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

***Российская Федерация***

**7 Данные, информация и знания. 2020-09-28**

Заголовок: Лабораторная работа №2 по системе Эйдос: Ввод данных и основы моделирования (ЛР 3.03)

Резюме текста:

1. Введение и организационные моменты:

Профессор Луценко Е.В. проводит второе лабораторное занятие по дисциплине "Интеллектуальные информационные системы и технологии" для группы ИТ 1822 (22.09.2020). Занятие является продолжением предыдущего, посвященного установке системы Эйдос. Проверяется наличие установленной системы у студентов.

2. Повторение: Установка и классификация лабораторных работ:

Напоминание о необходимости установки системы Эйдос и классификации лабораторных работ (ЛР) по типам:

Тип 1: Готовые ЛР для изучения (ничего делать не надо).

Тип 2: ЛР, где исходные данные рассчитываются системой (интернет не нужен).

Тип 3: ЛР, требующие ввода данных из внешних источников (текст, таблицы Excel).

Тип 4: ЛР, использующие парсинг данных из интернета.

3. Лабораторная работа 3.03: Идентификация предметов по признакам:

Цель: Изучение процесса ввода данных из Excel и формализации предметной области.

Процесс:

Удаление ранее созданных приложений (режим 1.11).

Добавление новой ЛР (режим 1.3), выбор ЛР 3.03 (тип 3).

Система создает пустое приложение (папка A0001) и копирует туда файл исходных данных Excel.

Файл исходных данных (Excel):

Структура: Строки - объекты наблюдения, столбцы - их описания (шкалы).

Типы шкал:

Классификационные (желтые): Определяют принадлежность объекта к классу (например, "Конкретный класс", "Обобщающий класс").

Описательные (зеленые): Описывают свойства объекта (например, "Цвет", "Материал", "Размер").

Градации: Значения внутри шкал (например, для цвета: "Красный", "Зеленый").

Требования: Названия шкал и градаций должны быть краткими и понятными, без повторений внутри одной шкалы. Числовые шкалы обрабатываются особо.

Формализация предметной области (Режим 2.1-2.3):

Режим 2.1: Чтение файла Excel, создание справочников шкал и градаций, кодирование исходных данных, создание обучающей выборки. Система автоматически определяет типы шкал (номинальные, порядковые, числовые).

Режим 2.2: Просмотр и редактирование справочников шкал и градаций.

Режим 2.3: Просмотр обучающей выборки (объекты и их кодированные описания).

4. Синтез и верификация моделей (Режим 3.5, 3.4):

Режим 3.5: Синтез статистических и системно-когнитивных моделей (баз знаний) на основе обучающей выборки. Рассчитываются различные информационные меры (количество информации, Хи-квадрат, мера взаимосвязи и др.).

Режим 3.4: Верификация моделей путем распознавания объектов обучающей выборки. Оценивается достоверность моделей с использованием метрик Ван Рисбергена (True Positive, True Negative, False Positive, False Negative) и их нечетких обобщений, предложенных лектором (суммирование уровней сходства вместо единиц).

Интерпретация результатов: Анализ графиков достоверности позволяет определить порог сходства для принятия решений и оценить качество модели.

5. Когнитивные концепции:

Данные, информация, знания: Лектор кратко поясняет различие: данные -> информация (понимание смысла) -> знания (информация, полезная для цели).

Конструкты: Понятия с противоположными полюсами и спектром промежуточных значений (например, температура, цвет). Система Эйдос формирует конструкты.

Мировоззрение и парадигма реальности: Система конструктов человека образует его мировоззрение и парадигму реальности. Сложность системы конструктов определяет глубину понимания.

Принцип Эшби: Более сложная система адекватно отображает более простую, но не наоборот. Это объясняет трудности в понимании между людьми с разной сложностью когнитивных систем.

Развитие познания: Обучение расширяет систему конструктов, добавляя новые понятия и углубляя существующие.

6. Технические проблемы и решения:

Во время демонстрации возникла проблема с отображением экрана в Webex.

Предложено использовать Skype для демонстрации экрана как временное решение.

7. Заключение и дальнейшие шаги:

Студенты должны повторить действия на своих компьютерах. На следующем занятии планируется изучение решения задач идентификации, прогнозирования и принятия решений с использованием созданных моделей.

Детальная расшифровка текста:

Раздел 1: Введение и организационные моменты

Здравствуйте, ребята. Здравствуйте. Здравствуйте. Ну, на чём мы остановились на предыдущей лабораторной работе с вами? Кто-нибудь может напомнить?

