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

***Российская Федерация***

**146 Математическое моделирование и анализ данных в садоводстве. Лекция 7. Численный пример решения задачи в Эйдос 2020-10-20**

**Заголовок:**

Системно-когнитивный анализ в Эйдос: Численный пример обработки данных и оценки моделей

**Резюме текста:**

Лекция посвящена демонстрации работы интеллектуальной системы Эйдос на численном примере. Рассматривается задача анализа влияния морфологических свойств помидоров на количественные, качественные и финансово-экономические результаты их выращивания в условиях неотапливаемых теплиц юга России.

**1. Подготовка и формализация данных:**
Лектор объясняет необходимость формализации предметной области для ввода данных в систему Эйдос. Это включает:

* Определение факторов (морфологические свойства) и результатов (количественные, качественные, финансово-экономические показатели).
* Создание справочников факторов и результатов с их градациями.
* Кодирование исходных данных (представленных в Excel, могут быть числовыми и текстовыми) с использованием созданных справочников.
* Подчеркивается важность правильной подготовки Excel-файла (input\_data.xls или .xlsx в папке indata) и определения диапазонов классификационных (результаты) и описательных (факторы) шкал.

**2. Системно-когнитивный анализ и обработка данных в Эйдос:**
Описываются этапы системно-когнитивного анализа, реализуемые в Эйдос:

* Преобразование данных в информацию (выявление причинно-следственных связей) и затем в знание (использование для достижения целей).
* Обсуждается метод автоматической квантификации числовых данных по интервалам с равным числом наблюдений (адаптивные интервалы), основанный на теореме Котельникова/Шеннона, как более предпочтительный по сравнению с интервалами равной длины.
* Демонстрируется запуск системы, ввод данных из Excel, автоматическое создание моделей.

**3. Интерпретация результатов и оценка моделей:**

* Представлены основные результаты анализа: корреляционная матрица (абсолютные частоты совместной встречаемости), матрицы условных и безусловных вероятностей, матрица количества информации (в битах), показывающая силу и направление влияния факторов на результаты.
* Вводится понятие оценки достоверности моделей с использованием метрик: Истинно-позитивные (TP), Истинно-негативные (TN), Ложно-позитивные (FP), Ложно-негативные (FN) решения.
* Обсуждается стандартный критерий оценки – F-мера Ван Рисбергена, его ограничения (моноклассовость, нечеткость не учитывается, зависимость от объема выборки).
* Предлагается авторское обобщение F-меры: мультиклассовое, нечеткое, учитывающее степень уверенности системы и инвариантное относительно объема выборки.

**4. Практические аспекты и дальнейшие шаги:**

* Лектор советует студентам подготовить свои данные (реальные или модельные) для следующего занятия.
* Кратко упоминается инцидент с отказом публикации теоремы Котельникова и аналогичная ситуация с Эйнштейном, подчеркивая важность новых идей, даже если они не основаны на "достижениях современной науки".
* Занятие завершается планом продолжить рассмотрение примера на следующей лекции.

**Детальная расшифровка текста**

**Введение:**

Здравствуйте, ребята.

Здравствуйте.

Да, здравствуйте, здравствуйте.

Ну что, у нас сегодня по плану численный пример решения задачи в системе Эйдос.

Давайте мы и рассмотрим численный пример.

Интересно, что у нас там в главе пять написано? Давай посмотрим.

В руководстве.

Так. Ну, давайте рассматривать.

Значит, вот у нас есть тут пример из области, которая вам более понятна будет. Это системно-когнитивный анализ силы и направления влияния морфологических свойств помидоров на количественные, качественные и финансово-экономические результаты их выращивания и степень детерминированности этих результатов в условиях неотапливаемых теплиц юга России.

Давайте этот вопрос и рассмотрим.

Там у нас страница 276. Но сейчас я сейчас выйду на свой сайт и найду эту статью просто. Будем по этой статье осваивать этот материал.

Совершенно аналогично и с другими культурами.

Так, это у нас… Так, значит, у нас 150-й номер. 150-й номер журнала. И там пятнадцатая статья.

И в ней есть модель.

Значит, что мы делаем? Мы сейчас сразу стираем систему Эйдос, ту инсталляцию, которая была. Копируем эту модель.

И запускаем систему Эйдос, ребята.

Значит, вы всё видите, да?

Когда мы её запустили, смотрите, что происходит. Сразу скачивается обновление, потому что это давно делалось. И запускается на развора… на разархивирование для всех файлов. Заменяем все файлы.

И смотрим, что у нас тут такое.

Запускается сразу сайт этот. Ну, это не обращайте внимания. Сейчас, сейчас я вам пошлю статью эту, которая… в чат пошлю, по которой мы сейчас рассматриваем вопрос.

Вот, можете ознакомиться с ней.

**1. Подготовка и формализация данных**

**1.1. Описание исходных данных примера**

И смотрим, какие у нас исходные данные для решения этой задачи. Значит, исходные данные – очень маленькая табличка, в которой есть… То, что жёлтым фоном – это выделены результаты влияния факторов. А вот это – сами факторы.

И мы видим, что у нас факторы есть и в текстовой форме написано. Это тоже текстовая форма, хотя вроде на вид как числа похожи, но там, видите, два числа через тире – это текст.

Вот. И потом результаты выращивания – это вот уже все числовые данные у нас.

И мы здесь видим, какие у нас есть гибриды и контрольная группа.

То есть совсем небольшая формочка.

Значит, я вам советую, ребята, подготовить к следующему занятию свои данные какие-то. Или прямо реальные, если они у вас есть, или какие-то модельные.

Что означает модельные? Ну, наподобие ваших данных, но где-то взятые в учебных пособиях, там, где-то в интернете. Понятно? Или вообще даже можно просто взять и придумать эти данные.

Вот. И посмотреть, как получается. Выявляет ли система зависимость или нет.

**1.2. Этапы формализации предметной области**

Значит, сначала нужно определиться, как мы… Да, во-первых, надо ввести эти данные в систему, произвести формализацию предметной области.

Формализация предметной области включает в себя следующие этапы работы.

Это создание справочников факторов и их значений и результатов их влияния. А потом кодирование исходных данных с помощью этих справочников.

Но сначала, на самом первом этапе, который не автоматизирован, вот этот на синеньком фоне, мы должны решить, что мы будем в качестве факторов рассматривать, а что в качестве результатов их влияния.

Вот. Потом мы вот эту осуществляем формализацию предметной области, то есть разрабатываем справочники и кодируем исходные данные.

**2. Системно-когнитивный анализ и обработка данных в Эйдос**

**2.1. Общий план системно-когнитивного анализа**

Потом создаём модели, проверяем, какая из них наиболее достоверна. И в наиболее достоверной модели решаем задачи идентификации, прогнозирования, принятия решений и исследования моделируемой предметной области.

Вот наш план. Это вообще план любого исследования в системе Эйдос. Он соответствует этапам системно-когнитивного анализа.

Можно эти этапы ещё описать следующим образом: что у нас есть исходные данные, и эти данные, они вообще, скажем так, не анализировались, и не очень понятно, что они там означают, не совсем ясно.

