Для больших данных предусмотрены опции, такие как Bootstrap (использование части выборки) и раздельный запуск синтеза и верификации.
* **Анализ моделей (Режим 5.5):** Рассматривается матрица абсолютных частот и матрица условных вероятностей (относительных частот). Объясняется, как сравнивать условные и безусловные частоты для выявления характерности признаков для классов, используя различные меры (отношение, разность, мера Харкевича, критерий хи-квадрат, коэффициент ROI).
* **Верификация моделей (Режим 7):** Достоверность моделей оценивается путем решения задачи идентификации (распознавания) объектов обучающей выборки. Используется F-мера Ван Рисбергена, основанная на точности и полноте, которые вычисляются из числа истинно положительных (TP), ложноположительных (FP), истинно отрицательных (TN) и ложноотрицательных (FN) решений. Объясняется, почему идентификация предпочтительнее прогнозирования или принятия решений для оценки достоверности.

**5. Естественный интеллект и моделирование:**
Проводится аналогия между работой системы Эйдос и процессом познания человека. Человек также использует внутренние (субъективные) модели для узнавания (идентификации) объектов, формирующиеся на основе опыта. Однако форма представления этих субъективных моделей науке неизвестна. Система Эйдос предлагает формализованный подход к созданию и использованию подобных моделей на основе данных.

**Заключение:**
Подчеркивается важность понимания всего процесса работы с данными и моделями в системе Эйдос для успешного выполнения лабораторных и контрольных работ.

### Детальная расшифровка текста

**Раздел 1: Введение и установка лабораторной работы**

**Подраздел 1.1: Приветствие и организационные моменты**

Здравствуйте, ребята.
Здравствуйте.
Здравствуйте.
Здравствуйте.
Сегодня 10 марта 2021 года.
И у нас третья пара.
Дисциплина интеллектуальные информационные системы с группой заочной ПИЗ 1822.
Занятие ведет профессор Луценко Евгений Вениаминович.
Тема занятия: первая лабораторная работа, которая называется 3.03.

**Подраздел 1.2: Обзор предыдущего занятия и цели текущего**

На предыдущем занятии мы увидели, как устанавливается система Эйдос, какие там есть типы лабораторных работ. И мы начали изучать, как устанавливать первую лабораторную работу и прямо ее и, собственно говоря, изучать, эту первую лабораторную работу.

**Подраздел 1.3: Запуск системы Эйдос и установка ЛР 3.03**

Вот. Для того, чтобы это сделать, мы запускаем систему.
Да.
Такое подозрение, что этот WebEx, он не дает через интернет работать системе.
Тяжело получается как-то.
Вот. И, как я просил напомнить мне...
Просил напомнить, на чем мы остановились. Да? Ксения, давай напоминай, на чем мы остановились на предыдущей работе.
Мы устанавливали лабораторную работу третьего типа. Нажали Окей. Указали диапазоны классификационных и описательных шкал градаций, тип исходных данных, тип интервалов для числовых шкал.

**Раздел 2: Типы шкал и формализация данных**

**Подраздел 2.1: Объяснение типов шкал**

Я рассказал вам про то, какие бывают типы шкал: описательные, классификационные, текстовые, числовые. Текстовые: номинальные и порядковые.

**Подраздел 2.2: Процесс формализации при установке ЛР**

Вот. И теперь мы нажимаем Окей. Работа эта устанавливается. Здесь обнаружено, что у нас есть описательная шкала числовая, поэтому вопрос возникает о числе диапазонов. Всё мы по умолчанию делаем, как здесь у нас задано. Но если мы хотим задать другое число диапазонов, тогда нужно после этого пересчитать шкалы и градации, тогда у нас будет размерность модели уже не 14 на 50 строк, не 14 колонок на 50 строк, а 14 на 43 строки. Но сейчас мы будем делать всё по умолчанию.
Идет создание модели. И сейчас, обратите внимание, какие выполнены работы.
Какие выполнены работы системой Эйдос?
Сформированы классификационные и описательные шкалы и градации на основе файла исходных данных. А потом исходные данные закодированы с помощью этих классификационных и описательных шкал и градаций, которые как справочники используются. И получена в результате база событий и обучающая выборка. И потом всё это переиндексировано, чтобы можно было работать. Можно было об этом и не писать, конечно.

**Раздел 3: Анализ исходных данных и структуры модели**

**Подраздел 3.1: Доступ к файлу исходных данных**

Вот. Теперь смотрим на сам файл исходных данных, который находится по пути, который здесь указан. На ваших компьютерах там он и будет находиться. У вас будет указан путь именно на ваших компьютерах на этот файл. Заходим мы в папочку исходных данных и открываем файл исходных данных.

**Подраздел 3.2: Структура файла исходных данных**

Вот, видим этот файл. Что мы видим? Что у нас есть описательные шкалы, в которых описаны свойства объектов, и классификационные шкалы, в которых (выделены желтым фоном для удобства), в которых приведены обобщающие группы, категории, к которым относятся объекты.

**Подраздел 3.3: Происхождение и особенности данных**

Объекты здесь использовались для формирования модели те, которые были на столе у студентов, когда они делали лабораторную работу. Студенты вводили эту информацию. Кое-какую информацию ввели некорректно. Сейчас я на это покажу вам, какая информация некорректно введена. Но, значит, я не исправляю это, чтобы объяснять вам, показывать, что там как.
Значит, у нас здесь все шкалы текстового типа, кроме одной шкалы – размер два, которую я добавил для того, чтобы показать, как обрабатываются шкалы числовые.

**Подраздел 3.4: Описание колонок файла**

Первая колонка шкалой не является, а просто содержит описание объектов обучающей выборки. Это описание может быть довольно длинным, там до 60 символов или даже больше. Сейчас я не помню, но длинным может быть наименование. Чуть ли не 256.
С ошибками орфографическими написано. Ну что, эпоха ЕГЭ, ЕГЭ, понятно, деградация происходит всей системы образования. Уже на федеральных площадках, совещаниях, конференциях, посвященных цифровизации общества, пишут, что цифровизация через "Ы". Ну, значит, уже, видимо, она так вот у нас и будет через "Ы". На федеральном уровне.
Значит, этот файл мы посмотрели.

**Раздел 4: Синтез и верификация моделей**

**Подраздел 4.1: Обзор этапов работы в Эйдос**

И смотрим теперь сами классификационно-описательные шкалы и градации, обучающую выборку и двигаемся дальше. Значит, путь, которым мы двигаемся, описан в помощи в режиме 1.3. Ввести исходные данные, посмотреть 2.1, 2.2, 2.3.1. Посмотреть сами модели, создать модели, посмотреть модели, посмотреть достоверность. Вот этот путь мы и проходим. Об этом говорится также вот здесь, в режиме 6.4. Есть Help, в котором описывается, что делает система. Она исходные данные преобразует в информацию, информацию в знания. Поскольку информация – это осмысленные данные, а смысл – это знание причинно-следственных зависимостей, то для того, чтобы преобразовать данные в информацию, создаются справочники прошлых и будущих событий или справочники факторов и классов. Вот. Потом исходные данные кодируются с их помощью, получается нормализованная база исходных данных, то есть обучающая выборка. И таким образом у нас уже появляются, кроме базы исходных данных, ещё много баз данных. То есть сама база исходных данных, база справочников классификационных шкал и градаций, база справочников описательных шкал и градаций, база обучающей выборки. А в каждую базу входят по две, по три таблицы, ребята, связанных отношением один ко многим. То есть получается уже довольно-таки много различных баз, ну около семи там, дополнительно.
А потом, чтобы уже преобразовать исходные данные в информацию, нужно выявить причинно-следственные связи между событиями, которые отражены в обучающей выборке. И вот эти причинно-следственные связи отражаются в форме семи моделей.

**Подраздел 4.2: Формализация предметной области**

Этап формирования классификационных и описательных шкал и градаций и обучающей выборки называется формализация предметной области. В системе Эйдос есть режим, который обеспечивает это. Видите, вот формализация предметной области. И демонстрацию также результатов формализации предметной области. По сути дела, это подготовка исходных данных к обработке в системе.

