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

***Российская Федерация***

***Ватамов Д.Н. ПИ2104***

**150 Теория информации, данные и знания. Лекция 8. Методика вычисления количества информации 2020-10-20**

L**Заголовок:** Методика расчета количества информации, системный анализ и верификация моделей

**Резюме текста:**

Лекция посвящена методике вычисления количества информации (Тема 11 курса "Теория информации, данные, знания"). Занятие проходит 20 октября 2020 года (шестая пара).

**1. Введение и контекст лекции:**
Лектор приветствует студентов и аспирантов. Обсуждается тема лекции: методика вычисления коэффициентов количества информации, основанная на подходе Харкевича. Рассматриваются вопросы синтеза и верификации моделей знаний, модели абсолютных частот, условных и безусловных процентных распределений, а также частные критерии знаний в системно-когнитивных моделях. Цель – научиться рассчитывать информацию на основе эмпирических данных и использовать это для задач идентификации, прогнозирования и принятия решений.

**2. Расчет количества информации: Матрица абсолютных частот:**
Основным инструментом является матрица абсолютных частот (или корреляционная матрица), которая строится на основе исходных данных. Для этого данные предварительно формализуются: объекты описываются набором признаков (свойств или факторов) и относятся к определенным классам (будущим состояниям). Признаки и классы кодируются. Матрица показывает, сколько раз каждый признак (значение свойства/фактора) встретился у объектов каждого класса.

**3. Подготовка данных и концепция системности:**
Подчеркивается важность формализации предметной области: разработки классификационных и описательных шкал и градаций, кодирования данных. Лектор связывает процесс расчета информации с повышением уровня системности модели объекта познания. Системность определяется наличием и силой взаимосвязей между элементами. Выявление этих взаимосвязей (причинно-следственных зависимостей) превращает данные в информацию (осмысленные данные). Модель с высоким уровнем системности способна адекватно отражать объект моделирования (принцип Эшби).

**4. Частотный анализ и вероятность:**
На основе матрицы абсолютных частот рассчитываются относительные частоты: условные (встречаемость признака в конкретном классе) и безусловные (встречаемость признака по всей выборке). Обсуждается связь между относительной частотой и вероятностью через предельный переход: вероятность – это предел, к которому стремится относительная частота при неограниченном увеличении числа испытаний (наблюдений).

**5. Верификация моделей и важность признаков:**
Для сравнения фактической и теоретической (ожидаемой при отсутствии связи) частоты встречи признака используются два основных подхода: вычитание (основа критерия Хи-квадрат) и деление (основа для расчета количества информации по Харкевичу и коэффициента взаимосвязи). Ценность признака для решения задач (например, идентификации) определяется вариабельностью количества информации, которое он несет о принадлежности к разным классам. Признаки с низкой вариабельностью (мало отличающиеся условные и безусловные вероятности) малоценны.

**6. Оптимизация модели и применение:**
Предлагается оптимизировать модель путем удаления малозначимых шкал (признаков), оценивая их значимость через среднюю вариабельность информации по их градациям. Обсуждается применение методов для задач идентификации и прогнозирования. Упоминается мера Ван Рейсбергена для оценки достоверности моделей и ее обобщения (мультиклассовые, нечеткие, инвариантные относительно объема данных), разработанные лектором. Подчеркивается, что любая задача решается на основе модели, даже если она интуитивная и слабо формализованная.