Сейчас я для записи говорю. У нас запись включена. Я скажу сейчас для записи. Сегодня 22 сентября 2020 года. Третья пара с 11:30 до 13:00 с группой ИТ 1822. Лабораторная работа по дисциплине Интеллектуальные информационные системы и технологии. Вторая лабораторная работа. Занятие проводит профессор Луценко Евгений Вениаминович. Это для записи, чтобы было понятно, что за запись. Открываешь, начинаешь слушать, ничего непонятно, почему, что там, откуда. Вот такая информация.

Учебные вопросы. Изучение лабораторной работы 3.03.

Раздел 2: Повторение: Установка и классификация лабораторных работ

Я вам рассказывал, ребята, о классификации лабораторных работ и установки системы Эйдос. Было дело такое, нет?

Сейчас вы должны видеть мой рабочий стол. Кто-нибудь видит экран моего компьютера? С экранной формой, да, сейчас вот, которая двигается вот. Видно. Видно, да? Ну хорошо.

Ребята, вы должны сейчас всё то же самое делать на своих компьютерах. На своих компьютерах. Ответьте мне в чате, у кого есть компьютеры, плюсики поставьте. А у кого нету, поставьте минусики. Все, 100% студентов, кто сейчас присутствует. Ну, хорошо. В основном есть. Вернее так, пока минусиков ни одного нету. Вот. А теперь мне, пожалуйста, скажите голосом, у вас у всех установлена система Эйдос на компьютерах или нет? У кого не установлена, можете сказать: "У меня не установлена". У нас на прошлом занятии мы должны были закончить тем, что должны были установить систему на компьютеры. Ну так я вижу, что у всех вроде установлено, у всех есть компьютеры. Значит, вы должны повторять на своих компьютерах то, что я показываю на экране. В такой форме у нас лабораторная работа будет идти. А сейчас, пожалуйста, отметьте, у кого установлена система Эйдос на компьютерах. Только не плюсиком, как-нибудь так: "уст", там, или "у". Установлено, установлено, установлено. То есть получилось у вас установить? Ну молодцы.

Раздел 3: Лабораторная работа 3.03: Идентификация предметов по признакам

3.1 Цель и общий процесс

Тогда вот сейчас что делаем? Запускаем систему Эйдос. Нажимаем 1.11 пунктик. Это пункт, который удаляет все ранее созданные приложения. Заходим в режим 1.3. Вам должно быть всё это видно. Сразу же нажимаем на кнопочку "Добавить лабораторную работу". Выбираем третий тип лабораторной работы и кнопочку на идентификация предметов по их признакам, работа 3.03. О'кей. Предупреждение, что всё будет стёрто. Нас это не волнует, у нас там ничего такого особого и нет, чтобы дорожить им в этой папочке исходных данных. И появляется экранная форма программного интерфейса, ребята, которая позволяет вводить данные из внешних источников данных экселевского типа файлов в систему Эйдос. Я вам это рассказывал или нет на прошлых работах, ребята? Ответьте мне. Такое что-нибудь помните наподобие, нет? Не рассказывал, да? Ну тогда, значит, давайте продолжим.

3.2 Файл исходных данных (Excel)

3.2.1 Структура и типы шкал

Значит, что у нас является исходными данными в этой задаче? Как как организованы исходные данные? Значит, вот смотрите, вот здесь вот есть вверху требования к файлу исходных данных на экранной форме. Хелпик есть. Значит, ребята, я прошу вас, если что-то не будет видно или какие-то будут проблемы, ну я же рассказываю, вы вы видите, видно или не видно. Вот, то сразу мне говорите, хорошо? Ну в Скайпе практически буду использовать только для того, чтобы показывать, потому что запись идёт в Вебексе. Чтобы я просто демонстрировал экран, его использую. Вот. Надо было на прошлом занятии догадаться. Там такая же ситуация была аналогичная. Пропал этот общий доступ к экрану.