**Подраздел 4.3: Синтез и верификация моделей (детально)**

А дальше выполняется режим синтеза и верификации моделей. Как это делается? Сначала создаются модели все, рассчитываются, а потом проверяются на достоверность. И выбирается наиболее достоверная модель. И если она есть, то в ней решаются задачи идентификации, прогнозирования, принятия решений и исследования моделируемой предметной области. А если достоверной модели нет, то никакие задачи не решаются. И происходит выход из системы и переход, наверное, на начало процесса постановки задачи, то есть когнитивно-целевой структуризации предметной области, где мы определяем, что у нас будет в качестве факторов рассматриваться, а что в качестве результатов их влияния. И потом заново мы всё это делаем, но уже на других данных, возможно, расширенных, больше описательных шкал, больше классификационных шкал, больше примеров.

**Подраздел 4.4: Режим 3.5 - Синтез моделей**

И вот здесь вот на этапе синтеза модели создаются три модели статистические – это база абсолютных частот, матрица абсолютных частот, и две матрицы условных и безусловных процентных распределений. А потом на их основе рассчитывается семь моделей, которые я называю системно-когнитивные модели, представляющие собой модели знаний.
Вот такой порядок наших действий. И хотя эти задачи я в виде отдельных блоков показал, но в развитой форме они все между собой, как скажем, пересекаются, взаимосвязаны.
Вот. Значит, сейчас мы выполняем режим 3.5. Переходим в соответствующие пункт меню. И видим, что здесь у нас три модели статистических создаются, семь системно-когнитивных. И мы можем работать, ребята, либо на центральном процессоре (ЦПУ), либо на графическом процессоре расчёты вести.

**Подраздел 4.5: Расчеты на ЦПУ vs ГПУ**

Ну, значит, надо вам сказать, ребята, что на графическом процессоре...
Ага, ну вот мы сейчас как раз и смотрим. Значит, модели...
Что-то я отвлекся и...
Создаются таким образом модели. И потом, да, вот что я хотел сказать про графический процессор и центральный процессор. На некоторых компьютерах расчёты на графическом процессоре не проходят. Это связано с тем, что видеокарта не стандарта Nvidia, не на чипсете Nvidia. Значит, если видеокарта на чипсете Nvidia или карты самой фирмы Nvidia или GeForce, то всё работает на графическом процессоре. Язык OpenGL стандартный используется. Расчёты тогда проходят в тысячи раз быстрее, чем на центральном процессоре. Не в два, не в три, а в тысячи раз, ребята, быстрее. Я фиксировал до 4.000 раз быстрее расчёты. Ну это зависит, правда, от того, сколько объектов обучающей выборки, сколько классов, сколько признаков. От многих параметров модели это зависит, насколько будет ускорение происходить.
Поэтому, значит, ставим на графический процессор, запускаем расчёт. Но если возникает ошибка, то надо прервать в диспетчере задач исполнение системы Эйдос, перейти сразу же опять, запустить её сразу же в режим 3.5 и задать центральный процессор, и дальше уже работать.

**Подраздел 4.6: Обработка больших данных (Bootstrap, раздельный синтез/верификация)**

Теперь, здесь есть несколько разных средств для того, чтобы обрабатывать большие данные. Большие данные. Значит, если у нас сотни тысяч объектов обучающей выборки, сотни классов, десятки тысяч признаков, и я сказал, ну, скажем, 700.000 объектов обучающей выборки. То для того, чтобы проверить достоверность модели, каждый объект выборки идентифицируется с каждым из 500 классов, а для каждой идентификации он складывается количество информации, там, скажем, в 10.000 признаков. Это огромный объём вычислений. На графическом процессоре это происходит довольно быстро, но всё ещё ж выходные формы рассчитываются. Это занимает всё время. Поэтому здесь сделано так, что можно только синтез модели осуществить, только верификацию модели, или можно синтез и верификацию. Для небольших моделей – и то, и то сразу же. Если же размерность модели большая, тогда осуществляется только синтез модели. Это проходит довольно быстро, даже когда 700.000 там примеров объектов выборки, к примеру. Ну, может там полчаса там или час, ну такое как бы приемлемое время проходит расчёт. А вот верификация, мы если запустим, она может неделю продолжаться при таких объёмах выборки и таком числе классов и признаков. Тогда здесь предусмотрен такой режим, называется бустрепный подход, бустреп – складной нож по-шведски, как в литературе говорят. В чём суть этого подхода? Что не вся обучающая выборка используется для верификации модели, для оценки её достоверности, а только часть. И эту часть можно использовать для создания модели, а можно и вообще не использовать, только для верификации использовать. Всё здесь это предусмотрено. Значит, как предполагается проверять достоверность модели при очень больших объёмах исходных данных, большие данные, Big Data? Ну, допустим, здесь вот я задаю 100 объектов обучающей выборки, выбранных случайным образом. Запускаю распознавание, верификацию. Процесс этот проходит, и система измеряет время этого процесса, которое на него было затрачено. И потом мы строим пропорцию: 100 объектов выборки, скажем, 8 минут использовались, то есть верификация шла 8 минут на таком числе объектов выборки. А у нас неизвестно, сколько объектов задать (X), но так, чтобы верификация шла, скажем, 8 часов, 480 минут. И находим этот X, получается, скажем, там 1.300. Запускаем вечером и идём спать себе спокойно. Утром приходим, у нас уже посчитана достоверность модели на максимальном количестве объектов выборки, которое можно посчитать за такое время. Вот для чего это сделано. Но у нас случай такой простой, что нам это не надо.

**Подраздел 4.7: Ограничения размера модели и опция прореживания**

Теперь, вот здесь внизу указывается, что вот эта модель, которую сейчас мы рассматриваем, она составляет 6 десятитысячных процента от того, что может обрабатывать система Эйдос. То есть вы можете понять, что она обрабатывать позволяет гораздо большие размерности модели. Но иногда напишет нам здесь вот, что модель в полтора раза превосходит максимальные возможности системы, скажем. Тогда что делать? Тогда, если мы нажимаем Окей, просто напишет, что не может быть выполнен расчёт с такими параметрами модели, с таким числом объектов, классов, признаков. Тогда вот здесь вот мы можем задать 5%, к примеру, или 10. Что это такое? Значит, в результате, значит, здесь есть узкое место – это база результатов распознавания, которая наибольший размер имеет из всех баз, когда создаётся модель. И вот она именно является ограничением, и размер этой базы данных, он не должен быть больше 2 ГБ по в связи с тем, какое, какие ограничения в языке программирования на размеры баз данных. Когда я проектировал систему, то я не думал, что может быть такого размера база распознавания, результатов распознавания. Вот. Ну это бывает очень редко, но бывает. Тогда мы вот здесь вот можем задать 5%, к примеру, или 10. И тогда из результатов распознавания будут оставлены только 5% наиболее значимых, с наиболее высоким уровнем достоверности. Сейчас мы это делать не будем. И система позволит обработать и данные, и ещё большего размера. При этом будет удалён именно шум, то есть те результаты, которые не представляют интереса. А критерий достоверности, там есть в системе очень хороший, убедительный и критерий достоверности результатов распознавания и прогнозирования, апробированный, хорошо, очень хорошо себя показавший критерий, там есть достоверный, адекватный критерий достоверности.

**Подраздел 4.8: Запуск синтеза и верификации**

Значит, нажимаем здесь вот Окей. Начинается синтез моделей. И после того, как этот синтез моделей завершён, начинается верификация модели.

**Раздел 5: Верификация моделей и F-мера Ван Рисбергена**

**Подраздел 5.1: Задачи для проверки достоверности**

Верификация модели, проверка её достоверности, может быть выполнена путём решения различных задач в этой модели и сравнения результатов решения этих задач с фактом. Какие есть типы задач, которые можно использовать для измерения достоверности? Прежде всего, это распознавание, идентификация, классификация – это одно и то же всё, диагностика. Вторая задача – это прогнозирование, третья – принятие решений.