**Детальная расшифровка текста:**

**1. Введение и контекст лекции**

* **Приветствие и обсуждение участников**
Здравствуйте, ребята.
Вы можете ответить.
Здравствуйте. Да.
Здравствуйте, здравствуйте.
Вот. Что-то вот Евгений Луцкий, почти мой однофамилец полный, тёзка. Он и аспирант, и первокурсник, и вообще он всё время на моих занятиях. Я заметил.
Евгений, вы кто? Бакалавриат, первый курс или кто вы? Или аспирант? А?
Ой, извиняюсь, я забыл выйти. Я аспирант, я просто...
А то я подумал, вдруг вам просто так интересно, что вы решили, так сказать, на все занятия приходить.
Не, я вот у нас лекция сейчас с вами была, я сейчас иду на следующую.
Я заметил. А, ну всего хорошего.
Всё, счастливо.
Угу.
* **Определение темы и расписания**
Так, ребята, у нас сейчас, э-э, сегодня 20 октября 2020 года. Шестая пара, которая длится 17:20-18:50. Восьмая лекция во всех группах, потому что это лекция, по дисциплине "Теория информации, данные, знания".
Тема 11 у нас сегодня: Методика вычисления коэффициентов количества информации. Методика расчетов, методика вычисления количества информации.
Вот у нас такая тема сегодня. То есть мы до этого рассматривали различные меры информации, а теперь будем рассматривать, как рассчитывается информация Харкевича. Вот, на основе эмпирических данных. И как на основе этого решаются в дальнейшем задачи идентификации, прогнозирования, принятия решений и исследования моделируемой предметной области.
* **Учебные вопросы**
То есть вопросы у нас учебные:
	+ синтез и верификация моделей знаний,
	+ модель абсолютных частот, статистические модели,
	+ э-э, условных и безусловных процентных распределений.
	Вот. А также частные критерии, э-э, знаний, э-э, в системно-когнитивных моделях.
	Это, дай бог, нам сегодня рассмотреть. Но следующий вопрос, задача, решение задачи идентификации и прогнозирования и аддитивные интегральные критерии. Ну это уже, видимо, на следующем занятии.
	А сейчас у нас вот эти вопросы. Сейчас я вам в чат эти вопросы отправлю. Это учебные вопросы этого занятия.
	Да, я вот забываю иногда об этом сказать. Занятие ведёт профессор Луценко Евгений Вениаминович.

**2. Расчет количества информации: Матрица абсолютных частот**

* **Начало рассмотрения методики**
Ну и начинаем рассматривать.
С вами эти вопросы о расчете моделей.
* **Определение матрицы абсолютных частот**
Первое рассчитывается модель на основе исходных данных, которую я называю матрица абсолютных частот или её называют ещё корреляционная матрица.
А, извините, вы ведь не включили экран?
Нет.
А.
Ну, шестая пара, это нормально. На шестой паре хорошо хоть звук включила. Да. Ну, шестая пара - это нормально для шестой пары.
Вот.
Я показываю вам учебное пособие.
Значит, для того, чтобы такую матрицу посчитать, нам нужны исходные данные определённым образом формализовать предварительно.