**2.2. От данных к информации и знанию**

Потом мы начинаем анализировать эти исходные данные, выявляем причинно-следственные события между тем, что мы приняли как факторы, и тем, что рассматриваем как результаты их влияния.

Ну, допустим, так, грубо говоря, берём мы растение помидор, да? И смотрим, какие у него свойства. Ну, смотрим на него просто и изучаем его свойства. И возникает вопрос: а какие будут результаты выращивания этого растения? Или на участке, например, или там где-то в теплице или так далее.

И этот вопрос, чтобы на него ответить, нужно выявить зависимости между морфологическими свойствами растения, то есть теми, которые наблюдаются внешне, на макроуровне, без особых там приборов каких-то. Ну, может быть, там линейка нужна, там, не знаю, такое вот что-то. То есть простым способом устанавливаются эти свойства.

Вот. И потом выявляются, какие связи между этими свойствами и результатами выращивания. В результате, если мы это сделаем, то у нас уже исходные данные получаются осмысленными. То есть мы уже не просто на них смотрим, как на некие числа там и тексты, а мы уже понимаем их смысл. А смысл в данном случае заключается в чём? В том, что мы понимаем, что будет в результате. То есть какие будут последствия. То есть когда мы понимаем последствия, значит, мы понимаем смысл.

Вот.

А если мы не понимаем последствий, что-то делаем, а не знаем, что в результате получится, значит, мы не понимаем смысл того, что мы делаем. Бездумно действуем.

Вот. И потом возникает вопрос. Мы уже знаем зависимости, которые… Да, у нас, когда появляется в данных смысл, то данные преобразуются в информацию. Возникает вопрос: а информация – это данные или нет?

Вот здесь вот по этой схеме видно, что информация – это данные, но не все данные, а только осмысленные данные. А смысл – это знание причинно-следственных зависимостей в событиях, между событиями, которые отражаются этими данными.

А данные не всегда являются информацией, только осмысленные данные.

Вот. То есть понятно соотношение.

И вот если мы эти зависимости попытаемся использовать для достижения цели, допустим, мы ставим цель – произвести качественную продукцию. Или ставим цель – произвести большой объём продукции. А мы знаем, как влияют морфологические свойства на объёмы производства и на качество. И мы тогда будем выбирать такие растения для селекции – там у нас идёт речь о селекции, да, я так понимаю, гибриды там всё – которые эти свойства обеспечивают.

И таким образом мы решим задачу поддержки принятия решений. То есть что нам нужно сделать, чтобы получить высший урожай, если мы это добиваемся это путём селекции и выбора растений для следующего поколения. Тогда нужно выбирать вот такие-то вот растения. Ну, можно просто по факту, сколько там они дали урожай. А можно на каком-то этапе ещё урожая нет, а мы уже можем сказать, что эти растения дадут такой-то урожай. Вот так примерно.

Ну вот примерно смысл того, что мы делаем.

**2.3. Анализ таблицы исходных данных**

И вот смотрим на эту таблицу, ребята. Посмотрите внимательно, что здесь у нас написано в колоночках. Я бы хотел, чтобы вы увидели, что там написано. Сейчас я попробую.

Вот. Значит, смотрите, у нас здесь есть колоночка: прибыль, рентабельность, общая урожайность (килограмм на метр квадратный), общая урожайность в процентах от контрольного примера, ранняя урожайность (в метрах на… на килограммах на метр квадратный и в процентах от контрольного примера), доля стандартных плодов, средняя масса плодов, содержание в плодах сухих пластических веществ, содержание в плодах сахара, содержание в плодах витамина С в процентах. То есть здесь у нас есть фактически… Сейчас я попробую их выделить. Вот это вот у нас экономические результаты выращивания. Это вот у нас урожайность. Это у нас… ну давай вот так как-то, я не знаю, что здесь. Вот. Урожайность – это у нас количественные результаты выращивания. А вот эти параметры – это у нас качественные результаты выращивания.

То есть у нас есть финансовые, количественные и качественные результаты выращивания.

А что в качестве факторов?

Количество локул в плодах, дней до цветения первого соцветия, второго, третьего, четвёртого, соотношение количества листьев к количеству соцветий, количество соцветий, количество листьев на одном… на стебле, количество листьев на одном погонном метре стебля, высота главного стебля, длина междуузлий, э-э, степень детерминированности… А нет, извините, степень детерминантности. И характер расположения соцветий на главном стебле.

**2.4. Подготовка Excel-файла для импорта в Эйдос**

Так, ребята, вы, пожалуйста, у нас сейчас лабораторная работа. Хотя вообще-то лекция, но я вот на примере… У нас же лекция – численный пример. Поэтому, если у вас возникают вопросы, то можете с помощью, значит, голоса, микрофона мне говорить, задавать эти вопросы. Я буду вам пояснения давать.

То есть мы видим, что у нас морфологические свойства, то есть внешние свойства, наблюдаемые на макроуровне без специальных каких-то особых усилий. Ну, ещё можно было бы какие-то характеристики генома здесь привести. Тоже было бы интересно, как связаны характеристики генома с прибылью, например. А это… они связаны, потому что геном определяет как раз вот эти морфологические свойства и, по сути дела, влияет на эти вот финансовые, количественные и качественные результаты выращивания.

Теперь смотрите, что нам нужно сделать. Чтобы ввести эту информацию в систему, э-э, нам… Мы уже проходили какие-то лабораторные работы, я вам показывал, да? А теперь мы вводим данные не из лабораторной работы, а прямо из своего файла.

Для этой цели нужно этот файл поместить в папочку, предназначенную для исходных данных.

Значит, папочка, предназначенная для исходных данных, на моём компьютере вот так путь на неё. Ну, начинается с диска C, но у вас он может начинаться с любого диска, но потом должна быть папочка системы, потом папочка данных в системе, потом папочка indata.

Вот, мы туда этот файл записываем, в эту папочку. У него должно быть имя определённое. Это имя input\_data либо .xls, либо .xlsx. Ну, разница в том, какой вы используете Excel.

Я могу вам сказать, что новый Excel, начиная с 2007 года, он больше строк обеспечивает, до миллиона, и больше колонок, до 16.000. А старый – 65.000 строк 536, а 256 колонок всего лишь. Если мы один и тот же файл запишем в старом и новом Экселе, то в новом он будет иметь размер, ну, примерно в два раза меньше. По-разному там получается, но заметно меньше.

Вот. И смотрим теперь на этот файл. Какие у него колоночки? Классификационные шкалы – вот это классификационные шкалы, потому что они классифицируют результаты действия факторов. Это описательные шкалы, они описывают ситуацию либо свойствами, значениями свойств, и либо факторами, значением факторов. Колоночка у нас – это свойство, какое-то свойство определённое. Ну, допустим, высота главного стебля – это свойство. А значение 108,3, 124,0, 185,6. Здесь уже усреднённые данные по этому гибриду за период наблюдений.

Ну, точно так же может быть и фактор. Это можно как фактор рассматривать, влияющий на результаты. Но я думаю, что в данном случае правильно рассматривать как фактор. Почему? Потому что это было во времени в одно время, а это потом уже позже гораздо, после получения урожая.