**Подраздел 5.2: Ограничения использования принятия решений и прогнозирования для верификации**

Принятие решений неудобно использовать для проверки достоверности модели, потому что представьте себе, мы не знаем, какова достоверность модели. Рекомендуем те или иные решения для достижения цели, ну, например, для перевода фирмы в более благоприятное состояние, более далёкое от банкротства. И рекомендуем эти решения применить. Фирма их применяет, и получается положительный результат. Действительно, она отдаляется от состояния банкротства. Тогда мы можем сделать вывод, что модель была достоверной. Я бы это не стал делать таким способом проверять достоверность модели. Почему? А потому что, может быть, она является недостоверной. Может же такое быть? Может. Тогда в этом случае наши рекомендации, они будут давать не тот результат, который хотелось бы. Вот мы, допустим, хотели бы отдалить её от состояния банкротства, а может быть, она как раз в результате этих решений наших, исполнения этих решений обанкротится. То есть это очень рискованно и авантюристично предлагать какие-то рекомендации на основе модели, достоверность которой неизвестна, и потом проверять, была она достоверна или нет, потому какой результат. Ну какой-то такой кровожадный способ очень, понимаете? Вот. То есть жестокий. Ну я считаю, что принимать решения на основе модели, достоверность которой неизвестна, вообще нельзя. Если же это делается, то это, я бы сказал, является профанацией вообще науки о принятии решений, управлении.
Прогнозирование. Для этого мы можем взять, допустим, данные ретроспективные, чтобы проверить достоверность модели путём прогнозирования. Ну, скажем, за январь взять данные и спрогнозировать ситуацию на февраль. Если получается, а февраль уже прошёл, сейчас уже март. И мы в состоянии определить, хорошо был прогноз сделан на основе январских данных на февраль или плохо. Если он был сделан хорошо, то есть прогноз оказался достоверным, значит, модель достоверная. Это не так опасно, как принятие решений. Почему? Потому что можно использовать ретроспективные данные, данные прошедшего периода. Но, однако, всё-таки это не самый лучший вариант. Почему? Потому что всё-таки прогнозирование – это задача, ну, как скажем, более сложная, чем задача идентификации. Потому что при задаче идентификации свойства и принадлежность объектов к классам, объектов к классам относятся к одному моменту времени, а при прогнозировании действие факторов относится к прошлому, а состояние объектов, в которое он переходит под действием этих факторов, к будущему. То есть эта задача логически более сложная.

**Подраздел 5.3: Идентификация как основной метод верификации**

Поэтому для проверки достоверности модели наиболее подходящим, на мой взгляд, и не только мой взгляд, а и других учёных-разработчиков, является решение задачи идентификации. Вот. И если задача идентификации решается достоверно, то можно надеяться, что и прогнозирование, и принятие решений тоже будет адекватным.

