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

***Российская Федерация***

**112 Интеллектуальные информационные системы и технологии.**

**Лабораторная 3. 2020-10-01**

Заголовок

Оценка Достоверности Моделей в Системе Эйдос: Критика F-меры и Обобщения

Резюме

1. Введение и Контекст

Лекция/лабораторная работа №3 для группы ИТ 18-21 по дисциплине "Интеллектуальные информационные системы и технологии" (1 октября 2020 г.). Занятие ведут профессора Луценко Е.В. и Аршинов Г.А. Цель – изучить достоверность моделей в системе "Эйдос". Подчеркивается важность использования актуальной версии системы (30.09.2020) и повторения материала на своих компьютерах для выработки навыков.

2. Достоверность Моделей: Необходимость и Проблемы Оценки

Объясняется, почему оценка достоверности модели критически важна перед ее использованием для решения задач (распознавания, идентификации, прогнозирования, принятия решений, исследования). Использование недостоверной модели ведет к неверным результатам (неправильный диагноз, неверный прогноз, неэффективное управление). Приводится аналогия с обучением плаванию: теория без практики (воды) не формирует навык.

3. Критерии Достоверности: F-мера Ван Ресбергена и ее Ограничения

Рассматриваются основные понятия: истинно положительные (TP), истинно отрицательные (TN), ложноположительные (FP, ошибка 1-го рода), ложноотрицательные (FN, ошибка 2-го рода) решения. Вводится классическая F-мера Ван Ресбергена, основанная на точности (precision) и полноте (recall).

Выявляются недостатки F-меры:

Зависимость от объема выборки: Значения TP, TN, FP, FN растут с увеличением выборки, делая F-меру зависимой от ее размера. Сравнение моделей по F-мере корректно только при одинаковых объемах выборок.

Моноклассовость: F-мера предназначена для оценки моделей, относящих объекты только к одному классу, тогда как система "Эйдос" работает с многоклассовыми объектами.

4. Обобщения F-меры в Системе "Эйдос"

Представлены три обобщения F-меры, разработанные для системы "Эйдос" и лишенные указанных недостатков:

Нечеткое многоклассовое обобщение: Учитывает степень уверенности системы в решении (уровень сходства), суммируя не единицы, а уровни сходства для TP, TN, FP, FN. Это делает оценку более адекватной при ошибках с низкой уверенностью.

Инвариантное относительно объема выборки обобщение: Использует относительные частоты (вероятности) вместо абсолютных чисел решений, что устраняет зависимость от размера выборки. Эта мера стабилизируется при достаточном объеме данных (около 500 объектов).

Комбинированное обобщение: Сочетает нечеткость, многоклассовость и инвариантность. Является наиболее предпочтительным критерием в системе "Эйдос".

5. Решение Задачи Идентификации и Выбор Модели

Демонстрируется процесс решения задачи идентификации (распознавания) в системе "Эйдос" (режим 4.1.3). Подчеркивается необходимость предварительного выбора наиболее достоверной модели (режим 5.6) на основе инвариантного нечеткого многоклассового обобщения F-меры перед решением прикладных задач. Объясняется, что система автоматически использует выбранную модель для идентификации.

6. Математические Основы: Количество Информации и Скалярное Произведение

Кратко поясняется, что в основе моделей "Эйдос" лежит расчет количества информации по Харкевичу (логарифм отношения апостериорной и априорной вероятностей) и сравнение моделей ХИ-квадрат и меры Харкевича. Отмечается, что итоговая формула для уровня сходства объекта с классом представляет собой скалярное произведение вектора, описывающего объект, на вектор, описывающий класс (из матрицы модели).

7. Заключение

Подводится итог: рассмотрена важность оценки достоверности моделей, проанализирована классическая F-мера и ее недостатки, представлены и обоснованы обобщения F-меры, используемые в системе "Эйдос". Показан процесс выбора достоверной модели и решения задачи идентификации.

Детальная расшифровка текста

1. Введение и Контекст

Всем, ребята, здравствуйте. Началась запись. Началась, началось занятие. Здравствуйте.