Так вот, ребята, исходные данные экселевские, они могут иметь вот такой вид в системе Эйдос. Строчки - это наблюдения объектов разных состояний или разных объектов. А колонки - это описание этих вот зелёным фоном - это описание этих объектов путём описания их свойств или факторов на них действующих. А жёлтые - это описание объектов наблюдения, только уже путём описания их принадлежности к каким-то обобщающим категориям. То есть, по сути дела, каждая вот такая строчка - это своего рода определение объекта наблюдения, определение по принципу подведения под более общую категорию и указания специфических признаков. Вот, допустим, кто такие млекопитающие? Определение. Это животные, выкармливающие своих детёнышей молоком. То есть выкармливающие молоком - это специфический признак. А вот здесь вот написано: животные, млекопитающие. То есть млекопитающих. Понятно, да? Так вот, а кошка Мурка выкармливает детёнышей молоком, а ещё там собачка Жучка выкармливает своих детёнышей молоком. И вот здесь можно написать: Мурка, Жучка, там, мышка там и так далее. А здесь написать: млекопитающие, млекопитающие. А здесь выкармливает молоком, есть хвостик, есть шерсть, там, есть четыре лапки. А здесь написать: земноводная. Вот. И здесь, ну, имеется в виду лягушка. Вот здесь в первом колоночке лягушка, там Мурка, Мурка, Жучка, лягушка, царевна, например. У неё э-э тоже четыре лапки, шерсти нет, и выкармливает не молоком своих детёнышей. То есть здесь будет нет написано в колоночке выкармливает молоком. Таким образом мы можем дать конкретное определение конкретных объектов. А на их основе система может сформировать обобщённое определение. То есть вообще определение млекопитающих, к примеру. Понимаете? Кто такие млекопитающие, кто такие земноводные? Вот вам, что она делает. Это один из примеров того, что она делает. Обобщает. То есть конкретное определение обобщает и получает обобщающее определение, обобщённое определение. То есть на основе конкретных примеров формирует обобщённый образ классов. На основе э-э фреймов экземпляров... Да, вот такая строчка, где даётся определение объекта, с одной стороны, путём описания его специфических признаков, а с другой стороны, принадлежности к обобщающей категории, называется онтологией.

3.2.2 Конкретные терминологии (Статичная и Динамичная)

Я использую универсальную терминологию и две конкретных терминологии для описания вот этих онтологий. Универсальная терминология. Вот эти вот зелёненькие э-э называются, которые описывают объекты э-э их свойствами, я называю описательные шкалы. А вот здесь вот в колоночках есть информация для формирования градаций описательных шкал, их значений. То есть здесь значения описательной шкалы. А это классификационные шкалы, описывают обобщающие категории, к которым относятся объекты наблюдения. Это обобщающее, а здесь информация для значений этих э-э классификационных шкал. А конкретная терминология - два вида. Один вид конкретной терминологии описывает статичную ситуацию, статичную. А другой - динамичную ситуацию. Для статичной ситуации признаки, вот эти объекты и принадлежность к категориям наблюдаются одновременно. То есть признаки сейчас и принадлежность к категории тоже сейчас. Тогда мы говорим так: колоночки обозначают свойства объекта, а в клеточках здесь написаны значения этих свойств. Например, колоночка - цвет, а значение - красный. Колоночка - вес, а значение - 60 кг. Вот. А то же самое касается обобщающих э-э классификационных шкал, обобщающих категорий. Тоже это в данном случае будущие состояния объекта моделирования. А нет, не будущие, а текущие состояния объекта моделирования. Динамическая интерпретация. Колоночка - это фактор какой-то, ну, допустим, температура. Значение этого фактора - 25°. Результат, или, допустим, там 100°. Результат - яйцо всмятку, например, сварилось. Или вкрутую сварилось. Вот. То есть мы понимаем, что фактор действовал до того, как объект перешёл э-э в это будущее состояние. Вот. То есть факторы относятся к прошлому периоду, а состояние объекта - к будущему периоду, которые возникли под действием этих факторов.

3.2.3 Типы шкал (Номинальные, Порядковые, Числовые)

Значит, если у нас нет необходимости, мы не будем э-э специально уточнять э-э статичная или динамичная у нас интерпретация, а будем использовать универсальную терминологию: описательные шкалы и градации, классификационные шкалы и градации. И описательные, и классификационные шкалы и градации могут быть как номинального типа, так и порядковыми могут быть, и числовыми. Сейчас я вам приведу ссылочку на работу, где это подробно описывается, что такое типы шкал, какие они могут быть, какие они могут быть виды. Номинальные, порядковые и числовые. В системе Эйдос все эти типы шкал могут быть использованы. Вот. И сейчас я вот сюда в чате вам в Вебексе пошлю ссылочку на статью. Увидели, да, ребята? Ну, могу и в Скайпе послать. Для гарантии. Чтобы уж точно вы получили её. Вот.