**2.5. Ввод данных в систему Эйдос (Режим 2.3.2.2)**

Теперь давайте зайдём в систему, и я вам показываю, где режим, который позволяет ввести эти данные. Это режим 2.3.2.2. Вот этот режим 2.3.2.2, я его называю универсальный программный интерфейс ввода данных в систему Эйдос из экселевских таблиц. Данные могут быть и числовые, и текстовые.

Что касается самой таблицы. Давайте я немножко вам дам пояснение. Значит, таблица должна быть выделена. То есть если мы ставим на клеточку А1 курсор, нажимаем Shift+Ctrl и нажимаем потом End, то выделяется область данных, видите? Эта область данных должна точно соответствовать фактической области данных. Если она не соответствует, вылазит куда-то вот сюда, сюда, то система будет выдавать всякие сообщения, что колонка наблюдается без наименования заголовка. Заголовок сделаете, она напишет: колонка без данных. Ну, в общем, будет вот там всё вам рассказывать.

Вот. Чтобы вот так сделать, нужно просто выделить блоком вот так таблицу. Вот. И потом скопировать её в новый лист. Вот. И всё. Вот, собственно говоря, что требуется сделать, чтобы у вас эти данные были теперь точно по таблице.

Следующий момент. Значит, вот здесь у нас в экранной форме спрашивается, какой тип данных? Ну, я здесь задал тот, который там фактически. Теперь смотрите, значит, у нас здесь задаётся диапазон классификационных шкал. То есть это результаты, шкалы, описывающие результаты действия факторов. Нам нужно посчитать, какая первая колонка с результатами, какая последняя. Но можно это не считать. Если таблица появляется у нас целиком на экране, тогда можно поставить курсор в клеточку А1, нажать кнопочку Shift и двигать. Вот я вправо сдвинул стрелочкой, видите, здесь написалось 1R 2C, да? То есть одна первая Row, вторая Columns – это вторая колонка. Вот я двигаю, колонка меняется.

То есть сейчас вот я вижу, со второй по двенадцатую – это у нас классификационные шкалы.

А с двенадцатой… с тринадцатой по двадцать пятую – это описательные шкалы.

То есть со второй по двенадцатую классификационные, а с тринадцатой по двадцать пятую описательные.

Вот и всё. То есть это та информация, которую нам нужно ввести в систему. Значит, если мы… если таблица очень большая, сейчас вот мне тут коллеги спрашивают там, когда 307 описательных шкал, то есть гораздо больше размерности задачи, то тогда есть смысл взять вот просто добавить строчку и пронумеровать колоночки. Вот. 1, 2, 3, вот просто пронумеровать. Только не вручную, а вот так поставили и тянете её. Вот. И мы видим, что со второй по двенадцатую классификационная, с тринадцатой по двадцать пятую описательная. Но, значит, э-э, вот так делать надо не в том листе, который мы будем вводить в систему, а в каком-то там другом месте. То есть где-то взять этот файл, записать в какую-то папочку, там это делать. Почему? Потому что если это сделать здесь, то система будет читать первой… первой строкой, то есть первую строку она считает строкой наименований. И она будет считать, что вот эти 1, 2, 3 – это и есть наименование колонок.

**2.6. Настройка параметров импорта и квантификации**

Значит, теперь мы задаём вот этот… эти диапазоны: со второй по двенадцатую, диапазон и с тринадцатой по двадцать пятую.

И здесь, я вам могу сказать, что… Помните, я вам рассказывал, или не знаю, помните вы или нет, было это вообще или нет, не знаю, если честно, с вами именно конкретно.

Значит, есть два вида интервалов числовых. Если у нас есть числовые шкалы, то можно двумя способами разбивать их на числовые интервалы. Один способ – одинакового размера интервалы, но тогда в них будет разное число наблюдений. И второй способ – это разные по размеру интервалы с примерно одинаковым числом наблюдений, плюс-минус один. Почему? Потому что наблюдение – целое число, поэтому может отличаться на единицу. Но я советую такой вариант, это наиболее хорошие модели получаются именно когда… Я их называю адаптивные интервалы. Это теоретически хорошо обосновано.

Ну, могу вам в двух словах объяснить, как это обосновано теоретически. Вот представьте себе, что у нас есть некая кривулечка, и нам надо эту кривулечку ввести в систему.

Вот если мы будем брать точки отсчёта, измерять её вот здесь вот, вот здесь, вот здесь, то у нас вот эти точки получатся. Вот. И мы можем их соединить, и получается аппроксимация этой кривой. А теперь представьте себе, что мы так вот точки дальше с такой… через такое же расстояние измеряем, ну через равное расстояние. У нас что получается? Что здесь вот у нас точек избыточное количество. Вот. То есть больше, чем нужно. Потому что очень низкая вариабельность этой кривой, то есть она почти прямая линия. А вот здесь у нас получится слишком редко эти точки. Вот одна, вот вторая, вот третья, четвёртая. То есть если мы их соединим эти точки, то у нас получится не совсем то, что хотелось бы.

Что у нас получится? У нас получится, что всё, что касается вот этих быстрых изменений, всё это пропадёт, понимаете?

Вот. То есть мы вот эти все гребешочки, этот вот, высокие частоты, так называемые, их потеряем. У нас их не будет в результате.

Отсюда такой вывод. Что нужно сделать? Что сделано в системе Эйдос? Это предложил вообще-то академик Котельников эту идею. Он высказался таким образом, что там, где у нас вариабельность этой кривой низкая, которую мы хотим оцифровать, там можно реже ставить точки. Вот. А там, где вариабельность высокая, там нужно гораздо чаще ставить, вот так вот прямо вот… э-э… прямо вот часто вот так ставить, потому что иначе мы вот эту гармонику высокочастотную потеряем.

Это и называется теорема Котельникова. За рубежом её называют теоремой Шеннона-Найквиста-Шеннона. Почему? Потому что академик Котельников, когда был совсем ещё молодым парнем, э-э, доказал эту теорему, написал статью о ней и не смог её опубликовать. Ему все рецензенты того времени, корифеи, доктора наук, профессора в области теории связи, они все ему дали отзыв отрицательный. В журналах были рецензенты. Сказали, что какая-то ахинея, вообще ничего не поймёшь. И не стали публиковать. И он попытался, два или три журнала попытался, везде ему дали полный отказ. И он после этого перестал… прекратил попытки и не стал публиковать. Ну, так она осталась неопубликованной. Он на какой-то конференции выступил, но это оказалось тоже неопубликованным, то есть её исключили из сборника.

Это известная история. В конце концов он плюнул на это всё, ну так я понимаю. Я так думаю, что он очень сильно расстроился. Вот. Ну и бросил эти попытки, понимаете, перестал пытаться опубликовать эти результаты. А через 15 лет Шеннон, э-э, я не знаю, сам он догадался или где-то узнал. Ну я думаю, что он был человек умный, и сам мог догадаться. Ну, факт такой, что через 15 лет эту теорему опубликовал Шеннон, и она стала называться теоремой Шеннона, понимаете?