Сейчас у нас сегодня 1 октября 2020 года. Первая пара с 8:00 до 9:30. Лабораторная работа номер три с группой ИТ 18-21 по дисциплине Интеллектуальные информационные системы и технологии. Занятия ведут профессор Луценко Евгений Вениаминович и профессор Аршинов Георгий Александрович.

Здравствуйте, ребята, что-то вы как-то там вообще в чате только здороваетесь. А, Аня, привет.

Микрофоном пользуйтесь, потому что… сейчас я включу совместный доступ и будем проходить лабораторную работу дальше.

Ребят, я бегло пройдусь по тому, что мы раньше делали для того, чтобы у вас в памяти освежить.

Обращаю ваше внимание, версия системы 30.09.2020. У кого не срабатывает Start Eidos, не подкачивает обновление, прошу установить, скачать и установить эту систему прямо с моего сайта, скачать и установить. Там есть важное обновление.

Скачать полную версию системы и развернуть, либо можно вот этот файл обновления скачать и развернуть в папочке с системой, это будет то же самое.

2. Достоверность Моделей: Необходимость и Проблемы Оценки

2.1 Важность оценки достоверности

Что мы делаем? Мы продолжаем изучать лабораторную работу 3.03. И сейчас у нас по плану мы изучаем вот эту вот… достоверность модели. Это сегодняшняя тема, тема сегодняшнего занятия.

Начинаем с чего? С того, что мы удаляем все приложения. Вы, значит, старайтесь это повторять на своих компьютерах.

Я хотел бы, чтобы у вас это очень как бы крепко закрепилось в вашем сознании и подсознании, что если вы создаёте модель, то это делается для решения каких-то задач: идентификации, прогнозирования, принятия решений и исследования. Но если модель недостоверна, то надо понимать, что результаты идентификации будут неправильными или результаты диагностики будут неправильными. То есть поставят врач диагноз, но неправильный, лечить будет от другого заболевания, не от того, которым человек болеет. Если что-то прогнозируется с помощью недостоверной модели, то оно может и не осуществиться. То есть это маловероятно даже, ну случайным образом может так получиться, что оно осуществится. А вообще, недостоверные результаты прогнозирования будут. Если будут приниматься какие-то решения с помощью недостоверной модели, то эти решения не будут приводить к достижению поставленной цели управления. Если вы будете исследовать модель и думать, и будете предполагать, что этим самым вы исследуете объект моделирования, то вы будете ошибаться, потому что модель-то неправильно отражает объект моделирования. Поэтому результаты исследования модели нельзя считать исследованием объекта моделирования. Поэтому достоверность модели – это очень важный момент, изучение достоверности модели, оценка хотя бы достоверности. И выбор достоверной модели, в которой вы видите, что получается хорошо решение задачи, тогда вы уже этой моделью можете спокойно пользоваться для решения различных задач.

2.2 Аналогия с обучением плаванию

Поэтому я ваше внимание обращаю на то, чтобы вы хорошенько запомнили. Человек учится только тому, что он сам делает. Ну я могу вам привести пример. Вот можно учиться плаванию. Как можно учиться плаванию? Положить вас на топчан, такой, знаете, как кровать без спинки, грубо говоря, там без это, без границ, как вот в медицинских этих учреждениях. И сказать, что для того, чтобы плыть, надо махать руками и ногами, вот так ногами вот так делать, а руками вот так. Понятно? Понятно. А теперь давайте ложитесь на топчан и делайте это. И вы будете это делать, и при этом вы совершенно не научитесь плавать. Почему? Потому что чтобы научиться плавать, нужно плавать, учиться плавать в воде, понимаете? Вот такая маленькая вот деталь есть существенная. Короче говоря, если идёт занятие лабораторное, и на нём парень сидит за компьютером и выполняет задание преподавателя, а рядом с ним сидит симпатичная девушка, положила голову ему на плечо, то кто чему учится, ребята? Парень учится решать задачу, которая изучается на этой лабораторной работе, а девушка учится сидеть рядом с парнем, положив ему голову на плечо, понимаете? То есть каждый учится тому, что он делает. Так вот, сейчас я веду, допустим, занятие, я хотел бы, чтобы вы повторяли всё на своих компьютерах. Почему? Потому что если вы будете просто слушать и смотреть, что я рассказываю, то это будет обучение лекционного типа, то есть вы будете это всё понимать там, всё, но у вас не будет формироваться умений и навыков. То есть вы будете вместо того, чтобы учиться работать в системе, решать задачи реально в системе, вы будете учиться смотреть на экран и слушать, что вам рассказывают, понимаете? Надо, чтобы вы учились именно работать. А чтобы научиться работать, для этого нужно именно вот работать, понимаете? То есть надо прямо на компьютере включить систему и работать. Тогда вы будете чему-то учиться. Именно вот будете умения и навыки приобретать те, которые должны приобретать на лабораторных работах. А иначе вы будете приобретать, ну, может быть, понимание будет, информационный уровень обучения, может быть, даже знания будете приобретать, но они не будут у вас доводиться до уровня умений и навыков. Поэтому давайте, включайте компьютеры, запускайте всё, что я вот делаю, повторяйте. Если какие-то возникают вопросы на лабораторной работе, совершенно нормальным является, если вы будете спрашивать микрофоном, задавать вопросы. И на лекциях тоже можно спрашивать вполне, ну, в пределах разумного. Если очень много начнёте спрашивать, то занятие, то есть трудно будет рассказывать.