**3. Подготовка данных и концепция системности**

* **Формализация исходных данных**
То есть если у нас исходные данные представлены в какой-то экселевской таблице, строчками которой являются наблюдения (то есть строчки используют для того, чтобы описать наблюдения), вот, а колонки для того, чтобы описать э-э, какие свойства объектов и значения, или какие факторы действовали на объекты и что в результате получилось. Вот, э-э, значит, факторы и свойства - это описательные шкалы и градации, а что получилось - это классы, обобщающие категории, к которым относятся будущие состояния объекта моделирования, в которое он перешёл под действием этих факторов или к которым он относится потому, что у него есть определённые свойства.
* **Связь с процессом познания и системностью**
Теперь, ребята, я хотел вам сказать, что вот то, что мы сейчас рассматриваем, это очень тесно связано с процессом познания и, в частности, с повышением уровня системности модели объекта познания. Системность, я вам помните, говорил, что система - это множество э-э, взаимосвязанных элементов, и э-э, множество имеет такие, то есть эта система имеет такие свойства, которых не было ни у одного из элементов множества. Это называется системные свойства. Эти системные свойства тем более ярко выражены, чем сильнее взаимосвязи между элементами. Вот эти взаимосвязи мы должны сейчас как раз и выявить. То есть я вам на таком уровне теоретическом рассказывал о соотношении понятий данные, информация и знания и говорил, что информация - это осмысленные данные, а смысл - это знание причинно-следственных связей, зависимостей. То есть если мы берём какие-то события, описываем в данных исходных, и находим между ними взаимосвязи, то мы повышаем уровень системности данных. То есть эти данные, э-э, в которых выявлены причинно-следственные зависимости, уже это осмысленные данные, это называется информация. Так вот информация, можно сказать так, это данные, у которой уровень, у которых уровень системности повышен до такого уровня, что уже это...
...адекватна, она может его отобразить, потому что есть э-э, принцип Эшби, который об этом говорит, что если две системы взаимодействуют, то более сложная система с более высоким уровнем системности адекватно отражает более простую, а более простая система неадекватно, упрощённо отражает более сложную. Поэтому для того, чтобы наша модель смогла адекватно отразить объект моделирования, для этого она должна иметь определённый уровень сложности внутренней, э-э, определённый уровень системности. А системность э-э, тем выше, чем выше интенсивность взаимосвязей между элементами множества. И вот эти взаимосвязи, э-э, причинно-следственные взаимосвязи между событиями, они и будут у нас формироваться, э-э, вернее так, количественная оценка этих взаимосвязей и будет формироваться в наших моделях. В результате наши модели будут иметь высокий уровень системности и будут адекватно отражать объект моделирования. Уловили, ребят, что я сейчас говорил, нет? Это очень интересные вещи я вам на самом деле говорю.
* **Кодирование данных и структура матрицы**
Так вот, чтобы добиться этого, э-э, того, что я сейчас рассказывал, повышения уровня системности, отражения взаимосвязей, для этого мы сначала рассчитываем матрицу абсолютных частот, которая выглядит следующим образом. Эта матрица... Ну это я перебрал, пожалуй.
Вот так можно. Значит, смотрите, ребята, значит, у нас э-э, строки этой матрицы - это значения свойств или значения факторов. То есть у нас есть фактор, а у него есть значение, например, степень интенсивности этого фактора. Или есть у нас свойство и значение этого свойства. Скажем, свойство цвет, а значение красный, зелёный, жёлтый.