Мы потеряли приоритет в таком чудесном деле. Ну у нас, э-э, как бы… с кровью в сердце, так не знаю, со слезами в сердце, её называют всё-таки теоремой Котельникова. Но я могу вам сказать, он не смог её опубликовать. То есть она в печати вышла уже потом, когда он стал академиком, там он опубликовал какие-то свои записки, там было написано, что вот в таком-то году там. А так фактически она миру не была известна до того, как Шеннон опубликовал свои работы. Она была известна только этим вот ретроградным рецензентам.

Кстати, с Эйнштейном была та же самая история. Он когда свою статью принёс об электродинамике движущихся сред в журналы, то ему стали все журналы отказывать, говоря, что это какая-то вообще ахинея, там вообще… Писали, знаете как интересно? Не основано на достижениях современной науки. Рецензенты, рецензенты писали: не основано на достижениях современной науки. А знаете, почему она не была основана на достижениях современной науки? Как вы думаете? Вот ответьте мне. Почему то, что принёс Эйнштейн в редакцию, не было основано на достижениях современной науки? Это так и есть, действительно. Ну так что ж, теперь не публиковать, что ли, если она не основана на достижениях современной науки? Вот тогда приняли решение не публиковать.

Вот если так вот публиковать только те статьи, которые основаны на достижениях современной науки, тогда эта наука двигаться вперёд не будет, понимаете? Почему она… его статья не была основана на достижениях современной науки? Потому что она не относилась к современной науке, она относилась к науке XX-XXI века, понимаете? Я могу вам сказать, сейчас дали недавно Нобелевскую премию за исследование чёрных дыр в центре галактики. Э-э, они движутся совершенно точно по теории Эйнштейна, которая не относится к достижениям… не основана на достижениях современной науки. Тогда, в 1905 году она не относилась. Уже двенадцатая Нобелевская премия за его работы, ребята, за их продолжение. То есть, ну, вообще нет другого случая такого, чтобы так много было Нобелевских по какому-то направлению определённому научному. То есть работы Эйнштейна привели к тому, что со временем… Он предсказал гравитационные волны, к примеру, 100 лет назад. Через 100 лет дали за… за их открытие, физическое измерение их, дали Нобелевскую премию. И вот так вот… Лазеры. Мало кто знает, но теория вынужденного излучения разработана Эйнштейном. Потом лазеры создали и, э-э, по теории Эйнштейна работающие, я подчёркиваю. И дали за это Нобелевские премии. Ну и так далее, и так далее. Там много можно перечислять. Э-э, начиная с того, что Ричарду Фейнману дали Нобелевскую премию за объединение специальной теории относительности, которую разработал Эйнштейн, с квантовой механикой, которую разработал Эйнштейн. Вот Ричард Фейнман объединил их и получил за это Нобелевскую премию. Называется квантовая теория поля. Вот. А сам Эйнштейн получил Нобелевскую премию за создание квантовой механики, потому что он первым в истории провёл расчёты с использованием того аппарата, который сейчас мы называем квантовая механика. Ну тогда она так не называлась, когда он это сделал. Но вот… но он первые реальные расчёты провёл с использованием аппарата квантовой механики, который сам он разработал для того, чтобы провести эти расчёты. И вот так и так далее, и так далее. Так вот, если мы будем основываться только на достижениях современной науки, тогда этот наука будет вечно, понимаете, современная. То есть она не будет меняться.

Поэтому, значит, ему отклонили в ряде журналов отклонили. А один журнал не отклонил, называется Nature. В то время этот журнал был малоизвестный научный журнал, такой периферийный. Сейчас это самый известный в мире журнал. Почему? Потому что они опубликовали статью Эйнштейна, но приписали там фразу, которая сейчас является уже во всех журналах эта фраза написана, что статья эта публикуется, потому что она интересна, в общем-то, но результаты, то есть выводы… результаты, выводы работы этой… работы автора могут не совпадать с точкой зрения редакции. Вот так вот они написали, что мнение автора может не совпадать с точкой зрения редакции. Ну то есть мы его как бы, говорится, просто опубликовали, потому что интересно, но мы сами так не думаем, но вот вообще-то вот интересно, поэтому опубликовали.

**2.7. Запуск процесса обработки данных**

Вот. Значит, теперь дальше смотрите, ребята. Значит, мы указали диапазон классификационных шкал, диапазон описательных шкал, указали тип данных. И теперь дальше идём, двигаемся. Указали адаптивные интервалы, указали, что нули и пробелы считать отсутствием данных. Но там их нету, но бывает, что данные такие фрагментированные, там где-то отсутствуют данные. Тогда, если числовая шкала, там нули надо писать, а если текстовая, тогда колоночка, я имею в виду, в таблице, тогда надо пробелы там, чтобы было пусто там в ячейке. Теперь смотрите, значит, здесь есть ещё что? Новый Excel, DBF файл и CSV DBF конвертер. Что это такое? Значит, Excel новый там не требует комментариев. DBF – это стандарт, который мало кому известен. Но если он неизвестен, то тогда и можно не читать, что это такое. Вот. Но, в принципе, это экселевский файл, э-э, ещё более древнего Ворда, то есть вообще более древнего Экселя, чем 2003. Вот. Ну, первых досовских версий Экселя, когда Windows ещё не было операционной системы. Вот тогда вот такие были экселевские файлы такого… не с таким расширением, но такого стандарта, вот так скажем. А вот CSV DBF конвертер – это режим, который я сделал вчера, позавчера, ещё его буду совершенствовать. Это режим, который обеспечивает преобразование форматированных текстовых файлов в DBF формат и запуск сразу же этого DBF… формата для… ну, в смысле, ввода данных из этого DBF формата. CSV формат – это текстовый файл, в котором поля, э-э, которые в Экселе являются колонками, разделены запятой просто-напросто. Вот, значит, я этот режим сделал. Он работает. Какой смысл в этом режиме? Почему… Зачем он нужен вообще? Значит, дело в том, что Excel, даже новый, позволяет не более тысяч… миллиона строк. Миллион строк. А DBF файл, он не ограничен по числу строк, он ограничен по размеру, то есть файл не может быть больше 2 ГБ по размеру. Значит, если взять этот размер и взять такой среднестатистический э-э файл с таким средним числом колонок, то это примерно 10-12 млн строк. То есть это значительно больше, чем в Экселе, понимаете? Вот. То есть намного больше, на порядок больше, может быть, даже. Вот. Так что может иметь смысл. А CSV файл, как текстовый файл, они имеют неограниченный размер. Они могут быть и там 10-30 ГБ, например, файл, понимаете, один. Вот. Не мегабайт, а гигабайт. И вот можно такой файл обрабатывать. Вот. Если э-э система будет его вводить и получится, что уже 2 ГБ достигнуто э-э DBF файла, то на этом процесс прекращается и сообщение выдаётся, что вот было такое вот ограничение э-э сработало. И дальше потом этот… эти данные, которые удалось ввести, вводятся в систему.