3. Критерии Достоверности: F-мера Ван Ресбергена и ее Ограничения

3.1 Основные понятия (TP, TN, FP, FN)

Добавляем лабораторную работу третьего типа. О'кей.

Так что можно теперь пользоваться графическим процессором при расчётах, ни на что, так сказать, ни о чём не вспоминать. Всё хорошо работает. Я понял, в чём моя ошибка. Слишком много ложных решений. Тогда я сообразил, как же, какой же прогноз можно сделать, чтобы ошибок было минимальное количество. И вот этот же самый кубик у вас, вы его кидаете, а я говорю: "Ребята, у вас сейчас ничего не выпадет вообще". Говорит: "Да ну ладно, как это не выпадет?" Вот сейчас я кину, раз, вот видите, выпала двойка. Я говорю: "Да, я признаю, я ошибся, действительно, выпала двойка. Но в основном-то я был прав, прав. И один не выпало, и три, и четыре, и пять, и шесть, ничего ж не выпало. А я ведь говорил, что не выпадет". Они говорят: "Да, но вы допустили ошибку". Я говорю: "Допустил. Но я только одну ошибку допустил. На предыдущем прогнозе я допустил пять ошибок прогнозирования. И одно было истинное предсказание, осуществившееся. А остальные были ложными". А вот сейчас-то наоборот, наоборот. Э-э, все истинные предсказания отрицательные, что этого не произойдёт. И только одно ложное, что всё-таки что-то у вас там выпало. Ну, теперь скажем о том, что является идеальным прогнозом. Идеальным является тогда, когда вы предсказываете, что произойдёт и чего не произойдёт. И это так и осуществляется. Это идеальный прогноз. Реально это бывает крайне редко, только на таких очень жёстко детерминированных моделях, где факторы очень жёстко определяют будущее состояние объекта моделирования. Реально бывает так, что вы говорите: "Скорее всего, вот это произойдёт. Вот прямо вот наиболее вероятный вариант". Вот это тоже может быть, но менее вероятно. А вот эти вот менее вероятные, скорее всего, не произойдут. Причём некоторые из них не поймёшь там, около нуля где-то колеблются уровень сходства с решениями положительными и отрицательными о том, что что-то произойдёт, что-то не произойдёт. А у некоторых прямо точно видно, что этого не произойдёт. Высокий уровень сходства с решениями отрицательными, что этого не произойдёт. Вот это реальный прогноз. Так вот, э-э, реальный прогноз, он чаще всего наблюдается на практике, когда модели статистические, когда они нечёткие, когда есть исключения из правил. Ну то есть в реальных условиях чаще всего осуществляется реальный прогноз. И задача системы, которая прогнозирует, что осуществится, в том, чтобы как можно было больше истинных прогнозов и как можно меньше ложных.