(Телефонный разговор: Привет, Игорь... Добрый вечер... Да я вот тут куча моментов... Вы завтра будете?... Ну я же дистанционно работаю, Игорь. Я вот сейчас у меня занятие, кстати, шестая пара, лекция... А там по поводу... Ну почему, знаю. Я тебе перешлю, перешлю там... Когда похороны завтра в час... Сейчас, да? Да. Я тебе перешлю информацию... А? Я думал, может подъедете сюда... Нет, я не смогу... У меня непрерывно пары идут. У меня по пять, по семь пар в день. Даже было и восемь один раз... Как вы с этим боретесь?... А как я борюсь? Я просто и читаю эти пары и всё... И под запись, да. И лекция, лабораторная, всё подряд... Их нет в расписании, но я их читаю, потому что в одном их расписании нет, а в другом расписании есть... У аспирантов, магистрантов, так бывает... Да там просто народ уже вот это вот всё исправлял... А я видел, ты ж прислал же, ты ж прислал... Ты ж прислал мне это, я ж видел... Да, да, так я ж одобрил... А вот реферат читали, вроде бы как более-менее... Ну слава богу... как бы это... Тему, темы внесли, они поменяли там... Я просто думал вот это вот, ну... вашу подпись под заключением просто с изменённой темой... А... Понял. Понял. Ну это можно сделать, это можно к Наташе подойти. Я ей позвоню, скажу, чтобы она поставила, она поставит... Наташа - это кто где?... Лаборантка... Какая?... А, во второй... Она подписывает так, что лучше, чем я сам. Я и не могу узнать, да... Хорошо... Ну вот. Ну тогда вы скажете, я завтра... Ну да, ну я по посоветуюсь там ещё узнаю, там уточню... Если, значит, всё, так сказать, решено, то чего там, конечно, проблем нет... А там что, что-то не решено?... А я не знаю. Я с этой с Татьяной Петровной не разговаривал же... Не, ну наше заключение мы ж... Ну я поговорю с ней всё равно... Ну хорошо... Тогда вы завтра не попадёте, значит... Нет, я точно не приеду. Я не приезжаю туда вообще... То есть я работаю из дома... Ну ладно, хорошо... Угу... Давай...)
Вот. Значит, для того, чтобы такую матрицу посчитать, нужно сначала подготовить исходные данные таким образом, чтобы у нас э-э, всем свойствам соответствовали коды, всем свойствам объектов, и всем классам, к которым они могут относиться, обобщающим категориям, тоже чтобы соответствовали коды. Это называется формализация предметной области, то есть разработка классификационных и описательных шкал градаций и кодирование исходных данных и получение обучающей выборки. Кодирование с помощью этих э-э, справочников классификационных описательных шкал градаций. Мы этот вопрос мы с вами рассматривали уже, как это всё делается, даже на лабораторных рассматривали, поэтому вы представляете себе. Вот, так я думаю, да, ребят? Правильно?
Вот. Поэтому у нас э-э, вот эта шапка горизонтальная, классы - это все классификационные шкалы и градации подряд. Э-э, то есть класс - это градация классификационной шкалы. Их может быть много этих шкал, и у каждой шкалы несколько градаций может быть. И вот здесь эти градации все идут от первой до последней. То же самое касается свойств и их значений или факторов и их значений. То есть идут факторы подряд или или свойства, а внутри них ещё идут их значения. Вот, скажем, шкала цвет описательная, они там идёт белый, красный, зелёный, синий, чёрный, там и пошло, понимаете? Вот. Ребят, вы реагируете на то, что я вас спрашиваю? Как вы считаете, понятно вам это, уловили, нет, смысл? То есть мы должны закодировать все эти вот э-э, исходные данные. Вот. После чего э-э, каждый объект э-э, выборки должен представляться системой кодов. С одной стороны, это э-э, наименование этого объекта, откуда он взялся вообще, э-э, откуда информация о нём взялась. И второе - это какие, какие у него свойства, их значения, или какие факторы на него действовали, и какие будущие состояния он перешёл.
* **Расчет элементов матрицы**
И вот эта матрица рассчитывается следующим образом. Берём мы, организуем цикл по объектам этим обучающей выборки. И э-э, когда э-э, какой-то текущий объект мы рассматриваем, то мы суммируем коды, то есть суммируем единички в строку, соответствующую какому-то коду признака, который у него есть, в колонку, соответствующую коду класса, который у него есть, которому этот объект относится. То есть мы э-э, когда рассматриваем какой-то объект конкретный, то организуем цикл по его признакам, то есть значениям свойств или значениям факторов, и по его классам, к которым он относится. И во все классы, к которым он относится, соответствующим строкам, соответствующим тем признакам, которые у него есть, суммируем единички.