**Подраздел 5.4: Результаты идентификации и F-мера**

Нажимаем здесь Окей. И смотрим на что? Когда у нас, я вам показывал порядок действий. Вот мы модели создали в режиме 3.5. Теперь надо посмотреть, а что ж мы там создали, что за модели? А потом уже посмотреть на их достоверность, пункт семь. Вам видно, да, экранные формы, ребята? Здесь может быть много приложений, и система работает с тем, на котором стоит W. Можно поменять. Здесь убрать, а здесь поставить на другом приложении. Вот.
Видно. Видно. Ну хорошо.
Смотрим на модели. Это делается в режиме 5.5, как я уже и показывал и в графическом хелпе, и вот в хелпе режима 1.3. Формируется, как вы поняли, несколько моделей: модели статистические и системно-когнитивные. Смотрим на простейшую статистическую модель – матрицу абсолютных частот. Что мы здесь видим? Мы видим колонки – это градации классификационных шкал. У нас их две таких шкалы – это конкретный класс и обобщающий класс. Конкретный класс и обобщающий класс. То есть один и тот же самый, один и тот же принцип группировки, только разная степень общности. А может быть, могут быть использованы различные способы группировки. Ну, например, эффективность в натуральном выражении: объём продукции, качество продукции. И, скажем, потребительские свойства, технические свойства. И каждое свойство в виде колоночки. Их может быть много таких свойств потребительских или технических. А потом идут колоночки, например, эффективность в стоимостном выражении. Там может быть прибыль, рентабельность, вот, выручка, убытки там, допустим, что там может быть там, самые разные показатели финансово-экономические, плановые. И эти показатели могут быть интегральные по всей фирме, а могут быть по подразделениям, по полям. Например, поле, там на нём выращивалась в этом сезоне пшеница, такой-то сорт. Какова рентабельность этого поля оказалась, понимаете? То есть может быть конкретно прямо. Какова рентабельность этого сорта? Вот на этом поле рядом с такими же условиями выращивались различные сорта. И там вот рентабельность выше, а там ниже. Там потребительские свойства выше, а там ниже. Там урожайность выше, а там ниже. И качество там и так далее, и так далее. Почему? А вот потому, что там сорт другой и потому что, возможно, другие способы были возделывания, выращивания. А может быть, там и условия несколько отличаются. Другой угол там наклона поверхности по отношению к солнцу, там склон какой-то там, плоский или наклонённый на юг, на север там и так далее. То есть всё это вот может иметь большое значение. Кроме того, что там росло в прошлом году, позапрошлом, и в течение последних 10 лет. То есть вы видите, что можно оценивать очень много различных параметров результатов деятельности в натуральном и стоимостном выражении. А здесь у нас могут быть различные факторы, которые влияют на эти результаты, которых может быть тоже очень много. Это, допустим, если мы будем продолжать пример с агротехнологиями, то это способы вспашки, параметры... А удобрений там десятки видов, и каждый из этих видов можно вносить разными способами, в разных дозах, на разных фазах и так далее, и так далее. Здесь получается тысячи, ребята, факторов, которые могут влиять на результаты, и сотни результатов влияния различных, в натуральном и стоимостном выражении. А также можно и в социальном, допустим, выражении. И вот мы видим здесь, что у нас на этой нашей маленькой выборке, мы видим здесь вот у нас классы, градации классификационной шкалы первой, градации классификационной шкалы второй, обобщающие классы вот. И здесь у нас свойства и значение этого свойства. Цвет свойства, материал. То есть это описательная шкала, её градация. А здесь классификационная шкала и её градация. И не одна шкала, а все шкалы классификационные одна за одной. А здесь все описательные шкалы одна за одной. У нас есть шкалы, как вы видели, текстовые, номинальные и порядковые, а также есть числовая шкала, которая в таком виде вот выглядят эти диапазоны.
И здесь мы видим количество встреч такого признака у объектов такого-то класса. А здесь всего мы видим, сколько объектов такого относящихся к этому классу, предъявлено в обучающей выборке. И мы видим, что вот по этому классу у нас семь объектов предъявлено, по этому три, по этому четыре. И вот, допустим, мы сейчас возьмём, поднимемся вот сюда, размер под руку. И обнаружим, что у нас оказывается этот признак обнаружен три раза у средств связи и три раза у элементов компьютеров. Видите, да? Можно ли считать, что это означает, что этот признак одинаково характерен для объектов этих категорий? Этого нельзя выводы сделать. Этот вывод некорректным является, что он одинаково характерен. Почему? Потому что средств связи у нас три, а элементов компьютера семь. То есть нам нужно перейти к относительным величинам, к процентам, например. И тогда уже можно будет сравнивать с учётом того, что у нас разное число объектов обучающей выборки по разным классам. И вот мы смотрим размер под руку, вправо. Видим, у средств связи 100% этим признаком обладают, а у элементов компьютеров 42%. То есть он менее характерен этот признак для элементов компьютеров, чем для средств связи. Но всё-таки он для них тоже более характерен, чем в среднем для всех объектов выборки. Потому что в среднем для всех объектов выборки около 30%, 30% обладают, то есть этих объектов выборки обладают этим свойством размер под руку. А у элементов компьютера 42%, а у средств связи 100. Это что означает? Что мы, сравнив эти величины друг с другом, условные относительные частоты в разных категориях. Условно – это означает относительная частота при условии, что объект относится к этой категории. Мы видим, что более характерно для средств связи, чем для элементов компьютеров. Но и для элементов компьютеров тоже это более характерно, чем для всех объектов выборки в среднем. И видим это количественно, насколько это различие большое. Как мы можем сравнить два числа, ребята? Вот мы их сравниваем: 42 и 100. Какое больше? 100 больше. А почему? Как мы это определили? Или вот 42 и 30. Какое число больше? Как мы это сравниваем? Есть два способа, ребята, основных таких нормальных и очень много ненормальных способов, бесконечное количество. Ну нормальные способы какие? Вычесть одно число из другого и посмотреть, какой знак разности. Или взять отношение этих чисел и посмотреть, больше единицы или меньше единицы, и насколько. И всё. Вот и все способы, какие можно использовать.
И я вам скажу вот что, что если мы будем этот это сравнение осуществлять с помощью своего естественного интеллекта и естественного зрения, вот просто будем смотреть на эти числа и сравнивать их друг с другом и с безусловной вероятностью встречи признака по всей выборке, то даже на маленькой модели нам это довольно быстро может надоесть. То есть это дело такое неблагодарное. То есть это довольно-таки трудоёмко и не творческая работа, скажем так. Поэтому я это автоматизировал. Я сравнил программно эти все условные, безусловные относительные частоты. Для этого использовал разные способы, которые вот здесь вот описаны. Вот эти способы. Значит, если у нас относительная частота встречи итого признака в житой группе B и житое, а по всей выборке P итое, то их отношение показывает, насколько этот признак характерен для объектов этой группы по сравнению с его характерностью для всей выборки. А если мы возьмём логарифм, то получим тогда нормированную величину. В случае, когда вероятность встретить признак в группе такая же, как и в среднем, то этот признак ничего не несёт, никакой информации не несёт о принадлежности или непринадлежности к группе. И логарифм как раз позволяет это нормировать с нулём в данном случае, если эта информация там отсутствует. Это называется мера Харкевича. Я вам на занятиях буду, может быть, рассказывать на лекционных, что Харкевич предложил, по сути дела, меру знаний, потому что он показал, предложил формулу, насколько изменяется вероятность достижения цели при условии сообщения информации некоторой объекту управления, скажем так. А вот это критерий хи-квадрат. Это разница между фактической частотой наблюдения этого признака в житой группе и теоретической. И именно разность, видите? А если мы возьмём не разность, а отношение, я вам говорил, что можно сравнивать путём вычитания, а можно путём деления. Тогда у нас получится вот эта величина, смотрите, N и Т житое, делённое на теоретическую частоту. То есть эта дробь переворачивается, получается N вверху, а N и Т житое внизу. N итое – это сумма по итой строке, а G – сумма по житой колонке. То есть мы видим, что мера хи-квадрат теснейшим образом связана с количеством информации по Харкевичу. Практически это вообще-то одно и то же, один и тот же смысл имеет, о чём и говорят Нейман и Пирсон говорили в своей лемме Неймана-Пирсона о том, что логично или корректно и обоснованно считать, что объект принадлежит тому классу, о принадлежности к которому в его системе признаков наибольшее количество информации. То есть они говорили об информации. При этом они использовали меру хи-квадрат. Но я могу сказать, что это, конечно, было для красного словца сказано информация. То есть в данном случае они применили этот термин как бы как бытовой термин, как бытовое слово, а не как научный термин. Но если вот сейчас вот мы видим на эти смотрим на эти формулы, то мы видим, что у них были все основания так считать, что это действительно количество информации. Но они об этом не написали, потому что тогда ещё в то время, когда они об этом писали, в начале XX века, ещё не было Александра Харкевича, академика Академии наук СССР, ещё не было предложено его мера вот эта вот, прагматическая мера целесообразности информации или мера знаний, ещё она не была известна. Но сейчас она известна, и мы видим, что она очень тесно связана с критерием Пирсона хи-квадрат, наличием взаимосвязи между признаками и принадлежностью объекта к классу. Это один одно слагаемое хи-квадрат, конечно, для одной клеточки таблицы. И есть также в экономике коэффициент возврата инвестиций, ROI называется, Return on Investment. Как он вычисляется? Тоже делится условная относительная частота на безусловную. И чтобы нормировать к нулю, не логарифм берётся от этого отношения, а вычитается единица. То есть есть разные способы нормировать к нулю, да? Вычесть единицу можно. А можно вычесть не единицу, а пять. Можно. Но тогда потом прибавить четыре надо ещё. А можно 1005 вычесть. Можно, но тогда надо прибавить 1004 потом. То есть это как раз то есть те самые вот ненормальные способы нормировки, которые я сказал. И сравнения тоже также можно аналогично выдумывать. И вот мы видим, что здесь каждая модель, формула повторяется два раза, но разный смысл переменных, которые туда входят. И сейчас мы можем посмотреть. Значит, если мы посмотрим на матрицу абсолютных частот, то здесь в самом низу мы видим, что у нас есть две строки суммарных. Суммарное число признаков у объектов какого-то класса, какой-то категории, и суммарное число объектов, предъявленных по этой относящихся к этой категории, предъявленных в обучающей выборке. Ясное дело, что чем больше объектов предъявлено, тем больше признаков будет у этих объектов суммарно наблюдаться. То есть эти две строки, сумма числа признаков и сумма числа объектов, они, ясное дело, между собой коррелируют тесно, взаимосвязаны. Но всё-таки они отличаются. И у нас, соответственно, получается два способа вычисления условной относительной частоты путём деления числа встреч признаков у объектов какого-то класса на суммарное число признаков у объектов этого класса или на суммарное число объектов этого класса. Получается два способа вычисления. И вот здесь вот мы видим эти два способа вычисления в каждой формуле повторяется два раза с разными смыслами этих вот переменных, которые входят в эту формулу. Вот. И вот мы видим результат этого сравнения. Смотрите, значит, если мы посмотрим на количество информации, то мы видим, что у нас есть и положительное количество информации, и отрицательное. Что означает отрицательное количество информации? Это значит, что этот признак менее характерен для объектов этого класса, чем для всей выборки в среднем. Ну я могу вам привести такой пример. Вот, допустим, у нас две группы: девушки, студенты и студентки. Вот. Для студенток очень характерен признак длинные волосы. То есть их вероятность их встретить в группе девушки значительно выше, чем в среднем по всей выборке. Почему? Потому что в среднем по всей выборке есть ещё ребята, у которых, как правило, короткие волосы. А есть ребята, у которых длинные волосы? Есть. Есть. Но очень мало. В среднем намного, то есть в группе э студентов признак длинные волосы встречаются гораздо реже, чем в среднем. Ну, скажем, в среднем он около 50%, будем считать, что девушек и ребят примерно поровну. И вот у длинные волосы в группе девушки встречаются у всех девушек практически. А, значит, вероятность встретить длинные волосы в группе девушки в два раза выше, чем в среднем. А в группе мальчики встречается 1%, к примеру. Это получается вероятность встретить этот признак в группе студентов намного меньше, чем в среднем. И если мы возьмём логарифм, получится величина отрицательная. Точно так же получится отрицательно, если мы вычтем единицу. То есть что это означает? Это означает, что обнаружение этого признака несёт информацию о непринадлежности объекта с этим признаком к этому классу. А о принадлежности. Помните, у нас под руку 100% было у средств связи и только 40%, там 42% элементов компьютера. Поэтому получаем в видах больше информации о том, что это средство связи и меньше о том, что это элемент компьютера. А по некоторым признакам получаем даже информацию количество отрицательное. Вот. И смотрим меру хи-квадрат. Эта мера гениальная, я вам скажу так, на мой взгляд, потому что она прекрасно сбалансирована. Сумма равна нулю по колонке и по и по колонке равна нулю сумма, и по строке. И среднее тоже равно нулю. Вы представляете себе? Насколько она хорошо у нас сбалансирована. На я могу вам сказать, что у меня довольно большой опыт расчёта различных моделей и применения. Модель хи-квадрат очень хорошо себя показывает. То есть всегда, как правило, это одна из лучших моделей. И вот смотрим, что у нас здесь получается. Размер под руку. Средств связи намного чаще фактическое наблюдение этого признака, чем теоретическая частота. И, соответственно, большое значение слагаемого хи-квадрат. А у элементов компьютера тоже больше фактическое, чем теоретическое, но не настолько, не в такой степени, не в два раза больше, то есть не 100%, а всего лишь 42%. Получается хи-квадрат значение поменьше. А у некоторых признаков, то есть для некоторых классов оно значение отрицательное имеет, значение этого свойства. То есть не характерность, а не характерность. И мы смотрим, я сказал, что сумма нулю равна, средняя нулю. А здесь мы видим иногда пустое место, иногда видим ноль, да? Что это значит? Это значит, что здесь на самом деле не ноль, а какое-то маленькое число, которое не помещается в разрядные сетки. Ну, например, там одна там стотысячная, например, там или там миллионная. Тогда это не ноль, хотя и близко к нулю величина, но это не ноль. Здесь отображается тогда это число, но все разряды, которые в формате отображаются, они равны нулю. Вот. Но ясно, что там какой-то есть, скажем так, погрешность расчётов, вот так я бы сказал.