3.2 F-мера, точность, полнота

А что такое… Давайте сейчас вам дам определение истинные и ложные прогнозы, положительные и отрицательные. Значит, если система оценивает сходство объекта с классом, уровень сходства как положительный, то есть система считает, что объект похож на какой-то обобщённый образ класса, то есть уровень сходства с ним положительный, тогда это называется положительный прогноз. Если же объект имеет отрицательный уровень сходства с классом, вот, то он на него не похож. Тогда это называется отрицательный прогноз. И положительные, и отрицательные прогнозы могут быть и истинными, и ложными. Истинные прогнозы: система видит, что уровень сходства объекта с классом положительный, то есть она считает, что он к этому классу относится объект. И действительно он к нему относится. Тогда это называется истинно положительный прогноз. А если система считает, что объект относится к какому-то классу, а на самом деле он к нему не относится, тогда это называется ложноположительный прогноз. И есть такой термин, который широко распространён, как бы на слуху, как говорят, ложноположительное решение, ложное срабатывание. Вот. Это очень опасная ситуация. Представьте себе, на самолётах есть транспондеры, так сказать, системы для определения свой самолёт, идентификации своих самолётов. И средства ПВО получают, э-э, шифр, зашифрованном виде получают, э-э, как должен, как должен делаться запрос на то, что свой самолёт, и как самолёт должен отвечать, если он свой. Это очень секретная информация, потому что если получит её противник, то он сможет спокойно через средства ПВО пролететь. И вот представьте себе, что эта аппаратура даёт сбой. То есть вы определяете, ваши средства ПВО, противовоздушной обороны, определяют свой самолёт или чужой. Вот. И свой самолёт летит, а он идентифицируется как чужой. Понимаете? И происходит пуск ракет и сбивается, если в боевых условиях. То есть это очень чревато всё. Ну и наоборот, если летит чужой самолёт и идентифицируется как свой, ошибки идентификации вот эти, то это тогда тоже очень опасно, потому что он пролетит со средствами поражения, со своими, и нанесёт удар. Вот. А средства ПВО не выполнят свои задачи по охране каких-то объектов. Ну вот и всё. То есть эти вот ошибки идентификации, не идентификации, они крайне опасные вообще-то, если реально они… в реальной жизни это всё, то это… огромные убытки, в финансовом выражении, и жизни могут быть, и в медицине тоже так же это очень чревато, когда диагноз не ставится, ставится неправильно, ставится другой диагноз. То есть вот ложноположительное решение – это когда ставится другой диагноз. Теперь, что такое истинно отрицательное решение? Значит, объект не относится к какому-то классу фактически, и система определяет, что уровень сходства его с этим классом ниже нуля, то есть тоже она определяет, что он не относится. Это истинно положительное решение, то есть истинно отрицательное решение. Теперь, если же система определяет, что объект не относится к классу, то есть уровень сходства отрицательный с классом, а на самом деле он к нему относится, ну тогда это ложно отрицательное решение.

3.3 Критика F-меры (зависимость от выборки, моноклассовость)

Это всё такие базовые понятия, связанные с оценкой достоверности моделей, убедительные вполне, вполне убедительные, то есть соответствуют здравому смыслу, логике, пониманию того, что понимается под достоверным решением, под ошибочным решением. Всё как бы правильно. Это то, что я сейчас говорю. Я сейчас рассказываю общепризнанное представление, в общем-то. И вот был такой учёный, ну он, по-моему, и сейчас есть, в очень пожилом возрасте, Ван Ресберген, который предложил посчитать эти значения сумматоров, на основе которых он предложил очень много разных выражений математических, основанных на этих сумматорах, которые он назвал точность модели, полнота модели, достоверность модели. И широко известна F-мера Ван Ресбергена. Это классическая мера оценки достоверности модели, которая используется и в системе Eidos, потому что она классическая. То есть это, в общем-то, ну, вещь широко распространённая мера. Хорошая мера, ребята, хорошая, обоснованная, убедительная. Но в ней есть, как говорится, и на солнце есть пятна. И в этой мере Ван Ресбергена есть тоже свои моменты, которые нас не устраивают. Почему они нас не устраивают, ребят? Сейчас я вам скажу. Извините за опоздание. Ну ладно. Что ж теперь? Ребят, а у кого стоит система Eidos, поставьте птичку. Кто вот поставил себе на компьютере и повторяет то, что я показываю на экране. Это ж лабораторная работа.