**4. Частотный анализ и вероятность**

* **Интерпретация матрицы: факты**
В результате мы узнаём в этой матрице, получаем такое, такую информацию: сколько раз какой признак встретился у объектов каждого из классов. А я вам, по-моему, раньше говорил, что является фактом. Фактом является наблюдение определённого значения свойства у объектов определённой категории. И мы здесь в этой таблице как раз эти факты накапливаем и подсчитываем. Каких фактов, сколько у нас появилось.
* **Расчет относительных частот (условных и безусловных)**
И потом мы начинаем эти факты анализировать и выявлять эти вот как раз э-э, зависимости, которые я вам, о которых я вам говорил. За счёт чего и повышается уровень э-э, системности модели. Как мы начинаем их анализировать эти факторы, ребята? Эти факты, извините. Факты. Мы э-э, вычисляем э-э, относительные частоты, условные и безусловные. Условная относительная частота - это фактическое число встреч итого признака в житой группе, разделённое на число вообще признаков в этой житой группе, в этом житом классе. Либо разделённое на число объектов обучающей выборки, относящихся к этому классу. То есть есть два способа расчёта этой модели. Здесь об этом говорится. И получается две модели условных и безусловных процентных распределений, отличающиеся тем, каким способом рассчитываются относительные, условные относительные частоты. И рассчитываются также безусловные относительные частоты. Безусловные относительные частоты рассчитываются следующим образом. Э-э, делится суммарная э-э, колоночка, сумма... э-э, из матрицы абсолютных частот... Вот эта вот колоночка сумма э-э, используется. Значит, э-э, сумма по строке, элемент по строке суммарный, делится на суммарное количество объектов обучающей выборки, либо суммарное количество признаков у этих объектов обучающей выборки. Получается два разных, ну для разных матриц, э-э, для модели, э-э, которая первым способом рассчитывается и второй. Соответственно, по-разному рассчитывается вот эта условная, безусловная относительная частота.
* **Обозначения и связь с вероятностью**
При этом используются следующие э-э, обозначения для фактического числа встреч итого признака в житом классе, модель Abs. И э-э, условных и безусловных относительных частот. Вот здесь у нас показаны условные относительные частоты. То есть это относительная частота наблюдения итого признака у объектов житого класса.
Вообще относительные частоты, они связаны с вероятностями через предельный переход. Понятие вероятности - это понятие э-э, связанное с понятием предельного перехода или предела.
Вот. Если мы э-э, проводим испытания какие-то, какой-то системы, которая может принимать несколько состояний, ну, например, игральный кубик. Бросаем этот игральный кубик, то мы понимаем прекрасно, что вероятность выпадения любой из граней равно 1/6, если она равновероятна для всех граней. А относительная частота будет совершенно меняться по-другому. Если у нас э-э, один раз мы бросили, выпала тройка, то относительная частота - это равно число исходов э-э, к числу испытаний. То есть это 1 / 1, получится единица относительная частота. Если мы ещё раз кинем, получится, скажем, четвёрка, тогда относительная частота выпадения тройки и четвёрки будет 0,5. Число испытаний два, и выпал один раз тройка, один раз четвёрка. И вот если мы теперь продолжим эти эксперименты с бросанием этого кубика и э-э, продолжим их до бесконечности, неограниченно будем увеличивать число таких испытаний, то постепенно относительная частота будет приближаться к некоторому пределу, всё больше и больше будет асимптотически приближаться к некоторой величине, вот к этой 1/6. Э-э, и погрешность будет, различие между ними будет э-э, плавно монотонно уменьшаться, но никогда не будет становиться равным нулю. И, в общем, э-э, вероятность - это предел, к которому стремится относительная частота при неограниченном увеличении числа испытаний. Ясно, что э-э, никакого неограниченного увеличения числа испытаний не может быть, оно всегда ограничено. Вот. Оно будет, может быть, очень большим. Вот, настолько большим, что погрешность будет, скажем, миллиардная доля процента. Ну тогда можно говорить о том, что практически относительная частота э-э, ну, практически стала равна вероятности. Вот. То есть практически. Но это значит, не математически она ей равна, но настолько близка, что можно вполне обоснованно использовать термин вероятность, а не э-э, относительная частота. На практике даже при не очень больших выборках уже пользуются термином вероятность. Вот, понимая, что это лукавство определённое, погрешность есть между, различие между относительной частотой и вероятностью, это все понимают, почти что, кто знает, что такое вероятность. Ну, этот термин применяется.