3.2.4 Требования к данным в Excel

Это описание требований к файлу исходных данных. Но для того, чтобы программа э-э система Эйдос смогла её этот файл ввести в систему, конвертировать его во внутренний стандарт данных и ввести. Для этого ей нужно указать диапазоны классификационных шкал, от второй, допустим, колоночки до третьей, и описательных шкал, скажем, от четвёртой до двенадцатой. А сколько строк, она сама определяет. Вот здесь мы и указываем. Обратите внимание. Значит, мы указываем, что э-э исходные данные имеют тип файла старый Excel. Классификационные шкалы имеют э-э номера от второй до третьей, диапазон классификационных шкал. Диапазон описательных шкал с четвёртой по двенадцатую.

3.3 Формализация предметной области (Режим 2.1-2.3)

И мы хотим произвести формализацию предметной области. На лекциях я вам говорил это, ну сейчас ещё раз повторю, что формализация предметной области включает в себя разработку справочников классификационных и описательных шкал и градаций и кодирование с их помощью файла исходных данных и получение таким образом в результате обучающей выборки. Если там встречаются в исходном файле числовые шкалы, тогда появляется возможность задать разные или одинаковые числовые интервалы, равные по величине с разным числом наблюдений или разные по величине интервалы, но с равным числом наблюдений. Если у нас уже созданы ранее классификационные описательные шкалы и градации обучающая выборка, то мы можем вот здесь поставить опцию и создать э-э тестовую выборку или распознаваемую. Вот, на основе файла такой же точно структуры, как исходные данные. Ну сейчас мы просто нажимаем О'кей. Происходит преобразование экселевского файла в DBS формат, и возникает вопрос о том, сколько диапазонов числовых использовать в числовых шкалах описательных. Слева мы имеем информацию по э-э классификационных шкалах и градациях, справа об описательных. Вопрос о том, сколько диапазонов возникает с той стороны, где описательные шкалы и градации. А там, где классификационные, про это не спрашивается. Почему? А там нет числовых шкал, видите, ноль числовых шкал. А здесь одна есть. И в ней 12 градаций сейчас. Если мы хотим их поменять, мы здесь меняем, и потом нужно пересчитать модель.

3.3.1 Просмотр справочников (Режим 2.2)

Теперь смотрите, суммарное количество градаций классификационных и описательных шкал. То есть сначала идут колоночки, сколько колоночек в файле, а потом сколько строк. Потому что колоночки - это классификационные шкалы, а строки - описательные шкалы и градации. Если мы здесь ничего не меняем или поменяли и нажали "Пересчитать шкалы и градации", то потом просто выходим на создание модели. У нас здесь появилось окошко. Вы меня видите, ребята, что я показываю экранные формы, видите? Скажите: "Да, нет", там в чате напишите где-нибудь. Да. Вот, видите. Ну, нормально значит. А у вас на компьютерах получается это сделать, у кого компьютеры там и система установлена? Да, делаем. Ну ладно. Как-то с каким-то сомнением возникло, я не знаю, смутно. Ну делайте, хорошо.

Значит, смотрите, ребята, что сделано было в этом режиме? Этот режим, о котором мы сейчас запустили, окошко, где вы диапазоны задавали, называется программный интерфейс ввода данных из экселевских файлов, универсальный программный интерфейс. Программный интерфейс - это программа, преобразующая данные из одного стандарта представления в другой. В данном случае данные были преобразованы из экселевского файла в DBS формат, внутренний для системы Эйдос. Потом этот файл был изучен, были сформированы классификационные описательные шкалы и градации, и с их помощью был закодирован исходные данные, и создана база событий и обучающая выборка. Потом всё было переиндексировано.

3.3.2 Просмотр обучающей выборки (Режим 2.3)

Потом появляется такая экранная форма. Студенты обычно имеют такую особенность, я уже замечал, когда вот они в аудиториях работали ещё, они просто сразу клацали О'кей, не читая, что здесь написано. Вот, а я это писал для них. Там написано, что мы можем посмотреть на сам файл исходных данных, путь на него указан. Вот давайте сейчас по этому пути на своих компьютерах открывайте файл исходных данных, по тому пути, который у вас конкретно указан на ваших компьютерах. Ну у меня вот такой получился файл исходных данных. Этот файл находится, знаете где? До установки лабораторной работы, я вам показывал. LabWorks и все работы пронумерованы. Каждая работа - 10 штук каждого типа. 10 штук работ первого типа, 10 штук второго типа и 10 штук третьего типа. Двадцать третья работа, наименование её, идентификация предметов по их признакам. Файл исходных данных. Вот этот файл исходных данных, он был скопирован вот отсюда. Вот, из этой папочки. LabWorks 23. Было создано пустое приложение, папочка А0001, и папочка System, и сюда создаётся, здесь создаются все файлы приложения. А скопировано было вот сюда файл исходных данных, папочку ImpData. Это папочка исходных данных. И вот здесь был этот файл помещён при установке лабораторной работы. А потом то, что я сказал, созданы классификационные описательные шкалы и градации, и с их помощью закодирована обучающая выборка.