**Подраздел 5.5: Вероятность vs Относительная частота**

Теперь, различие нужно, ребята, обязательно сказать о различии между относительной частотой и вероятностью. Вероятность – это предел, к которому стремится относительная частота при неограниченном увеличении объёма испытаний, числа испытаний. Кубик кинули, выпала четвёрка. Вероятность выпадения четвёрки – 1/6, относительная частота равна единице, потому что одно испытание было, и одна и выпала. А остальных граней – нулю. Кинули ещё раз, выпала пятёрка. Относительная частота выпадения четвёрки стала 0,5, пятёрки тоже 0,5, а вероятность – 1/6. Если мы будем кидать вот так вот 100 лет, например, этот кубик, то у нас асимптотически будет приближаться относительная частота к вероятности. При 400 испытаниях, 480 испытаниях, по-моему, точнее, получается при равномерном распределении, получается погрешность около 5%, различие между относительной частотой и вероятностью. А при 1200 испытаниях получается 2,5%. Ну, дальше там, чем мы больше число испытаний, тем медленнее приближается относительная частота к вероятности, асимптотически. И никогда к ней не приближается, никогда не становится ей равной. Но на самом деле она при реальном числе испытаний, если сделать программу имитационную, которая это делает, разыгрывает равномерное распределение, будем считать, что кубик без просверленных там дырочек внутри там, что он симметричный и равновероятно выпадение всех сторон, то можно сказать так, что относительная частота то выше, то ниже, чем вероятность, чем теоретическая вот эта вероятность 1/6. И, в принципе, может быть и равной. Но потом при следующем испытании она опять становится либо больше, либо меньше. Вот эта вот погрешность, отклонение относительной частоты от вероятности, оно постоянно уменьшается при увеличении числа испытаний. Ну я вам скажу так, что при большой выборке можно считать, что это и есть вероятность, потому что само слово вероятность означает нечто абстрактное, математическое понятие абстрактное, ну наподобие бесконечно малых, материальной точки, геометрической точки там или линии – это всё абстрактные математические понятия. Реально в жизни мы не встречаем ни математической точки, ни линии. Вот, но те объекты, которые мы реально встречаем, они могут быть довольно похожими на эти объекты математические абстракции. Почему эти математические абстракции, по сути дела, эффективны? Потому что они отражают реальность. Хотя мы и не встречаем прямо вот математических точек в реальности и так далее. Вот. Понятно, да? То есть, по сути дела, мы всегда имеем дело с конечными величинами, а не с математическими абстракциями. Поэтому мы можем вполне и говорить о том, что это вероятность, понимая при этом, что фактически это не вероятность, а относительная частота, которая приближается к вероятности, и всегда есть некая погрешность в наших выражениях. Ну я вам могу сказать так, что у нас всегда есть некая погрешность в наших выражениях. То есть мы всегда немножко отклоняемся от истины. Вот. Ну мы должны это понимать, знать, контролировать, и тогда это, в принципе, и приемлемо. Если мы не считаем, что это вот истина последней инстанции, абсолютная истина.

**Раздел 6: Естественный интеллект и моделирование**

**Подраздел 6.1: Идентификация объектов человеком**

Теперь мы выполняем следующий этап. Мы модели посмотрели, ребята. Выполняем следующий этап. Вот по той схеме преобразования данных в информацию, её в знания, которые в режиме 6.4 мы смотрели. Мы исследуем достоверность моделей. Достоверность моделей в системе Эйдос оценивается с помощью стандартного классического критерия, который применяется большинством исследователей для этих целей. Это F-мера Ван Рисбергена. На чём основана эта мера? На том, что решается задача идентификации. Я вам уже говорил, что можно различными задачами проверять достоверность модели. Но задача идентификации, она как бы, ну, наиболее подходит для этого, скажем, по общему мнению. И что это значит? Это значит, что в соответствии с моделью мы должны правильно отнести объекты к тем категориям, обобщающим классам, к которым они относятся, и правильно не отнести к тем, к которым они не относятся, и не допустить ошибок ни в первом, ни во втором случае.

**Подраздел 6.2: Модели в естественном интеллекте**

Теперь я хотел бы вас спросить, ребята, вот скажите, пожалуйста, вот вы смотрите сейчас, допустим, на стул, да, перед собой, и видите там мышку, ну, в смысле, компьютерную. Вот. И вы получаете информацию перцептивную, чувственную, результат чувственного восприятия с помощью органов восприятия. И зрительный контакт у вас есть с этой мышкой, и тактильный контакт. И вы идентифицируете этот объект как мышку. Вот я хочу вас спросить, как вы считаете, вы это делаете на основе какой-то модели, саму задачу идентификации решаете, или просто вы берёте и узнаёте, что это мышка? Как решается задача узнавания объектов? Вот мы смотрим вокруг себя и узнаём, какие объекты мы видим. И в том числе и себя самих тоже узнаём, если всё в порядке с головой, то мы смотрим в зеркало, скажем, и понимаем, что это мы. Не какой-то там придурок на нас смотрит там, или какая-то красавица. Вот как мы это понимаем? На основе модели какой-то или просто вот мы понимаем, что это тот или иной объект? Как вы считаете? Ответьте мне на вопрос. Кто ответит на вопрос, идёт на самоэкзамен. Что вы какие-то пассивные? Вы наверное подсоединились и сами там где-то там вообще чай пьёте, там, телевизор смотрите, да? Гешов, как ты считаешь?
На основе модели всё-таки мы понимаем.
Вот. А что за модель?
На работе не получается телевизор смотреть. Да ну ладно, не получается. Взяли там, включили на телефончике трансляцию, звук потише сделали, наушники одели и смотришь себе. Положил его там на коленке и смотришь, а сам делаешь вид, что ты работаешь. Ну, ребят, хорошо, да, я с вами согласен, что это всегда делается на основе модели. А вот когда человек сам узнаёт эти объекты?
А! То есть я у вас вместо кино, короче, да? Вместо телевизора. Понятно.
Ну тогда ещё один телефончик рядышком положите, там можно смотреть кино. Так вот, ребят, смотрите, как получается. Мы решаем задачу узнавания, то есть идентификации, с помощью модели, с помощью своего естественного интеллекта. А вопрос возникает такой: а что это за модель? Откуда у нас взялась эта модель? Вот почему я, допустим, узнаю, где мужчины, где женщины, где молодые, где старые, где дом, где дерево, где собака, где кошка, где машина? Как я это, как получилось, что у меня возникла модель, которая отражает характеристики этих объектов? Когда она возникла, каким образом? Какие есть соображения по этому поводу?
Думаю, она возникла, потому что мы с ней сталкивались много раз. То есть, ну как, например, выглядит мышка, мы уже сразу знаем.
Ну, я вам скажу, да, конечно, так, в общем так. Но дело в том, что вот эти все модели, они у нас формируются с раннего детства. Вот когда там кладут ребёночка в кроватку и вешают погремушки. Зачем это делают? Знаете, погремушки зачем вешают? Когда ребёнок совсем маленький, ну несколько дней, там, скажем так, то он эти погремушки видит как пятнышки такие цветные. Он на них резкость не наводит, не фокусирует зрение на них. Вот. Но они там дребежжат, звенят, двигаются перед ним. Он их видит в размазанном таком виде, примерно как вот в воде, например, если глаза открыть под водой. Вот такое, размазанные такие пятна цветные. Но эти пятнышки, они звучат ещё, понимаете? В принципе, они на таком расстоянии находятся, что можно до них дотянуться, тронуть их. И вот он хватает ручками, машет, машет, хватает ручкой за эту игрушку. Она у него вызывает определённые тактильные ощущения, поверхность там гладкая там, допустим, да? Цветовые ощущения и звук, звуковые ощущения. И раз, и фокусирует на них зрение. И у него формируется образ у ребёнка вот этой игрушки, зрительный образ. И точно так же потом приносят мама там мячик ему, там другие игрушки, куколку там, ещё машинку. И говорит: "Это вот мячик". Первый в жизни мячик он увидел. Потом он увидел второй в жизни мячик, а он оказался не таким, как первый, понимаете? Он другого размера, другого цвета, другой материал. Ну, скажем, первый был резиновый, половина синяя, половина красная, а второй был такой сиреневый с разливами разных цветов, там зеленоватый, сиреневатый, светлый, белый, пластмассовый. Вот. То есть тоже это мячик. Вот мама сказала, что это мячик. Это слово мячик, оно означает некоторый род, обобщающую категорию объектов. И когда мы увидели эти два этих мячика, то у нас уже возник обобщающий образ. У ребёнка возникает обобщающий образ. Потом, когда он начнёт уже говорить что-то, можно его спросить, что наиболее характерно для мячика? Он скажет: "Они все кругленькие, маленькие, пустые, ими можно играть". Понимаете? Как он это определил? На основе большого числа примеров. А если его спросить, он красный или там синий, там, или фиолетовый, там, розовый? Он скажет: "Да это вот может быть любого цвета вообще-то. Вот те, которые я видел, они самых разных цветов бывают". Понимаете? То есть он сразу же определит, что некоторые свойства всегда наблюдаются у мячиков, а некоторые иногда наблюдаются. То есть некоторые являются характерными для этого рода объектов, другие нет. То есть у него возникает модель, в которой отражены характерные, нехарактерные свойства объектов различных видов. Вот мы вышли на улицу, там собачка идёт на верёвочке с тётей, да? Тётю ведёт. Вот. Мама говорит: "Собачка". Вот. А потом кошка там сидит. Вот это кошка. Вот. А иногда подшучивают: "Это кошка или собака?" Вот бежит собачка, говорит: "Это кошка". А ребёночек говорит: "Нет, это не кошка, это собачка". То есть у него уже сформирован обобщённый образ, и он уже идентифицирует конкретные объекты с этими обобщёнными образами.