**5. Верификация моделей и важность признаков**

* **Сравнение условных и безусловных частот (вероятностей)**
Теперь смотрите, ребята, значит, у нас возникает вопрос о том, как сравнить друг с другом условные и безусловные относительные частоты. Вот, допустим, у нас есть, э-э, ну или давайте будем слово вероятность говорить, потому что это одно слово, а относительная частота - это два слова. Вот, допустим, P и Т житое - это вероятность встречи итого признака или значения свойства, или значения фактора у объектов житой категории. А, значит, P итое - это вероятность встречи этого признака по всей выборке, э-э, по всем объектам, всем категориям, всем классам. Вот как их сравнить? Значит, я могу вам сказать, ребята, что есть всего два правильных способа сравнения разумных двух чисел, и неограниченное число, там тысячи способов неразумных. Вот когда я про это рассказываю, то мне всегда вспоминается дебаты Рейгана с сенатором Макфарлейном, который был претендентом на пост президента. Это было очень давно, ребята, когда был времена Горбачёва, вот, когда ещё был Советский Союз, потом он распадался, вот в это, в эти времена. И Рейган, он был артистом, прекрасно чувствовал публику, свободно ориентировался э-э, в зале, с людьми большим, вот большим числом людей. Э-э, прекрасно был бы, кстати, этим, э-э, капитаном команды КВН. То есть он остроумный, находчивый, не боится публики, чувствует наоборот подъём и чувствует себя совершенно свободно, раскованно и очень интересно, неожиданно выдаёт повороты мысли оригинальные. И вот он как-то раз на этих дебатах этому говорит этому сенатору Макфарлейну: "Существует э-э, бесконечное количество, ну или там он сказал, существует тысяча, тысяча способов э-э, нечестных заработать деньги. И только один честный способ существует". А Макфарлейн, будь идиотом, он, ну просто по-другому я даже сказать не могу, он спрашивает: "А какой честный?" Вот простачок, просто вот, я не знаю, на президента претендовал идти, шёл. Ну Рейган ему говорит: "А я так и думал, что вы не знаете, какой честный". И там рейтинги у него сразу раз-раз, там начали подниматься, у того опускаться. Ну кончилось тем, что Рейган стал президентом. Вы про это слышали, наверное. Вот. Так вот, какие существуют два честных способа сравнить два числа и тысяча нечестных? Значит, честные способы - это вычесть одно из другого и разделить одно на другое. Вот. А ещё можно вычесть одно из другого, а потом взять, прибавить 10 и отнять 10. А можно вычесть одно из другого, а потом прибавить семь и отнять семь. Ну вы поняли мою мысль, да? Вот таких способов существует неограниченное количество, счётное, но бесконечное, но счётное количество. Вот. Так вот, э-э, получается очень интересная ситуация, что, конечно, мы все понимаем, что вот это делать бессмысленно, прибавлять и вычитать какое-то число там и так далее.
* **Хи-квадрат и информация Харкевича как меры сравнения**
Вот. Так вот, э-э, посмотрим то, как сравнивается друг с другом э-э, фактическая частота встречи итого признака в житой группе и теоретическая частота его наблюдения. Это э-э, выражение является одним из слагаемых матрицы Хи-квадрат. Значит, смотрите, ребята, значит, мы вычитаем из фактической частоты наблюдения этого признака в житой группе теоретическую, которая вот таким выражением описывается. А теперь смотрите, если мы э-э, это делаем путём вычитания, сравниваем фактическую, теоретическую частоты. А если путём деления? Тогда надо вот эту фактическую частоту разделить вот на эту, э-э, на эту дробь. А можно умножить на обратную дробь. Вот мы это видим. Смотрите, n и Т житое умножить на n и разделить на n итое, разделить на n житое. Видите? То есть можно сравнить э-э, фактическую и теоретическую частоту встречи признака либо вычесть их друг из друга, либо разделить. Вот если разделить, тогда это, если вычесть, то это Хи-квадрат. Если разделить, то это коэффициент возврата инвестиций. Или количество информации. Смотрите. Вот я вам показываю. Вот Хи-квадрат, а это количество информации по Харкевичу. Если мы возьмём и разделим фактическую частоту встречи признака на теоретическую и возьмём логарифм, то это количество информации по Харкевичу называется.
* **Связь с условной и безусловной вероятностью**