**2.8. Результаты обработки: Справочники и обучающая выборка**

Значит, дальше я сейчас нажимаю О'кей. И что происходит? Файл сам этот преобразуется во внутренний формат данных DBS в системе. И пишет система рекомендацию, что необходимо либо уменьшить число градаций в описательных шкалах из-за недостатка данных, появляющ… появляются интервалы без наблюдений. Удалить шкалы либо вообще без наблюдений, то есть без вариабельности наблюдений, либо с очень малым числом наблюдений. И здесь вот у нас информация о том, как там у нас что получилось. Значит, эта информация может выводиться и на экран.

Вот у нас получается, что… Ну давайте здесь вот тоже три сделаем. Пересчитаем.

Значит, у нас здесь информация такая: числовых шкал классификационных – 11 и описательных – 11 числовых. Ещё две описательных есть шкалы текстовых. Матрица получится 33 колонки, соответствующих классам, результатам выращивания, и 37 строчек, соответствующих значениям факторов, которые мы применили.

Смотрим, ребята. Смотрите, у нас сейчас нарисовалось, каких… какие будут диапазоны и сколько в них наблюдений. Классификационные шкалы идут. Мы видим по три наблюдения на градацию. Потом идут описательные шкалы, тоже по три наблюдения на градацию. Градации разного размера, исходя из тех принципов, которые я вам говорил, что они разного размера, но число наблюдений одинаковое. И в шкалах текстовых там не пишется, сколько наблюдений на интервальные значения.

Ну вот, значит, с этим мы соглашаемся и идём на создание модели.

Что произошло? Созданы были, разработаны классификационные и описательные шкалы и градации на основе анализа этого файла исходных данных. А потом исходные данные были закодированы с использованием этих классификационных и описательных шкал и градаций. У вас дисциплина э-э базы данных не было, наверное, да? Ну, я вам могу сказать, что эта операция, которая сейчас вот произведена, процедура, называется в науке о базах данных, называется нормализация баз данных. А сейчас мы, по сути дела, что делаем? Мы провели формализацию предметной области, то есть повысили степень… то есть уровень системности модели. То есть у нас были исходные данные неформализованные, числа просто. Вот. А теперь у нас есть справочники событий прошлых и будущих и есть база, которая отражает эти события уже в виде закодированном.

Смотрим, что же у нас там за события? Смотрите, значит, прибыль определённая, минимальная градация… минимальное значение 1,3, максимальное 3,3. Как получены эти диапазоны? Значит, найдено минимальное и максимальное значение, и потом они на такие… такие размеры даны этим диапазонам, чтобы там было одинаковое число наблюдений. Вы видели, что это число наблюдений – это три. У нас все классификационные шкалы, все являются числовыми.

Вот. И смотрим описательные шкалы. Количество локул. Вот эти вот все наименования, которые мы здесь видим слева, наименования шкал, это непосредственно взято из экселевского файла. А вот текстовые у нас значения.

Значит, обращаю ваше внимание, ребята, что вот эти названия всё-таки желательно, чтобы были покороче. Потому что они потом будут в выходных формах присутствовать, и если они будут очень длинные, то может даже не помещаться градации.

**3. Интерпретация результатов и оценка моделей**

**3.1. Модели знаний: Корреляционная матрица и матрицы вероятностей**

И смотрим ещё, есть другой формат обучающей выборки. Вот такой.

Что мы здесь видим? Мы видим абсолютно то же самое, что мы видим в Экселе. Вот. Давайте даже посмотрим.

Вот, смотрим экселевская форма и форма системы Эйдос. В чём различие, ребята? Различие в том, что в экселевской форме мы видим числа и тексты, а в системе Эйдос мы видим коды, соответствующие справочникам этих чисел и текстов.

Что здесь интересного? Вот после этого этапа получения обучающей выборки уже коды и текстов, и чисел, они выглядят совершенно одинаково. Вот. И дальше уже никакого нет различия в обработке чисел и числовой информации и текстовой информации. То есть дальше в системе Эйдос они обрабатываются одинаково.

**3.2. Синтез и верификация моделей**

Дальше у нас по вот этому плану нашему, который мы видим здесь, порядок преобразования данных в информацию, а её в знание. Мы видим, что следующий этап – это синтез моделей.

Сергей Семёнович, я занятие веду.

Ну, я… нету одной пары сегодня. Первая, вторая, третья, пятая. А четвёртой нету. И шестая. А вот четвёртой нету пары. То есть у меня первая, вторая, третья, пятая и шестая пара. То есть я практически всё время сижу за компьютером и рассказываю что-то монитору.

Ой, вы знаете, я как-то против вообще-то. Против этого. Потому что я не тот… не являюсь тем, вот кто вот соответствует этим требованиям. Я сам это вижу, понимаю и думаю, что я не соответствую. Это первое. Второе, ну как, я работаю, как могу, стараюсь всё делать, но я знаю, что я не всё делаю. Вот. И потом второе, что вот этот ЦИТ наш, Центр информационных технологий, они имеют абсолютно полный доступ ко всему абсолютно, что у нас там есть. То есть они зря вот это говорят, что там надо нам имена, пароли. Пусть сами туда входят под именами и паролями. Потом я скажу так, что в ЦИТе работает много преподавателей. Вот тот же самый Алексей Семёнович Кремер, Крепышев Дмитрий Александрович, они все преподаватели. Ну пусть они свой дадут пароль, тот же самый Алексей Семёнович, и, пожалуйста, пусть смотрят там они всё. Пусть он войдёт как сам, как преподаватель, и показывает. Хотя они как руководи… как они как наши администраторы, они могут войти, я вам заверяю вас, куда угодно, под…

Вот я сейчас аккредитацию мы готовим пароль такой. Ещё один момент, мы должны предоставить логины и пароли преподавателя, который у нас конкретно вот на ЭБСе работает, да, за который я отвечаю. Преподаёт, чтобы они зашли, посмотрели на головной пароль. Да никогда это требование… это то же самое, что требование дать личные данные. Эти данные, имена и пароли для того и существуют, чтобы их никому не давать. Банк когда сообщает вот пароль входа в банк, он пишет там: никому не предоставлять его. Зачем эти вот имена и пароли, если их предоставлять? Их потому и дают, чтобы не предоставлять. Там огромный объём работы в моих этих… Там у меня 54 дисциплины. Вот если сейчас я дам имя и пароль, это возьмут и постирают, удалят эти данные. Я весь год этим занимался, долбился с утра до вечера, понимаете, чтобы это всё делать. Я тут получу имя и пароль, я не знаю теперь, после… после этого я не знаю, что там будет, понимаете? Там может быть что угодно. Там может быть написано название лекции: "Пошёл ты нафиг", например. Название лекции. Понимаете? И что, придёт мне скажет: "Петрух, а что вы там у вас вообще бред знает, что… не знаю, что написано". А я скажу: "Да я не знаю, я отдал имя и пароль". Понимаете? Это нельзя делать вообще. Если… это то же самое, что пойти к бухгалтеру, вот планово-экономический отдел, расчётный отдел и сказать: "Дайте имя и пароль от системы, где начисление зарплаты". Ну вы что, с ума посходили, что ли? Ну, ну дали вам имя и пароль, всё, я могу… я снимаю с себя полную ответственность, так сказать, полностью снимаю с себя ответственность за то, что там будет в системе, понимаете? Там могут нули кому-то дописать, кому-то убрать эти нули. Я уже не несу ответственность за это, понимаете, после этого.