Значит, мы видим, что здесь есть и текстовые колоночки, и числовая есть. Теперь выполняем эту рекомендацию, которая там у нас была. Какая у нас была рекомендация? Э-э, зайти в режим 2.1, 2.2, 2.3. Ну давайте это и сделаем. И потом посмотреть, э-э, сами модели, осуществить синтез моделей в режиме 3.5, посмотреть на сами модели, которые получились в последующих режимах и посмотреть, как это осуществляется преобразование данных в информацию, а её в знания, согласно алгоритму, приведённому э-э в режиме 6.4. Сейчас мы здесь увидим, что возникла лабораторная работа здесь в диспетчере приложений. Э-э, и она является текущей. W - это Work, значит, она рабочая в данный момент. Значит, мы э-э переходим в режим 2.1, как нам было рекомендовано, и видим слева названия двух классификационных шкал, а справа - градации этих шкал. Эти шкалы у нас номинального типа. То есть в них есть э-э названия текстовые, которые никак не упорядочены между собой. Но они здесь рассортированы по алфавиту, но смысл этих вот э-э в этом порядке отсутствует какой-либо. Если бы мы хотели, чтобы эта шкала была порядковой, тогда нужно было пронумеровать объекты, то есть градации этой шкалы. Допустим, монитор 1.монитор, 2.клавиатура, 3.мышка и так далее. Тогда при сортировке был бы тот порядок, который мы задали. Вот. И тогда бы эта шкала была порядковая. Это классификационные шкалы. Название классификационных шкал полностью повторяют заголовок экселевского файла. А градации полностью повторяют э-э градации этих шкал текстовые. Но э-э при этом каждая градация встречается один раз. То есть это э-э сортировка с признаком Unique и э-э получается нормализация исходных данных, по сути дела. То есть исходные данные кодируются с использованием этих справочников, нормализуются. Описательные шкалы абсолютно аналогично. Название из экселевского файла. Вот. Здесь на что надо обратить внимание? Значит, здесь есть э-э шкалы, которые могли бы быть порядковыми. Вот, скажем, цвет вполне можно было бы упорядочить его по спектру, например. Хотя здесь есть цвета, которые в спектре вроде как отсутствуют, но можно было бы упорядочить по спектру. Вот. Материал едва ли можно как-то разумно упорядочить, ну, может быть, по твёрдости, не знаю. Размер один - это точно, однозначно совершенно порядковая шкала. Но здесь она у нас не является порядковой по той простой причине, что у нас сортировка по алфавиту, и те э-э градации, которые по смыслу должны быть первыми, они у нас в середине находятся, те, которые должны быть последними, они находятся в начале. Потом мы видим, смотрите, здесь есть две градации: большая, большой. Мы должны, конечно, убрать либо большая, либо большой. Ну, лучше, наверное, первый убрать вариант - большая. Почему? Потому что какой предмет? Большой, маленький, средний. То есть ответ на вопрос, какой предмет. Поэтому здесь мы вот эти два объединить бы должны были бы большой-большая. Для того, чтобы это сделать, для этого нужно в экселевском файле контекстной заменой в колонке большая поменять на большой. А потом поставить номера. Допустим, маленький - один, наклонная черта - четыре, потому что будет четыре градации получится, - маленький. Один, то есть два, наклонная черта - четыре, средний. Три, наклонная черта - э-э четыре {тире} под руку. И четыре, наклонная черта - четыре {тире} большой. Потом при сортировке это будет упорядочено по смыслу. Минимальное значение будет первым идти, потом среднее, а потом максимальное. Берём размер два. Это та же самая шкала, но здесь уже числовая она. Система сама нашла минимальное, максимальное значение в шкале и разделила на заданное число числовых диапазонов. Причём равных по величине. Семь знаков э-э расчётных в системе Эйдос. Наличие экрана есть, нет. Вот здесь на что я ваше внимание обращаю? Что мы не только есть указываем, но и когда нет, тоже указываем. Есть или нет, указываем. Э-э, часто психологи что-то спрашивают, например, или социологи, и там у них нет варианта. Ну, допустим, там, вы хотели бы, чтобы там сделали остановку ближе к вашему этому дому? И, значит, если вы хотели бы, вы птичку ставите, и всё. А если птичку не ставите, то как бы тогда получается, что вы на этот вопрос не ответили. Вот. А в системе Эйдос на, если задан вопрос, то всегда есть ответ: положительный или отрицательный, или развёрнутый. Как вот, допустим, здесь вот, э-э, размер, например. Ну и то же самое касается наличия кнопок. Э-э, форма. Форма - это, конечно, не порядковая шкала, это номинальная шкала. То есть здесь нет порядка этих вот градаций разумного.