Если вы посмотрим на выражение для условной и безусловной вероятности, вот они здесь есть у нас. Выражение для условной вероятности. То это у нас что? Э-э, фактическая частота, разделённая на сумму по колоночке. Вот она у нас есть, фактическая частота, разделённая на сумму по колоночке. А что представляет собой теоретическая, э-э, безусловная частота? Безусловная. Это условная частота, это безусловная. Или условная вероятность, безусловная вероятность. Это сумма по строке n итое, разделённая на число объектов обучающей выборки n, либо число признаков э-э, объектов обучающей выборки. Если два способа расчёта. Мы получаем вот это выражение, ребята. А это выражение фактически оно означает отношение условной и безусловной вероятности. То есть тоже сравнение путём деления условной и безусловной вероятности. А есть ли их сравнение путём вычитания? Есть. Это модель N6, коэффициент взаимосвязи, который в электротехнике известен.
* **Оценка значимости признаков (вариабельность информации)**
То есть мы видим, что и абсолютные частоты мы сравниваем путём вычитания и деления, и относительные частоты, вероятности мы тоже сравниваем путём вычитания и деления. Вот. Но если мы вычитаем, сравниваем путём вычитания, то у нас само собой получается ноль, и в случае абсолютных частот, и в случае э-э, условных и безусловных вероятностей. Само собой получается ноль, если они равны и нет взаимосвязи между наличием этого признака и принадлежностью объекта к житому классу. А если мы это делаем сравнение путём деления, то у нас не получается ноль. А нам бы желательно иметь ноль, потому что мы хотим использовать аддитивный интегральный критерий, в котором хотим складывать частные критерии друг с другом, вот, и получить суммарное количество информации, скажем, о принадлежности объекта к классу по всем его признакам. Поэтому вот эти вот э-э, эту таблицу я и назвал частные критерии. Таблица частных критериев. И вот э-э, мы можем нормировать к нулю это отношение. Ну, когда связи нет, получается отношение равное единице. Как мы можем нормировать его к нулю? Да просто вычесть единицу можем. А можем вычесть не единицу, а 10. Да можем вычесть и 10, но тогда надо будет потом прибавить девять. Ну я вам рассказываю вот эти вот нечестные способы сравнения. Поняли, да, и нормировки? Ну, понятно, что это совершенно безумие, вычитать 10, а потом прибавлять девять. То есть можно просто вычесть единичку и всё. Вот. И м-м, можно просто взять логарифм от этого отношения. Что и сделал Александр Харкевич, когда предлагал свою семантическую меру целесообразности информации. То есть эта мера, она имеет очень глубокий смысл. По сути дела, можно сказать так, что смысл тот же самый, что Хи-квадрат, но другие единицы измерения. Хи-квадрат - это абсолютные частоты измеряются в количестве, ну, в штуках, грубо говоря. Сколько раз встретилось то или иное значение свойства у объектов той или иной, той или иной категории. А количество информации - это вероятности сравниваются, э-э, встречи этого признака в группе, у объектов этой группы, житого класса, и вероятность встречи этого признака по всей выборке. Вот. Есть и другие формы этих выражений, которые я в своих статьях описал. Но здесь я не буду на этом останавливаться. Сам Александр Харкевич описывал свою меру следующим образом. Он говорил так, что если мы сообщаем какое-то количество информации субъекту, например, а этот субъект хочет достичь определённой цели, ну, например, найти помещение, где договор заключается, то при этом мы, когда ему эту информацию сообщаем, то у него либо увеличивается вероятность достижения цели, либо уменьшается, либо ничего не меняется. Если у него увеличивается вероятность достижения цели, значит, мы ему сообщили положительное количество информации. Если уменьшается, значит, сообщили отрицательное количество информации. То есть это тогда означает, что мы им сообщили ему дезинформацию. Ну и если ничего не меняется, значит, мы ему как бы вроде как что-то сообщили, на самом деле ничего не сообщили. То есть что-то сказали, но то, что мы сказали, никакой информации не содержит о том, что его интересует. Вот такие вот меры, меры, э-э, количество информации. Так они вот рассчитываются по этим частным критериям.