Нам идёт речь на логины и пароли доступа к ЭБС. И вот всякие электронные… А что мы там… а что мы там делали? Нет, не только, там ещё Moodle система, где вся… вся информация методическая по преподаванию дисциплин.

Ну, конечно, я его прочитал. Как на него не ругаться? Я не знаю. Ну, бывают письма руководства, которые как бы странные какие-то, я бы сказал так. Но я не знаю, как… Ну, допустим, если Алексей Семёнович будет спрашивать, то можно ему сказать, что неужели вы не имеете доступа к этим всем ресурсам, как руководитель ЦИТа? А в библиотеке там они сами тоже могут взять, завести там, э-э, зайти в эту систему под своим именем и паролем. Они учат там аспирантов всех. Это делать. Пусть они покажут, как это делается проверяющим, что вот так вот мы даём имена и пароли, входим в эти системы, вот так вот мы заказываем книжки, вот… пусть покажут. Библиотеки сами, работники библиотеки.

Ладно. Буду что-то думать. Спасибо. Угу. Да, до свидания. До свидания.

Вот, ребята. Значит, теперь по вот этому плану нашему мы должны осуществить… Мы выполнили этап формализации предметной области, создали справочники классификационных и описательных шкал и градаций, и с их помощью закодировали исходные данные. Теперь мы по нашему плану должны осуществить синтез моделей и проверку их на достоверность.

Значит, здесь, э-э, вы сейчас видите, ребята?

Видим, видим.

Вот, этот этап – синтез и верификация модели. Синтез – это создание, а верификация – это проверка на достоверность.

Вы мне, ребята, отвечайте голосом, когда я спрашиваю, вы отвечаете.

Видим, видим.

Хорошо. Потому что ситуация такая: я целыми днями, вы уже поняли, да, вот так вот сижу, рассказываю что-то монитору. И к концу начинаю чувствовать себя полным идиотом, потому что я не знаю, сидите вы там или нет, слушаете или нет. Поэтому мне нужно знать, что вы хотя бы там, понимаете? Поэтому я иногда спрашиваю что-то вас, чтобы проверить, что вы меня слышите реально. Вот я экран дал, предоставил вам доступ к экрану, чтобы вы его видели. А я не знаю, смотрите вы на него или нет, понимаете? Я этого не вижу. То есть когда я в аудитории, я вижу, как вы занимаетесь. А когда здесь, я вообще не вижу ничего, я вижу только свой экран. И вижу там отметки, что вы там есть на занятии. Но я прекрасно понимаю, так сказать, с помощью своего естественного интеллекта, что можно вот так соединиться по телефончику и потом идти себе заниматься своими делами спокойно, понимаете?

Можно, но мы сидим, слушаем.

Ну хорошо, хорошо. Очень приятно.

Вот. Вот если бы я видел все ваши лица, так сказать, на экране, что вы там слушаете, смотрите, ну было бы как-то, может, приятнее. Вот. Но, э-э, тогда было бы тяжелее, конечно, всё рассказывать, потому что было бы на экране меньше места. Ну, наверное, это было бы более правильно.

**3.3. Оценка достоверности моделей: Метрики и критерии**

Значит, смотрите, ребята, значит, эта форма, она показывает, что мы сначала разрабатываем три модели статистических: матрицу абсолютных частот и матрицы условных и безусловных процентных распределений. А потом семь моделей знаний, системно-когнитивные модели. Вот. При этом мы можем, если очень большая выборка, можем разделить операции на только синтез модели, только верификация. А сейчас, когда модель маленькая, мы синтез и верификацию проведём. Сейчас мы будем использовать графический процессор. А теперь дальше слушайте внимательно. Вот здесь внизу написано, что вот эта модель, которую сейчас мы обрабатываем, составляет собой 7 десятитысячных процента от того, что может обрабатывать система Эйдос. Вы слушайте, 7 десятитысячных процента. Понятно, да? То есть она может обрабатывать значительно большие объёмы информации. А если, значит, э-э, превышается эта возможность системы, то вот здесь вот мы можем задать, скажем, 5%, не 100, а 5%. И будет 5% наиболее значимых результатов идентификации обрабатываться, остальные будут игнорироваться. А существует прямая связь между… То есть система может оценивать степень достоверности своих решений, короче говоря. И эта оценка очень высокую степень адекватности имеет, то есть она прямо вот правильно определяет это. И получается так, что мы можем оставить наиболее достоверные результаты, тогда можно гораздо большие объёмы данных обрабатывать. Если очень большая выборка, то можно создать модель на всей выборке, а потом проверить её достоверность на каком-то подмножестве. Ну, скажем, там 100 объектов, 1000 объектов выборки. Ну такое дело.

Значит, сейчас модели сначала создаются, а потом проверяются на достоверность.

Способ проверки на достоверность, ребята. Значит, когда мы в системе с помощью модели решаем какие-то задачи, задача, допустим, идентификации, прогнозирования, принятия решений, то мы можем решение любой задачи использовать для оценки достоверности модели. Допустим, мы взяли объект, отнесли к каким-то категориям, идентифицировали, классифицировали их. И посмотрели, сколько правильно, а сколько ошибочно. Это проверка достоверности модели. Второй вариант. Взяли, спрогнозировали что-то, оно что-то там осуществилось, что-то не осуществилось. Посчитали, сколько осуществилось, сколько не осуществилось. Оценили достоверность модели. Взяли, э-э, приняли решения какие-то управляющие о том, чтобы перевести объект моделирования в нужные целевые состояния, заданные. Вот. Потом сделали это всё, осуществили эти решения и смотрим, сколько раз объект перешёл в целевое состояние, сколько нет. Определили достоверность модели. Но из всех этих вот вариантов, э-э, легче всего проверить идентификацию. То есть относится объект к классу или нет. Почему? Потому что и прогнозирование, и принятие решений требуют времени для проверки. Идентификация ничего вообще не требует. То есть у нас есть уже полностью обучающая выборка, просто можно её использовать и всё для этой цели, для проверки достоверности. И там есть информация о том, к каким классам относятся объекты. Прогнозирование требует времени. То есть мы спрогнозировали, и потом смотрим, что получится. То есть это более сложно. А принятие решений ещё сложнее, потому что нужно что-то ещё делать, нужно оказывать воздействие управляющее на объект моделирования, на объект управления, и потом смотреть, какие последствия. То есть это ещё сложнее, потому что нужно реально осуществлять управляющее воздействие на объект управления. То есть это вообще более проблематично. Но, в принципе, да, вот если мы систему эксплуатируем в реальном, так сказать, образе, то тогда можно, конечно, и прогнозирование использовать для оценки достоверности, принятия решений. Но это в случае, если система находится в эксплуатации, и там эти задачи постоянно решаются. А вот в таких вот учебных целях, конечно, задача идентификации наиболее удобна для этих целей.