И есть две формы, где мы выводим результаты кодирования исходных данных. Вот здесь вот у нас форма из трёх окошек. В верхнем окошке - наименование объекта выборки, в левом нижнем окошке - коды классов, а в правом нижнем окошке - коды признаков. То есть признаком называется градация описательной шкалы, э-э, значение свойства или значение фактора. А классом называется градация классификационной шкалы. И в некоторых случаях, вот когда мы вводим из экселевских файлов данные, то форма, файл исходных данных ещё представляется вот в таком вот виде. Здесь мы видим что? Э-э, тот же по структуре файл исходных данных, что был в Экселе, как мы помним. Вот он он. И сейчас система Aidos. Вот смотрите, очень похожие файлы. Вот. Ну я вот так сделаю, чтобы ещё более похожими они были. Вот. Мы видим, смотрите, что э-э колоночки классификационных шкал жёлтым фоном, наименование объекта выборки, а потом вместо наименований и чисел, ребята, коды приведены этих градаций шкал соответствующих. Причём э-э для числовой шкалы это коды числовых диапазонов. И вот уже на этом уровне, если вы посмотрите, то здесь уже не видно никакого различия между числовыми шкалами и текстовыми. О чём это говорит? О том, что система Aidos все дальнейшие, вся дальнейшая обработка в системе Aidos, она выглядит совершенно одинаково и для числовых шкал, и для текстовых шкал. Ничем не отличаются. То есть дальше обрабатываются они единообразно.

Раздел 4: Синтез и верификация моделей (Режим 3.5, 3.4)

4.1 Синтез моделей (Режим 3.5)

Вот это я хотел сказать. Теперь, если мы посмотрим внизу вот здесь, здесь есть э-э шапка ещё одна с нумерацией колонок. Колонка с номером нулевая, такой колонки нету в экселевском файле. Ну, вернее так, она там есть, но это вот это слева вот сами эти номера строк. Э-э, и потом дальше идёт колонка наименования объектов обучающей выборки, классификационные шкалы, они имеют э-э номера 2-3, со второй по третью. А наименование объектов выборки - первая э-э колоночка. А потом с четвёртой по двенадцатую идут э-э описательные шкалы. То есть эти вот номера внизу, они точно соответствуют прямо вот экселевскому файлу э-э и соответствуют э-э тем параметрам, которые мы задавали в интерфейсе программном 2322. Вот. Так что мы уже осуществили формализацию предметной области.

Теперь следующий этап у нас, ребята, - это синтез и верификация моделей. Для этой цели мы используем режим 3.5. Здесь есть много разных нюансов, но я не знаю, стоит ли вам все это рассказывать сейчас. Ну, во-первых, нюанс такой: мы можем расчёты вести на графическом процессоре, а можем на центральном процессоре. На графическом процессоре одна модель рассчитывается неправильно, но это не влияет на её достоверность. Там другие единицы измерения. Поэтому я сейчас ставлю на центральный процессор. Теперь, эта модель, которая сейчас будет создана, она составляет 60% от того, что могла бы создать система Aidos по размерности модели. То есть вот эта модель, которую мы сейчас будем создавать, она по своей размерности 60% от того, что можно было бы обработать в системе Aidos. Как вам это нравится? То есть там можно обрабатывать модели очень большой размерности в ней. Ну, конечно, для этого желательно графический процессор использовать. Теперь, если модель очень большой размерности, очень много объектов обучающей выборки, ну, скажем, миллион. В Экселе вполне можно 700.000, там 900.000, миллион строчек сделать. И создать модель на основе соответствующего числа примеров. Так вот, если создать модель - это процесс достаточно быстрый, ребята. То есть даже на центральном процессоре, ну займёт там 15-20 минут. Ну то есть это такие, ну, такое время как бы обозримое.