**Подраздел 6.3: Сложность и эффективность естественной идентификации**

Теперь, значит, он решает довольно сложную задачу на самом деле. Задача идентификации – это сложная задача, это интеллектуальная задача. Это определить по совокупности признаков объекта, к какой категории он относится. Мы эту задачу будем решать с помощью естественного интеллекта, это одна из основных задач, задача идентификации. Но я могу сказать, что для того, чтобы её решать, нужно, по крайней мере, иметь эти обобщённые образы. Вот мы хотим идентифицировать объект с какими-то категориями, с классами. Так надо же иметь образы этих классов. Если у нас их нет, то не с чем и идентифицировать, понимаете? То есть какие-то у нас должны быть образы. И возможно, я не исключаю, что у нас тех образов, к которым относятся объекты, нету. Вот, допустим, там НЛО пролетело, а у нас образа НЛО нету. Но у нас есть только образ тарелки, тазика. Мы говорим: "О, тарелка пролетела". Почему? Потому что вот это, что пролетело, похоже на тарелку, понимаете? Но образ тарелки. Но мы при этом прекрасно понимаем, что это никакая не тарелка, правильно же? А что это такое? А мы не знаем, у нас нет обобщённого образа, к которому этот объект относится. Не сформирован он у нас, понимаете? Мы его никогда в жизни не видели и больше не увидим. Поняли, да, о чём я говорю? То есть мы можем идентифицировать объекты только с теми образами, которые у нас созданы в нашей биографии, скажем так, в нашей жизни предыдущей. Не в предыдущей реинкарнации, а той, которая вот в жизни, которая вот до этого момента прошла. И у нас постоянно формируются эти образы и уточняются, и формируются новые. И сейчас это тоже продолжается этот процесс, он всю жизнь продолжается. Уточнение и формирование новых образов. Потому что мы всё новые и новые объекты видим, и у нас эти образы обобщаются, уточняются, адаптируются к тем объектам, к которым эти объекты относятся. Ну, например, когда я увидел на Евровидении Кончиту Вурст с бородкой, помните, да, это омерзительную эту? Вот, то у меня образ девушки немножко скорректировался, что я понял, что и такие тоже могут быть. Хотя я уже не знаю, что она там такое, это Кончита. Сейчас я даже затрудняюсь понять, что. Вот. Ну, в общем, вы поняли, о чём я говорю. То есть когда мы что-то видим такое совершенно необычное, то оно может изменять те образы, которые у нас до этого были сформированы. Точь-в-точь там вот эти передачи эти. Вы ж поняли, о чём я говорю.

**Подраздел 6.4: Неизвестность формы представления субъективных моделей**

Так вот, теперь вопрос на следующий. Значит, ясно, что у нас есть модель, в которой сформированы обобщающие образы тех категорий объектов, которые мы встречаем в жизни, и в быту, и в учёбе, и в науке, они у нас сформированы. А какая форма представления этой модели, ребят? Какая модель у нас, в которой отражены эти образы? Что это за модель? Как её можно назвать? Это дифференциальные, интегральные уравнения или системы уравнений, уравнения в частных производных, или что это такое вообще? Какие это у нас модели там?
Математические?
Ну вот я очень сомневаюсь, что они математические, понимаете? Ну, в общем, я вам скажу так, если ответить на этот вопрос на уровне современной науки, скажем так, то форма представления и тип этих моделей субъективных, которые мы используем для ориентации в реальном мире и в субъективной сфере, и в объективной, форма представления этих моделей, форма представления знаний в этих моделях, она неизвестна науке. То есть неизвестно, что это за модели, понимаете? Ну я бы сказал так, это субъективные модели очень низкой степени формализации. То есть они не выражены в словах и не выражены в каких-то математических выражениях, в числах, уравнениях. А в чём они там выражены, никто не знает. Вот мы как просто как каким-то образом, непонятным, создали эти модели и пользуемся ими. И я вам могу сказать, ребята, что вообще-то довольно эффективно пользуемся. То есть они довольно-таки неплохо отражают действительность. Если бы эти модели плохо отражали действительность, знаете, что бы с нами стало? Мы бы не смогли в этой действительности существовать. То есть, ну, я бы сказал так, выживать не смогли бы.