**3.4. Результаты верификации и интерпретация метрик**

Модели созданы, ребята. Теперь мы можем на них посмотреть. А что ж там за модели-то созданы?

Вот это модель называется корреляционная матрица или матрица абсолютных частот. Здесь строки – это значение факторов, то есть какой-то фактор и его значение. А колонки – это результаты действия этих факторов, это градации классификационных шкал. Здесь их довольно много у нас. 33 градации классификационных шкал. А факторов – 37.

И на пересечении вот здесь мы видим, сколько раз то или иное значение фактора встретилось в том случае, когда объект моделирования перешёл в состояние, соответствующее тому или иному классу. И потом смотрим, сколько всего было примеров по каждому классу. Здесь в данном случае было по три примера, я смотрю. Ну не всегда. Вот здесь вот один пример был. Здесь вот пять примеров, четыре примера, то есть так вот. Ну, примерно, значит, ну разное количество, но так вот… не один раз редко такие. Ну и такие один раз есть. Теперь смотрите, у нас здесь указано, что у нас суммарное число объектов выборки – 99. Почему так получилось? Потому что мы-то знаем, что у нас объектов-то гораздо меньше. Вот. Но как меньше? Ну давай посмотрим сейчас. У нас всего-то э-э девять строчек в таблице. Видите? Ну что ж это 99 получилось? А вот смотрите. Первая колоночка классификационная, вторая, третья, четвёртая, пятая, шестая, седьмая, восьмая, девятая, десятая, одиннадцатая. Вот умножьте 11 на 99… или 99… 9 умножьте на 11. У вас получается 99. Что это значит? Что у нас физических… на физических наблюдений девять, но каждое физическое наблюдение относится к… к одиннадцати классам. Этот, этот, этот, этот, этот. То есть можно считать, что одно вот такое наблюдение, на самом деле это 11 наблюдений. Если так рассматривать, тогда получается, что у нас 99 наблюдений. И вот… Подскажите, вот что в этих девяти колоночках нужно писать? Примерно, что это такое, что обозначает? Это название гибридов помидоров, сортов, новых сортов. Это новые сорта. Просто что если в наших примерах вы нам задали, что писать, допустим? Ну, у вас может быть… это это вообще-то строчка – это наблюдение. Наблюдение может быть как наблюдение прямо вот самого растения, к примеру, конкретного, так и может какая-то быть усреднённая информация по сорту, например, или по району, или ещё там что-то, по саду. То есть это уже зависит от того, какую мы задачу решаем. Ну… В задаче вот этой может быть и не девять сортов, да? Не девять колоночек, а 29. Может быть, этих строчек может быть до миллиона, а колоночек до 16.000. Спасибо.

Значит, теперь что здесь интересно ещё во всём этом? Что когда мы посчитаем всё в количестве информации, то уже единицы измерения, в которых измерялись эти вот сами параметры, которые у нас в шкалах, то есть значения, значения факторов, они там у нас одни в процентах, другие в миллиграммах там, третьи там ещё в чём-то там. Вот. Миллиграммах и процентах. Я такого даже не знаю, такой единицы измерения. Вот. Уже эти единицы измерения перестают играть роль. Вот, скажем, стоимостные единицы измерения, рубли, например, там, да? А здесь у нас проценты какие-то, там килограммы на метр квадратный, а тут проценты. Вот как мы это всё вместе обрабатываем? А потому что, вот, ребята, нас интересует только одно обстоятельство: какое количество информации содержится о том, что сколько-то там рублей, там, или сколько-то там процентов, или сколько-то там килограмм на квадратный метр в том, что действует такой фактор. То есть нас везде интересует количество информации. То есть мы всё преобразовали в количество информации. Все эти вот единицы измерения, они все исчезли, стало количество информации. После этого вполне корректно всё это сравнивать.

Вот. Сейчас я вам покажу статью, которая об этом… написано про типы шкал, про модель и подробно объясняется вот этот момент, который сейчас вы спросили. Почему измерение происходит в количестве информации? Почему на этом всё основано? Это обеспечивает сопоставимость обработки различных данных, которые были представлены изначально в различных единицах измерения.

То есть мы вот получили модели, которые не зависят от того, числа там или тексты, от того, сколько объектов представлено по той или иной категории обучающей выборки, с теми или иными результатами выращивания. То есть сколько объектов по классам, мы от этого уже отошли, потому что мы взяли относительные величины. А потом отошли ещё и от единиц измерения, когда всё преобразовали в количество информации или хи-квадрат.

**3.5. Оценка достоверности моделей (продолжение)**

И теперь мы хотим посмотреть, какова достоверность модели. Ну давайте сначала э-э зададим название модели. Вот видите, здесь написано: приложение создано путём ввода данных из источника данных внешнего. Вот я беру, копирую отсюда. Надо русский регистр включить. Да, получается копировать.

Вот. И у нас это приложение, которое сейчас нами создано, оно вот третье, последнее. Вот. Вот здесь у нас есть Help, где описано вот то, что я сейчас вам рассказывал, здесь всё это по пунктикам всё расписано. Сейчас мы подошли к пунктику 5.5, что мы модели посмотрели. А теперь к пунктику… извините, к пункту шесть. Мы рассмотрели модели в режиме 5.5. А теперь подошли к пункту семь и будем рассматривать, какова достоверность модели. Что такое вообще достоверность?

Вот такую форму получаем. Значит, сейчас мы э-э сразу посмотрим… помощь по описанию того, как оценивается достоверность решения задач в системе Эйдос. Ну я вам говорил, что можно решать э-э оценивать достоверность модели путём её применения для решения различных задач. Но мы это сделали для… на примере задачи идентификации. Вот. То есть мы по морфологическим признакам относим э-э эти наблюдения к тем или иным классам, э-э которые у нас там являются градациями классификационных шкал. То есть это финансовые результаты, количественные результаты и качественные результаты. И модель должна что обеспечивать? Она должна обеспечивать правильное отнесение наблюдения к тем классам, к которым оно относится, правильное неотнесение к тем, к которым оно не относится. И она может ошибаться модель и система на ней основанная, реализующая эту модель, может ошибаться и в первом случае, и во втором.

Значит, э-э теперь нужно вам запомнить немножко терминов, ребята. Вот какая терминология тут сейчас потребуется вам.

Значит, если система оценивает уровень сходства какого-то наблюдения, объекта обучающей выборки, э-э что он больше нуля при сравнении этого объекта выборки обучающей с классом, с обобщённым образом класса, то тогда мы говорим о том, что это решение позитивное, положительное решение. Если же уровень сходства меньше нуля, тогда это называется решение отрицательное. И эти и положительные, и отрицательные решения, позитив-негатив, они могут быть истинны и ложные. Ну, может быть, вы знаете, что истина – это True обозначается, ну, в программировании, в английском языке True – это истина. И False – это ложь.