4.2 Верификация моделей (Режим 3.4)

А вот верификация модели связана с распознаванием. Что такое распознавание? Это сравнение каждого объекта распознаваемой выборки со всеми классами. А если этих классов, допустим, там 400, к примеру, или 1.000? В системе Aidos можно до полутора тысяч классов использовать э-э в моделях. Э-э, а самих объектов, там 800.000. Вы представляете себе, каждый из этих объектов из 800.000, каждый из них сравнивается со всеми 400 классами. И определяется степень сходства с каждым из классов. А чтобы определить сходство объекта с классом, для этого все признаки этого объекта э-э используются, вычисляется количество информации суммарное, которое содержится во всех признаках объекта, которые у него есть, о принадлежности к каждому из классов. То есть мы берём сначала все признаки, суммируем и делаем это для каждого из классов, то есть 400 раз, и для каждого из объектов выборки, то есть 700.000 раз или 800.000 раз. Это очень вычислительном отношении очень трудоёмкая операция. На графическом процессоре она выполняется секунды, ребята, эта операция. На центральном, э-э, может выполняться, вот для всего этого 700-800.000 объектов выборки, она может выполняться, скажем, минуты полторы-две эта операция распознавания. На центральном процессоре это может выполняться месяц, например, там или там неделю там. То есть вот такие вот дела. Поэтому я сделал здесь возможность разделения синтеза модели и верификации модели. Потому что верификация - это очень трудоёмкая операция. Если, значит, у нас модель небольшая, тогда синтез и верификация выполняются. Если модель большая и трудоёмко оценить её достоверность, то тогда мы осуществляем только синтез модели, а потом только верификацию. Причём верификацию осуществляем на небольшом числе примеров. Допустим, берём вот здесь вот 100 примеров или 1.000. И в течение 10 минут э-э система осуществляет верификацию этих моделей на тысячи случайно выбранных объектов обучающей выборки. После этого мы составляем пропорцию: 1.000 объектов - 15 минут, 800.000 объектов - X минут. Понимаете? И вычисляем, что 800.000 объектов она будет проверять неделю. А нам хотелось бы это сделать за 8 часов. Мы берём, делим неделю на э-э там на три, грубо говоря, там, да, или что там, получаем э-э в сутках там сколько часов там? 24 э-э на семь, да, это получается сколько часов в неделе? А мы хотим, чтобы за 8 часов. Вот мы делим число часов в неделе на три. Вот. И получаем, сколько объектов выборки мы можем взять. Мы можем взять, скажем, 200.000 объектов. И запускаем на таком количестве верификацию и получаем контур результата. Эти объекты обучающей выборки, на которых мы проверяли достоверность модели, мы можем удалить из обучающей выборки, а можем не удалять. Если мы их удалим, тогда получается какая задача решается? Часть обучающей выборки используется для синтеза моделей, а другая часть - для верификации. Такой подход, когда вот мы берём часть выборки, используем для одних целей, а другую часть - для других целей, называется bootstrap-ный подход, bootstrap. Это на каком-то скандинавском языке обозначает складной нож. Имеется в виду, что вот мы взяли там лезвие, вытащили то одно, то другое. Вот. И вот в ноже у нас есть много лезвий, разных инструментов. Вот примерно идея такая, что мы часть выборки использовали для одной цели, другое лезвие вытащили, а другую часть для другой цели. Потому что это разные инструменты для разных целей. Вот. Но я думаю, что никакого смысла нет э-э удалять эти объекты. Почему? Потому что если там 800.000 примеров, то каждый из них, из этих 800.000, будет оказывать очень малое влияние на результат. Одна восьмисоттысячная его влияние будет, одна восьмисоттысячная на результат. Практически это то же самое, что его там и не будет. То есть будем ли мы удалять один из 800.000 или не будем его удалять, ничего не поменяется.