**Подраздел 6.5: Сравнение с возможностями ИИ**

Так вот, что предложил Ван Рисберген? Он предложил вычислять достоверность модели по результатам решения задачи идентификации. То есть узнавания объектов. То есть система должна в соответствии с моделью правильно относить объекты к тем классам, тем обобщающим категориям, к которым они относятся, и правильно не относить к тем категориям, к которым они не относятся, и не допускать ошибок ни в первом, ни во втором случае. Как это определить количественно? Он предложил сделать четыре сумматора, обозначить их, причём двухбуквенное обозначение у них. Это не характерно для математики обозначения, а больше это применяется в экономике, может быть, в психологии, ну в экономике чаще всего. То есть переменная обозначается не буквой с индексами, а двумя буквами, например. Это обычная практика в программировании. То есть похоже, что этот Ван Рисберген программист, наверное. Потому что такой стиль обозначения переменных – это совершенно естественным является для программиста. И он обозначил четыре таких вот сумматора, в которых подсчитывается количество истинных и ложных, положительных и отрицательных решений. Ну сначала нужно дать определение, чтобы эти понятия, может быть, вы имеете, но когда есть определение, то как-то легче ориентироваться. Значит, положительное решение. Что такое положительное решение? Это решение о принадлежности объекта к классу, когда уровень сходства объекта с классом, с обобщённым образом класса, больше нуля. То есть он положительный. Вот когда уровень сходства положительный, то это положительное решение. А решение о непринадлежности объекта к классу. Вот, например, берём мы, сейчас вот я вам приведу пример с вами, со студентами. Вот я беру, например, Алексея. И могу вам сказать, что конкретный образ Алексея имеет положительный уровень сходства с обобщённым образом студент. Фамилию я не говорю, но это неважно. Вот. А вот на обобщённый образ студентки он не похож. Уровень сходства отрицательный. А вот Ксения, например, она имеет положительный уровень сходства с обобщённым образом студентка. А с обобщённым образом студент у неё уровень сходства отрицательный. То есть если система в соответствии с моделью относит Алексея к студентам, то это положительное решение. Но эта же модель, она и не относит его к студенткам, потому что у него уровень сходства с образом с образом студентки обобщённым отрицательный. Поэтому принимается системой отрицательное решение о том, что он принадлежит к образу, что относится к категории студентка. То есть он относится к студентам, а к студенткам не относится. А ещё к кому он не относится? К профессорам он не относится, к трактористам не относится, к олимпийским чемпионам, к сожалению, тоже не относится. То есть он не относится ко многим образам, а ко многим относится: студент, сын, брат, может быть, там, может быть, в будущем муж. Понятно, да? То есть к каким-то категориям он относится, а к каким-то не относится. Система должна определить. Для каждой категории вычисляется уровень сходства, для каждого класса. Определяется, насколько этот конкретный объект сходен и насколько он отличается от каждого из классов. Вот идём мы по дорожке в университете, видим вокруг себя тысячи, десятки тысяч, сотни тысяч или даже миллионы объектов каких-то. Ну, деревья, листики там, вороны, голуби там, чайки там, собачки, студенты, профессора, урны, машины там, ну масса всего, здания, конечно. Заходим вовнутрь, а там ещё компьютеры там, столы там, стулья, да? То есть мы видим огромное количество различных объектов. Вы когда-нибудь замечали, что вы устаёте их идентифицировать? Что вот вы устали определять, а что это дерево, а не машина, а это собака там, а не ворона? Какое-нибудь усилие вы испытывали, когда вы их идентифицировали, нет, внутреннее? Нагрузку какую-то, напряжение? То есть вы чувствуете, что какую-то работу вы выполняете, когда вы узнаёте, что это за объект, или нет? Скажите мне. А, ребят?
Ну, это да, если не после 8 марта, то, конечно, мы можем отличить собаку от машины. А после 8 марта, ну, вот если прямо вот девятого, то там ещё кто его знает. Вот десятого опять начинаем уже отличать собаку от машины, да? Ребят, я хочу вам сказать, что наши модели субъективные, которые мы используем для узнавания объектов, процессов и явлений, они имеют высокую степень совершенства. И они великолепны по своим возможностям вычислительным. Великолепны просто. То есть мы узнаём тысячи объектов, которые видим. Вот мы, допустим, едем на машине, видим тысячи объектов в динамике. И мы их узнаём без какого-либо внутреннего усилия в реальном времени. Значит, когда стали делать машины, которые ориентируются, ну, автоматически автопилоты автомобильные там, которые машину ведут вместо водителя, то оказалось, что это довольно-таки проблематично сделать. То есть чтобы машина правильно идентифицировала объекты, для этого нужны колоссальные вычислительные мощности. Причём этих вот классов не очень много этих объектов у робота, который ведёт машину. Он знаки идентифицирует, полосу, по которой идёт там машина и так далее, препятствия идентифицирует, и вид препятствия, и расстояние. Вот. Ещё там используются всякие подсказки, ему там по GPS-навигации говорят, где он там находится с точностью до сантиметра там почти что и так далее, и так далее. Так вот, ребята, нужны колоссальные вычислительные мощности для того, чтобы эту идентификацию в реальном времени осуществить, даже с небольшим числом объектов и небольшим числом классов. То есть сначала же нужно ещё выделить эти объекты из фона, нужно признаки этих объектов идентифицировать, нужно потом создать образы конкретных объектов, конкретные образы, а потом эти конкретные образы идентифицировать с обобщающими. Это колоссальные вычислительные затраты, ребята. Вот сейчас мы будем потом считать задачи в интеллектуальных системах, и мы будем видеть пример того, как эти задачи решаются вычислительными системами. Они решаются, но пока, мне кажется, что далеко им до того, чтобы как человек ориентироваться в окружающей среде. Хотя прогресс технологии настолько быстрый сейчас и стремительный, что не исключено, что через некоторое время они даже и превзойдут человека, эти технические системы. Ну я, например, видел робота, который играет в теннис, так там у него уже ракетку не видно. Она как пропеллер у самолёта гудит там, свистит. И он отбивает тысячи шариков в минуту, которые стреляют из этих пушек. Знаете, вот пушки такие для тренировки теннисистов, шариками стреляют. Вот мне стоят таких пять пушек с разных сторон, стреляют под разными углами, ещё наклоняются, повыше, пониже. Динамики, он их всех отбивает. И не просто отбивает, а они все ложатся на столик противника, на сторону противника, и тушатся, то есть они улетают на большой скорости, кручёные под разными углами, то, что их невозможно отбить. То есть я не исключаю, что такой робот, он скоро может победить на Олимпийских играх, например, там, если там такой вид есть, я не помню, есть или нет. Ну вы поняли, да, к чему я говорю? То есть это всё в процессе развития находится.

**Подраздел 5.6: Формулы F-меры и их ограничения**

Так вот, что предложил, собственно, Ван Рисберген? Он предложил посчитать, сколько у нас истинных и ложных положительных решений. То есть объект отнесён моделью к какому-то классу, и это так и есть, он действительно относится к этому классу. Это решение истинное. Или объект относится к какому-то классу системой, то есть уровень сходства больше нуля, но это решение ложное, ложное срабатывание. То есть объект отнесён системой к классу неправильно, ошибочно. Это очень, ребята, опасная ситуация, очень, прямо крайне опасная ситуация. Представьте себе, что средства ПВО стоят противовоздушной обороны. И на средства разведки дают информацию о воздушных целях. А интеллектуальные системы идентифицируют эти цели по их поведению, по их различным характеристикам, которые системы разведки смогли измерить. И есть там, конечно, транспондеры, то есть есть там какие-то сигналы, я свой самолёт там и так далее. И может и не быть иногда там бывает, что и нету из-за каких-то помех там и средств радиолокационной борьбы. И вот представьте себе, двигается цель, да? Попадает, входит она уже в зону поражения. Система идентификации этой цели дала такое решение, что это похоже, что это цель, ну, объект противника, подлежащий уничтожению, приближается к базе Хмеймим, да, в Сирии. Подлежит уничтожению. И, значит, шандарах по нему из этого Панциря 1С там, и сбили его. Вот. А это было ошибочное решение. Это решение было неправильным. Часто провоцируют средства ПВО на то, чтобы они совершили подобные ошибки. Например, известно, что ракета ПВО наводится на цель с большим с большей величиной сигнала. Если есть две цели, то она перенаправляется на ту, которая даёт больше отражения, больше сигнал. И вот представьте себе, взлетает на аэродроме гражданский Боинг, там 300 человек сидит или Аэробас какой-нибудь. А истребитель противника пролетает между средствами ПВО и этим вот гражданским судном воздушным большого размера. И совершает манёвр такой и уходит на малой высоте. А средства ПВО идентифицируют этот истребитель как самолёт противника. И совершенно правильно идентифицируют, ребята, правильно. Он действительно является самолётом противника. Потом вдруг этот самолёт противника куда-то исчезает из поля зрения, потому что высота 15 м, сверхзвуковая скорость, использование складок местности, он уходит, его средства обнаружения уже не видят вообще. А вот этот эта средства ПВО, ракета летит прямо на этот Боинг и сшибает его. Все погибают. Значит, такие случаи есть. И у нас на базе Хмеймим такой случай был, когда сбили наш самолёт радиолокационной борьбы и разведки, Ил-18 там какой-то модификации. Вот. Именно вот так вот пролетели израильские самолёты, и его на него перенавели средства ПВО сирийские, и сбили наш самолёт российский. И точно так же было через неделю или там через две в Иране. Абсолютно по такому же сценарию. Пролетел самолёт военный израильский между этим вот средствами ПВО и аэродромом гражданским. И был сбит гражданский самолёт. Очень большой скандал был, люди поуходили со своих постов там генералы. Вот. То есть неверная идентификация, то есть false positive, то есть положительное решение ложное, ложное срабатывание – это бывает крайне негативные последствия может иметь. И решение отрицательное о том, что объект не относится к такому-то классу. Есть тоже истинное негативное решение (true negative) и ложное негативное решение (false negative).