Значит, в чём недостатки этой меры? Сначала просто описание этой меры, что она собой представляет. Потом какие у неё есть недостатки, на мой взгляд. И потом, э-э, значит, суть в чём, ребят? Ван Ресберген предложил посчитать эти значения сумматоров, посчитать, каких решений сколько было: истинно положительных, истинно отрицательных, ложноположительных, ложноотрицательных. Вообще, я вам скажу так, что модель, э-э, должна правильно относить объекты к тем классам, к которым они относятся, и правильно не относить к тем, к которым они не относятся. И может ошибаться и в первом, и во втором случае. Вот все эти сумматоры, они называются интуитивно понятно: True Positive, True Negative, False Positive, False Negative. Ну, True – истина, False – ложь, Positive, Negative – особо в таком комментировании не нуждаются, объяснения. Но я уже сказал, что это уровень, знак уровня сходства. Если он больше нуля, тогда позитив, если меньше нуля, то негатив. Это, если сходство больше нуля, то это система считает, что этот объект относится к этому классу в соответствии с моделью. А если меньше нуля, считает, что он не относится. А фактически может быть и так, и не так может быть тоже. Так вот, э-э, Ван Ресберген предложил посчитать эти значения сумматоров. То есть вот сколько таких решений было: истинно положительных, истинно отрицательных. То он суммирует именно единички к этим сумматорам. Он подсчитывает, каких решений сколько было: истинно положительных, истинно отрицательных.

Вот, э-э, если посмотреть на эти вот, э-э, характеристики достоверности модели, F-меру, точность модели, полноту модели, для разных объёмов выборки, то у нас получится, ребята, разные значения достоверности. Это что значит? То есть мы можем исследовать и я вот, собственно, призываю к этому, и в этой статье описываю, описываю это. Я просто взял и исследовал зависимость критерия Ван Ресбергена от объёма выборки от нуля до 2.500 и построил вот такую кривульку жёлтенькую, видите? Сначала, когда объектов выборки не очень много, сравнительно, ну где-то вот, как я оцениваю, где-то около 300, то вообще критерий F-мера Ван Ресбергена является неустойчивой, то есть она резко падает, резко увеличивается. То есть видно, что здесь такие затухающие колебания какие-то. То есть она является неустойчивой, то есть она быстро меняется при изменении небольшом изменении объектов обучающей выборки. Это вообще плохое качество. Это означает, что при небольших объёмах выборки вообще нельзя ей пользоваться. То есть она может дать какой-то фонарный результат совершенно. То есть она там колеблется сильно, видите? А потом выходит на такую более плавную кривульку, которая видно, что она стремится к некоторой асимптоте, стабилизируется, можно сказать, значение этой кривой где-то в районе около 2.000-2.500 объектов обучающей выборки, она стабилизируется. Значит, я вам могу сказать, что это большой количество объектов обучающей выборки. Вот из тех задач, которые я решал, э-э, очень часто задачи там ограничиваются тем, что там 500 там или там 700 объектов обучающей выборки, а не вот эти вот тысячи – это редкий случай вообще-то. Ну бывает такое, но это редко. И вот если посчитать эти вот, э-э, характеристики достоверности модели, F-меру, точность модели, полноту модели для разных объёмов выборки, то у нас получится, ребята, разные значения достоверности. Это что значит? То есть мы можем исследовать… То есть это означает, что сравнивать модели по этому критерию можно только в том случае, если они получены на обучающих выборках одинакового объёма. А если они получены на обучающих выборках разного объёма, то сравнивать их по этому критерию некорректно. Потому что есть зависимость самой этой меры от объёма выборки.

4. Обобщения F-меры в Системе "Эйдос"

4.1 Нечеткое многоклассовое обобщение

Я понял, в чём моя ошибка. Слишком много ложных решений. Тогда я сообразил, как же, какой же прогноз можно сделать, чтобы ошибок было минимальное количество. И вот этот же самый кубик у вас, вы его кидаете, а я говорю: "Ребята, у вас сейчас ничего не выпадет вообще". Говорит: "Да ну ладно, как это не выпадет?" Вот сейчас я кину, раз, вот видите, выпала двойка. Я говорю: "Да, я признаю, я ошибся, действительно, выпала двойка. Но в основном-то я был прав, прав. И один не выпало, и три, и четыре, и пять, и шесть, ничего ж не выпало. А я ведь говорил, что не выпадет". Они говорят: "Да, но вы допустили ошибку". Я говорю: "Допустил. Но я только одну ошибку допустил. На предыдущем прогнозе я допустил пять ошибок прогнозирования. И одно было истинное предсказание, осуществившееся. А остальные были ложными". А вот сейчас-то наоборот, наоборот. Э-э, все истинные предсказания отрицательные, что этого не произойдёт. И только одно ложное, что всё-таки что-то у вас там выпало.