Соответственно, у нас… если обозначить True, False – это первая буква, а позитив, негатив – вторая буква, то у нас получается, что все решения, которые может система на основе модели э-э принять, их можно классифицировать как True Positive, True Negative, False Positive, False Negative. То есть True Positive – это система относит объект к классу, и действительно он относится к этому классу. True Negative – система не относит объект к классу, то есть уровень сходства меньше нуля. И действительно он к нему не относится. И она может ошибаться. Может быть False Positive. Что это значит? Это система отнесла объект к классу, а на самом деле он к нему не относится, к этому классу не относится этот гибрид 1907. Но при этом система-то отнесла, конечно, но она это отнесла э-э… это вот если рассматривать уровень сходства как степень уверенности, то она сама очень сильно сомневается о том, что он относится к этим классам, этот гибрид. Всего лишь 1% степени уверенности из 100.

Вот. И ещё она допустила ложноотрицательное решение. То есть на самом деле фактически этот гибрид относится к прибыли максимальной, а она его не отнесла к этой категории. То есть это ложноотрицательное решение. Вот. Но уровень различия очень невысокий. То есть она не отнесла, но сама сомневается, что он не относится к этой категории. И есть истинно отрицательные решения, здесь всё правильно, никаких птичек нет, с более высоким уровнем различия.

Теперь, э-э, я хочу вот что сказать, что Резберген, он, когда истинные решения, суммирует единичку к сумматору, в данном случае True Positive, вот эти просуммирует. А этот просуммирует единички к сумматору False Positive. А этот просуммирует единички к сумматору False Negative. А эти вот просуммируют единички к сумматору True Negative.

Теперь возникает вопрос: насколько это справедливо, ребята? Вот насколько справедливо наказывать систему и модель, считая, что суммируя единички к сумматору False Positive, если система сама очень не уверена в том, что действительно объект относится к этому классу. Всего 1% степени уверенности из 100.

Я посчитал, что это вообще-то несправедливо, что надо суммировать не единички к сумматору этим, а вот эту степень уверенности системы в этом решении. Вот это надо суммировать. То есть в данном случае нужно просуммировать 1/100 к сумматору False Positive. А это надо просуммировать 18/100 к True Positive. И здесь тоже False Negative нужно просуммировать, но не единички, а 0,9, 0,9, понимаете, нужно… э-э… э-э… даже не девять, а 9/1000 нужно просуммировать. Это 1/100 примерно, это 9/1000, понимаете? Если суммировать единички, тогда это 9/1000. А это 1/100, 1%.

В общем, если вот это учесть то, что я сейчас сказал, тогда меру Ван Резбергена нужно обобщить и разработать мультиклассовое и нечёткое обобщение меры Ван Резбергена.

И следующий момент. Смотрите, вот эти вот числа – True Positive, True Negative, False Positive, False Negative – они линейно растут при увеличении объёма выборки.

Вот если взять у нас какую-то… Так, не то я сделал.

Вот я сейчас вам нарисую, как они растут. Вот берём прямые линии и смотрим. Вот у нас вот это объём выборки. Вот, а это число решений.

Смотрим, какие решения могут быть. Берём… Быстрее всего растёт… Ну, в общем, они растут с разной скоростью эти… Это различные сумматоры.

И пишем здесь вот. Вот это что такое у нас? Это у нас сумматор True Negative. Он быстрее всего растёт фактически. А это True Positive. Это хорошая модель, когда вот так. А это False… Negative. А это False Positive. Что это значит? При увеличении объёма выборки быстро растёт число истинно отрицательных решений. Чуть-чуть помедленнее растёт число истинно положительных решений. Ещё медленнее растёт число ложноположительных, ещё медленнее ложноотрицательных. Вот.

Смотрим. Э-э, эти сумматоры используются при расчёте точности модели, полноты модели и достоверности модели. А если они постоянно меняются в зависимости от объёма выборки, постоянно меняются их значения, то к чему это может привести, ребята? К тому, что сами эти вот оценки точности, полноты и достоверности могут меняться в зависимости от объёма выборки. Что является не совсем правильным, потому что достоверность модели определяется самой моделью, самой математикой этой модели, а не тем, какое количество её использований было для решения задач различных. То есть если мы взяли модель какую-то, то эта модель, она вот есть, она обладает такими-то свойствами. И сколько бы мы её не использовали, эти свойства не меняются, понимаете? Если эта модель не адаптивная. Вот просто мы одну и ту же модель используем. И у нас определённые получаются результаты, определяемые тем, какая модель. Ну и какие данные, конечно. Но данные, они будем считать так, однородные, одни и те же данные, только объём разный. А закономерности в них те же самые. И вот получается что, что мы должны… то есть я так решил, подумал, что будет обоснованным, если мы будем говорить о том, что эта модель… результаты идентификации… результат оценки достоверности модели не должны зависеть от объёма выборки. То есть чем больше выборка, тем достоверность оценивается точнее, но это понятно. Но модель должна быстро сходиться к тому пределу, который можно считать достоверностью модели при неограниченном увеличении объёма выборки. У вас статистика была, ребята, нет?

Сейчас уже конец занятия у нас.

Ну, я вам могу сказать, что вероятность – это предел, к которому стремится относительная частота успешных исходов к числу всех испытаний, относительная частота называется, при неограниченном увеличении объёма выборки. То есть для того, чтобы модель быстро сходилась, нужно абсолютные частоты вот этих вот сумматоров True Positive, True Negative заменить относительными, которые стремятся к вероятностям. И я это сделал и разработал инвариантное относительно объёма данных нечёткое мультиклассовое обобщение F-меры достоверности Ван Резбергена.

**Заключение**

Значит, ребята, на следующем занятии мы продолжим рассмотрение этой модели с того момента, на котором мы сейчас остановились. А мы остановились на моменте, что мы рассматриваем достоверность моделей. Вот. После того, как мы рассмотрим достоверность моделей, мы рассмотрим и наиболее эффективные модели с точки зрения достоверности. Ну, это я сейчас не буду объяснять, но смысл такой, что это вот у нас положительное значение, это отрицательное, это их количество. Красное – это истинное решение, а синее – это ложное. То есть мы видим, что у нас, начиная с определённого уровня сходства, гораздо больше истинных решений, чем ложных. То есть эта модель очень хорошая получилась с точки зрения достоверности.

Всё, ребята, на этом мы наше занятие сейчас заканчиваем. Можете задать какой-нибудь вопросик. Вот. Ну, в принципе, вот я вам рассказал всё, то, что мог на за это время.

Значит, я вам очень рекомендую попробовать собрать такие вот данные, как в этой статье. Я же вам дал ссылки на статьи. Но смысл такой, что, конечно, можно не по гибридам, а по вашим работам. Там могут быть растения, могут быть какие-то там другие поля, например, могут быть, понимаете, там. Или сады. Строчка – сад, характеристика сада, понимаете? Вот. И мы попробуем с вами провести такие исследования на следующих занятиях с тем, чтобы вы потом, закончив обучение по этой дисциплине, сделали своё приложение. Я вам уже рассказывал, что вы в результате должны сделать своё приложение. Я вам сейчас дам ссылочку на это приложение, то есть на… на задание по разработке приложения и инструкцию о том, как его разрабатывать. Ну, это, по сути дела, то, что я вам рассказываю.

Ну, на этом, наверное, занятие заканчиваем. До свидания, ребята. Всего самого хорошего вам.