**Подраздел 5.7: Нечеткое мультиклассовое обобщение F-меры**

Что предложил Ван Рисберген? Он предложил формулы, в которые используются эти вот сумматоры. Все эти виды: true positive, false positive, true negative, false negative. Предложил формулы для оценки точности модели, полноты модели и с использованием точности и полноты модели среднее геометрическое – достоверность модели. На самом деле есть статьи по критерию Ван Рисбергена, где вот эти вот различные характеристики модели, которые получаются на основе вот этих вот сумматоров, их настолько много, что они на целую страницу там одни формулы вот эти вот характеристики модели. Я видел такие статьи. Вот. Но я могу сказать так, что всё это может развиваться неограниченно все эти меры достоверности. Вот, но на практике обычно используется самый простой вариант – симметричная средняя геометрическая достоверность модели, F-мера Ван Рисбергена. Вот. Но мера Ван Рисбергена, она очень убедительна и логична, и обоснована. Однако для системы Эйдос она слишком примитивна. Сейчас я объясню почему. Смотрим на результаты идентификации в системе Эйдос. Вот, допустим, мы видим мышка один отнесена к классу, конкретному классу мышка и к обобщающему классу элемент компьютера. Видите, да? А ещё она отнесена к классу телефон, к классу средства связи. Вот там, где истинно положительное решение, там птички стоят. А там, где ложно положительное, там уровень сходства больше нуля, но птичек нет. А ещё есть отрицательные решения, там, где уровень сходства меньше нуля, синеньким здесь показано. И вот так мы можем по разным посмотреть объекты. И вот мы берём клавиатура два, ребята. Смотрим. На обобщённый, на конкретный класс клавиатура похожа на 95%, на обобщающий класс элемент компьютера на 56%. А ещё похожа на сумку на 4%, на аксессуар на 4%, на средства связи на 22%. Вот. А вот, то есть мы видим и истинные, и ложные решения. Вот. Но мы-то видим, ребята, смотрите сейчас, что решение о том, что клавиатура два является аксессуаром или сумкой – это ложно положительное решение. И по методологии, которую предложил Ван Рисберген, мы к сумматору false positive должны просуммировать единичку для каждого ошибочного решения положительного. А я говорю: "Нет, ну это несправедливо, ребят, ну зачем так наказывать, так жёстко прямо вот подходить к этому вопросу?" Ведь система Эйдос, она не просто говорит о том, принадлежит объект классу или не принадлежит. Она ещё говорит о том, в какой степени он принадлежит. Это нечёткая принадлежность, в нечётких множествах, понимаете, степень принадлежности. Или можно сказать так, она интерпретирует, можно по-другому это интерпретировать. Она показывает уровень сходства, который отражает степень уверенности системы в том, что этот объект, клавиатура два, принадлежит к этому классу, обобщающий класс аксессуар. Степень уверенности, надо сказать, чрезвычайно низкая, ребята, всего 4%. Что такое 4%? Это 4 сотых. 4 сотых. Почему Ван Рисберген суммирует в сумматору false positive не 4 сотых, а единицу? Я так думаю, что это неправильно. Поэтому я предложил нечёткое обобщение меры Ван Рисбергена.

**Подраздел 5.8: Мультиклассовость и инвариантность к объему выборки**

Теперь, у него F-мера предполагает, что объект может относиться только к одному классу. У нас так бывает или нет в системе Эйдос? Бывает. Вот если бы у нас была только одна классификационная шкала, то каждый объект принадлежал бы к одному классу. Но у нас их две. Получается, что каждый объект принадлежит двум классам. А вообще их может быть там сотни этих классификационных шкал. То есть он может принадлежать одновременно к сотням классам объектов. И это будет правильно, ребята, это не будет какое-то ошибочное принадлежность. А мера Ван Рисбергена предполагает, что объект принадлежит только к одному классу. Это тоже нас не устраивает. Поэтому я предложил мультиклассовое нечёткое обобщение меры Ван Рисбергена. Следующий момент. Вот эти вот все эти переменные: true positive, true negative, false positive, false negative – это что такое у нас? Это абсолютное количество. Если мы возьмём... Вы видите экран мой, да, ребят? Если мы возьмём... Скажите, видите или нет? Да, мы видим. Видите. Слава Богу. Вот, значит, теперь смотрите. Вот у нас, допустим, это объём выборки. Вот, а вот это у нас количество решений. Количество решений. Вот мы берём, допустим, истинно отрицательные решения. И смотрим, как они зависят от числа объектов выборки. И я вам говорю, сообщаю, что истинно отрицательные решения, true negative, их число очень быстро растёт при увеличении объёма выборки. А ложно положительные, то есть ложно отрицательные решения, они тоже тоже растут в зависимости от объёма выборки, но медленнее. Медленнее. False negative. А истинно положительные решения, если взять, они медленнее растут, чем ложно отрицательные. И, значит, мы видим false positive решение. А false negative, мы нарисуем, они ещё медленнее растут их число с увеличением объёма выборки. Что мы можем сказать? Мы видим, что все эти сумматоры, естественно, увеличиваются пропорционально объёму выборки. Чем больше объём выборки, тем больше значение всех этих сумматоров. Но быстрее всех растут истинно отрицательные решения. Потом, ну, может быть, немножко по-другому. Но истинно отрицательные точно быстрее всех растут. Потом истинно положительные решения растут быстро, а потом ложные решения. То есть это что значит? Что если мы возьмём какие-то формулы на основе этих сумматоров и будем по ним считать, как вот предложил Ван Рисберген, точность модели, полноты, достоверность, то надо тогда будет, получается, возникает вопрос: будет ли зависеть эти вот точность и полнота модели, достоверность модели от объёма выборки? И оказывается, ребята, что да, будет. Будет. Это очень-очень плохо. Потому что это означает, что эти меры дают различные результаты при различных объёмах выборки. То есть зависят не только от самой модели, но и от объёма выборки. Поэтому я написал статью, где предложил обобщение меры Ван Рисбергена, нечёткое, мультиклассовое обобщение. И к тому же ещё и инвариантное относительно объёма выборки. И вот эти меры все достоверности и используются в системе Эйдос.

**Раздел 7: Заключение**

И мы можем посчитать, какое количество каких решений содержится в той или иной модели по результатам распознавания. И вот я вам показываю, что вот, допустим, берём модель хи-квадрат. Мера Ван Рисбергена 0,503. Я знаю, что конец занятия. Мера L1, которую я предложил, 0,864. Мера L2 874. Строим график вот такой числа истинных и ложных, положительных и отрицательных решений в зависимости от уровня сходства. И обнаруживаем, что решения отрицательные всегда истинные, и гораздо больше истинных, чем ложных. А положительные решения у нас при низких уровнях сходства есть ложные решения. А чем выше уровень сходства, тем больше доля истинных решений. Начиная с какого-то уровня сходства, вообще ложных решений нет, все решения истинные. То есть у нас мера, уровень сходства является адекватной мерой степени истинности решения, ребят. Это важнейший вывод. И мы имеем, так сказать, инструмент в системе, позволяющий оценивать достоверность модели. И дальше мы можем уже решать задачи в наиболее достоверной модели.
На этом наше занятие заканчивается, ребята. Кроме этого, может быть, кто-нибудь пришёл ещё из тех, кто публицистикой занимался у нас? Нет? Ну, на этом у нас занятие заканчивается. Всего самого хорошего. До свидания.
До свидания. Спасибо. До свидания. Ксения, я сейчас это тоже под запись рассказывал, и это будет в облаке находиться видеозанятий сегодняшним числом. Вечером, я думаю, будет уже. Надо будет дать ссылочку всем студентам, тем, кто отсутствовал на занятии, обязательно, чтобы они понимали это всё, могли делать контрольную работу. Всё хорошего. До свидания.
До свидания.