Вот. Это называется нечёткое обобщение меры Ван Ресбергена. Третий момент. Значит, эти вот все сумматоры, они содержат какие-то варианты решений, подсчитывают, их сколько решений, да? И понятно, что сумма значений всех этих сумматоров равно количеству объектов этой распознаваемой выборки или обучающей выборки, которые были скопированы распознаваемы для, когда осуществлялась оценка достоверности модели. То есть когда мы какой-то объект предъявляем системе, она относит его правильно или неправильно и не относит к тем или иным классам, то к какому-то из сумматоров вот этих вот True Positive, True Negative, False Positive, False Negative, к какому-то одному из них, одному, я подчёркиваю, из них суммируется единичка по Ресбергену. Это означает, что если сложить значения всех этих сумматоров, у нас как раз и получится число объектов распознаваемой выборки.

4.2 Инвариантное относительно объема выборки обобщение

И у нас получается очень интересная ситуация, что когда мы увеличиваем объём распознаваемой выборки, то у нас растут значения всех этих сумматоров. Ну, практически линейно, но с разной скоростью. То есть значения всех этих сумматоров увеличиваются при увеличении объёма выборки. Ну, почти линейно я сказал, они увеличиваются, там колебания есть какие-то, связанные с неоднородностью данных. Но, в общем, если посмотреть вот на больших данных, то они прямо вот выходят на линии, на прямые линии, которые под разными углами находятся к оси X. И будут толсти вверх, потому что при увеличении объёма выборки к какому-то одному из этих сумматоров суммируется единичка, то есть они все будут увеличиваться. И вот если посчитать эти вот, э-э, характеристики достоверности модели, F-меру, точность модели, полноту модели для разных объёмов выборки, то у нас получится, ребята, разные значения достоверности. Это что значит? То есть мы можем исследовать… То есть это означает, что сравнивать модели по этому критерию можно только в том случае, если они получены на обучающих выборках одинакового объёма. А если они получены на обучающих выборках разного объёма, то сравнивать их по этому критерию некорректно. Потому что есть зависимость самой этой меры от объёма выборки. А я это увидел, посчитал недостатком это и решил от этого недостатка избавиться и избавился, ребята. При объёме выборки выше 500, зависимость отсутствует от объёма выборки в тех мерах, которые я предложил. То есть это устойчивые модели, устойчивые меры достоверности, сопоставимые. То есть если будут пользоваться моими обобщениями меры Ван Ресбергена, которые я предложил, вот этим инвариантом, то можно сравнивать э-э модели, э-э достоверность моделей, полученных на разных объёмах выборки. Это корректно. А на классической мере это некорректно. Это сравнение, потому что есть зависимость самой этой меры от объёма выборки.

4.3 Комбинированное обобщение (нечеткое, многоклассовое, инвариантное)