**6. Оптимизация модели и применение**

* **Пример матрицы информативности и оценка значимости шкал**
Смотрим теперь, э-э, что у нас получается в результате того, что мы используем эти частные критерии для сравнения условных и безусловных вероятностей, условной вероятности встречи признаков в каком-то классе, к безусловной вероятности встречи его по всей выборке. Если мы применим на практике эти вот частные критерии, э-э, которые разные способы сравнения условных и безусловных вероятностей и абсолютных частот описывают, то что у нас получится в результате? В результате у нас может получиться э-э, ну, 10 матриц различных: матрицы абсолютных частот, матрицы условных безусловных процентных распределений ПРЦ1, ПРЦ2 и матрица Abs, я их так назвал когда-то в своё время, и так вот оно и прижилось. И э-э, матрицы системно-когнитивных моделей, которые я назвал Inf1, Inf2, Inf7. То есть мы вот таких матриц системно-когнитивных моделей можем получить семь разных матриц, в зависимости от того, какие формулы будем использовать для расчёта вот этого количества информации и и Т житое. А таких формул вы видели семь от Inf1 до Inf7. А ещё мы уже рассмотрели, как рассчитывается матрица абсолютных частот и матрица условных безусловных процентных распределений. В результате мы получаем семь вот таких матриц. Сейчас я вам покажу размер получше сделаю эту матрицу.
Как-то она мало увеличивается, не знаю, что-то не понял.
Вот. Ну, наверное, видно. Значит, что в этой матрице интересно, ребят? Мы получили, э-э, какое количество информации содержится в каждом значении свойства или в каждом значении фактора о принадлежности объекта с этим, с этим значением свойства к каждому из классов. Это и есть те связи, ребята, между элементами, э-э, которые увеличивают уровень системности. То есть уже дальше, как говорится, э-э, сложно себе представить. То есть у нас получилось что? Что мы э-э, исходили из эмпирических данных, э-э, в которых вообще неизвестно о каких-либо закономерностях в этих данных. Потом мы разработали справочники классификационных описательных шкал градаций, э-э, и закодировали исходные данные с их помощью, получили нормализованную базу данных. Уровень системности резко увеличился, потому что отношения один ко многим возникли. Эти отношения, они уже э-э, отражают взаимосвязи между элементами в системе и фактически они уже повышают уровень системности модели. А потом мы взяли и выявили ещё взаимосвязи между наличием тех или иных свойств у объекта и принадлежностью его к тем или иным классам. И ещё у нас увеличилось число взаимосвязей между элементами модели и повысился уровень её системности. Улавливаете, ребята, что я говорю, а? А уровень системности модели, он э-э, обеспечивает, скажем так, определяет э-э, и возможность её использования при познании. Потому что модель э-э, тем более является адекватной, чем она имеет высший уровень системности. И если, значит, адекватна модель отражает только те э-э, объекты познания, которые имеют уровень системности ниже, чем уровень системности модели. Вот такие дела. То есть эти вот взаимосвязи, которые там у нас есть, они повышают уровень системности, и модель становится адекватной даже для довольно, для описания довольно сложных объектов моделирования. То есть она хорошо их описывает, проще выражаясь, если так простым языком сказать. Вот, это в соответствии с принципом Эшби. То есть сейчас я стал излагать эту тему немножко в более развитом виде, чем раньше.
* **Оценка ценности признаков для задач классификации**
Теперь посмотрим, что у нас, э-э, что можно сказать о ценности тех или иных э-э, значений э-э, свойств для того, чтобы отличить объекты одного класса от объектов другого класса. Представьте себе, что у нас такая задача решается. Стоят учащиеся за дверью, и мы можем задавать вопросы о том, э-э, э-э, какие у них, какая длина волос у этих учащихся, э-э, брюки у них или нет, и есть ли у них брюки или нет, и э-э, есть ли у них мобильный телефон или нет.
И вот вы догадываетесь, что если есть, если если отвечают да на вопрос о том, есть ли длинные волосы у этого учащегося, то скорее всего, мы можем сказать так, что скорее всего, это студентка. То есть в этом признаке большой объём информации о том, что это э-э, студентка, что она относится к классу студентки. А вот признаки брюки, э-э, наличие брюк, есть небольшое количество информации о том, что это студент, и небольшое количество информации о том, что это не студентка. А признак мобильный телефон вообще никакой информации ни о чём не содержит в этом плане, об этих классах. И мы можем определить, э-э, значит, ну и, соответственно, есть признаки, которые являются очень ценными для решения задачи идентификации, типа признаки длинные волосы, например, в этом примере. Отсюда вытекает очень интересный вывод, что мы можем измерить ценность признака, измеряя вариабельность количества информации вот в этой матрице системно-когнитивной модели. Если у нас вариабельность признака, количество информации низкая, то есть очень мало информации о принадлежности, очень мало информации о непринадлежности. То есть здесь написано брюки, например, одежда, да? Вот. Признак практически бесполезный, ну почти бесполезный. Потому что девушки тоже чаще всего в брюках ходят, да? А вот признак длинные волосы, он очень является полезным для идентификации по этим категориям. Почему? Потому что у девушек он встречается почти стопроцентная вероятность его встречи, там 99, скажем, и девять, да? А у ребят очень низкая, там 1%, например. Получается отношение вот этот э-э, логарифм отношения вот этот, если мы его посчитаем, для девушек получается логарифм двух, то есть у них встречается почти 100%, а по всей выборке 50, около 50. Это единица, логарифм двух по основанию два - это единица, 1 бит информации. То есть мы узнаём э-э, к классу э-э, девушки или к классу ребята относятся учащиеся, значит, получаем 1 бит информации. А если признак э-э, брюки, то мы получаем очень малое количество информации по модулю, причём отрицательное. Значит, смотрите, значит, э-э, у ребят маленькая отрицательная величина, э-э, да, а что касается длин

.