4.3 Интерпретация результатов верификации

Ну вот, значит, рассчитывается при этом режиме, при этой экранной форме, значит, сейчас мы рассчитаем три модели статистических - это матрица абсолютных частот, матрица условных и безусловных процентных распределений. У нас на лекциях было про модели, рассказывал про модель подробно. И семь системно-когнитивных моделей баз знаний. Значит, сейчас идёт процесс какой? Сначала э-э Так. Сейчас мы запустим заново всё это. Что-то я там, когда показывал, накуралесил немножко. Вот. Сначала осуществляется синтез модели, ребят, а потом проверка их на достоверность. И при этом здесь вот выводятся проценты исполнения каждого пункта. При проверке моделей на достоверность рассчитывается одновременно 11 выходных форм. Ребят, если бы они не рассчитывались, тогда бы было бы намного, ну, скажем так, мы бы достоверность, допустим, определили, но при этом бы не знали очень многих вещей, которые в этих выходных формах э-э определяются. Вот. И вот здесь вот есть хелп этого режима. Исследование достоверности модели. Значит, э-э теперь слушайте внимательно. Я вам объясню, как измеряется достоверность модели сейчас. Значит, э-э модель должна обеспечить правильное отнесение объектов к тем классам, к которым они в действительности относятся, и правильное не отнесение их к тем классам, к которым они на самом деле не относятся. При этом модель может ошибаться. Если она не очень достоверна, не стопроцентная достоверность, она может ошибаться. Она может иногда относить объекты к тем классам, к которым они не относятся на самом деле. Называется ложное срабатывание или ложноположительное решение. А может э-э не относить объекты к тем классам, к которым они относятся. Называется ложноотрицательное решение. То есть на самом деле объект относится к этому классу, а система его туда не относит к нему. То есть есть истинно положительные, истинно отрицательные и ложноположительные, и ложноотрицательные решения. Ну, наверное, вы с такими словами знакомы: True, False. True - это что такое? Истина. А False - это Ложь. Ложь. Вот. А положительного, отрицательного решения называется позитив и негатив. Положительное - позитив, отрицательное - негатив. Что означает положительное решение? Это решение, когда уровень сходства объекта с классом выше нуля. Это положительное решение. Решение о принадлежности объекта классу. А когда уровень сходства ниже нуля, то это решение о непринадлежности объекта классу. Понятно? Понятно, да? Так вот, теперь смотрите, ребята. Значит, э-э система должна правильно относить объекты к классам и правильно не относить. И она может и ошибаться и в первом, и во втором случае. Если она правильно относит и правильно не относит, то это называется True Positive и True Negative решение. Истинно положительное, истинно отрицательное решение. А если она относит объект к классу, к которому он не относится, то это называется ложноположительное решение. А если она не относит объект к классу, к которому он относится, то это называется ложноотрицательное решение. То есть уровень сходства отрицательный, а объект относится к этому классу. Уловили, ребят? Вот. Так вот, э-э Ван Рисберген, э-э математик, э-э сейчас он вроде бы как работает ещё, но такого старшего поколения, он предложил ввести сумматоры. Четыре сумматора: True Positive, True Negative, False Positive, False Negative. И берём мы какое-то количество объектов распознаваемой выборки, ну, обучающей выборки, и распознаём их, копируем их в распознаваемую выборку, и получаем результат. Значит, любой объект, который мы распознаём, он увеличивает один из сумматоров. Он True Positive, True Negative увеличивает, или False Positive, False Negative. Потом Рисберген предложил формулы, которые вот здесь вот приводятся в этом хелпе, по которым рассчитывается достоверность модели. Эти формулы, ребята, они обладают одним э-э свойством или даже, скажем так, у них есть некоторые недостатки в этой мере Ван Рисбергена. Хотя вообще она сама по себе очень э-э такая убедительная, очень-очень убедительная. Значит, э-э суммируется всегда единичка к этим сумматорам Рисбергеном. Система дала ложное срабатывание, ну то есть отнесла объект к классу, суммируется единичка к ложноположительному решению. А на самом деле он к нему не относится этот объект к этому классу. Единичка к этому сумматору добавляется. То есть он просто подсчитывает количество таких решений. А в системе Эйдос определяется степень сходства объекта с классом. То есть если, допустим, система относит объект к классу, но уровень сходства, если в процентах считать, допустим, 3%, а максимальное сходство - 100%, она ошиблась, ребята, но она сама очень не уверена в том, что она относится к этому классу, всего 3% сходства. Я предложил нечёткое обобщение меры Ван Рисбергена, которое заключается в том, что суммировать не единички, ребята, а суммировать уровень сходства, нормированный к единичке. Я это всё описал в статье. Обобщение меры Ван Рисбергена. Вот, описал в статье. Инвариантное относительно объёмов данных. Что это значит, ребята? Сейчас я вам объясню. Значит, дело вот в чём. Дело в том, что э-э мера Ван Рисбергена рассчитывается на основе частот. Сумматоры там, там там просто сумматоры используются, сколько раз встретилось то или иное решение, понимаете? А я так посчитал, что более правильно было бы не не суммы использовать, а вероятности. Какова вероятность э-э истинно положительного, истинно отрицательного, ложноположительного или ложноотрицательного решения? Но вероятности мы посчитать не можем, потому что это предел, к которому стремятся относительные частоты при неограниченном увеличении числа испытаний. Но мы можем посчитать относительные частоты. И я предложил заменить э-э сумматоры относительными частотами. И предложил относить инвариантное относительно объёмов данных нечёткое мультиклассовое обобщение меры э-э Ван Рисбергена. Вот вам статью сослался на эту статью. И эту меру применил в системе Эйдос.