Вот это синее – это второе обобщение меры Ван Ресбергена, полученное из первого. То есть это и мультиклассовое, и нечёткое, и ещё независимое, инвариантное с точки зрения математики, физики, такие случаи используют, в таких случаях этот термин используется, инвариантное, то есть независящее от объёма выборки. И вот посмотрим, как они эти меры ведут себя. Значит, при малых объёмах выборки, вот эти нечёткие мультиклассовые меры и инвариантная, и неинвариантная, они близки и показывают близкие значения, гораздо более высокие, чем у классической меры Ван Ресбергена. Второе, они более устойчивы, чем классическая мера Ван Ресбергена при малых значениях числа объектов обучающей и распознаваемой выборки. При больших они все устойчивые, а вот при малых они неустойчивы, а эти вот обобщения, которые я предложил, устойчивые, особенно вот эта инвариантная устойчива. Видите, там вообще нет никаких-то колебаний, она прямо сразу плавно идёт с самого начала, при малых объёмах выборки тоже плавно идёт, понимаете? Без каких вот таких серьёзных колебаний. А дальше что происходит? А дальше получается очень интересная вещь. Где-то в районе 400 объектов обучающей выборки, распознаваемой, происходит разделение первого обобщения меры Ван Ресбергена, мультиклассового, нечёткого, от второго, которое ещё, кроме этого, и инвариантное. И происходит разделение какое? Вот эта инвариантная, она, во-первых, сразу же стабилизируется очень быстро и ползёт практически горизонтально, да, параллельно оси X. Это происходит при 500 уже объектах обучающей и распознаваемой выборки. То есть, что это значит? Если у нас объём выборки около 500 объектов наблюдения, то вот эти меры, то есть инвариантная мера уже даст устойчивое значение достоверности модели, которое дальше практически не меняется, то есть при увеличении объёма выборки это значение не меняется. А остальные продолжают падать при увеличении объёма выборки. Вот то, что они то, что они падают, это, знаете, что означает? Что есть зависимость, ну так по-простому можно сказать, формально математически, есть зависимость достоверности модели от объёма выборки, наблюдаемая выше 500, ну она всё время наблюдается, но у инвариантной меры она выше 500, эта зависимость отсутствует объектов выборки. А у этих вот меры классической и первого обобщения неинвариантного, мультиклассового, нечёткого обобщения, зависимость наблюдается от объёма выборки.

5. Решение Задачи Идентификации и Выбор Модели

Запускаем как раз вот синтез и верификацию модели.

Мы дошли до того места, на котором мы остановились на прошлом занятии. Сейчас я вам покажу порядок преобразования данных в информацию и в знания в системе Eidos. И мы видим, ребята, что первый этап системно-когнитивного анализа автоматизированного – это называется этот этап когнитивно-целевая структуризация предметной области. Это единственный неавтоматизированный этап системно-когнитивного анализа. На этом этапе мы определяем, что мы рассматриваем в качестве факторов или признаков, свойств объектов, значений свойств. А что мы рассматриваем в качестве результатов влияния этих факторов? То есть это такая просто идея, что ли, постановки задачи, даже не сама постановка, а идея, что вот мы хотим изучить. Потом осуществляется формализация предметной области. Этот этап и все последующие, они автоматизированы в системе Eidos, то есть выполняются они в самой системе путём выбора соответствующих режимов. Э-э, на этом этапе анализируются файлы исходных данных, выявляются в нём классификационные, описательные шкалы и градации, а потом файлы исходных данных кодируются с их использованием и получается обучающая выборка, которая представляет собой нормализованную базу исходных данных, нормализованную с помощью вот этих вот справочников. После этого мы осуществляем синтез и верификацию моделей в режиме 3.5. А эти вот процедуры выполняются формализация предметной области в режиме 2.3, 2.2 программный интерфейс ввода данных из внешних источников данных. Ну и в других режимах он может выполнять, эти функции могут выполняться. В частности, когда вот мы лабораторную работу устанавливаем, то там кое-что делается для того, чтобы установить лабораторную работу, создаётся приложение, копируются туда файлы соответствующие, создаются пустые исходные данные, копируются, а потом запускается программный интерфейс 2.3, 2.2, который и, собственно говоря, и вводит эти данные в систему и формируется таким образом приложение. Когда мы рассчитываем модели, то сначала рассчитывается матрица абсолютных частот, в которой отражено, какое количество раз каждое значение описательной шкалы встретилось у объектов, относящихся к соответствующему какому-то классу. А классами являются значения классификационных шкал. Потом рассчитываются матрицы условных и безусловных процентных распределений, на основе которых рассчитываются системно-когнитивные модели или базы знаний. Потом оценивается достоверность всех этих моделей. Вот что мы сейчас и будем изучать с вами подробно. И в наиболее достоверной модели решаются задачи распознавания, идентификации, прогнозирования, задачи принятия решений и задачи исследования предметной области путём исследования её модели.

Переходим, ребята, для исследования достоверности модели, переходим в режим 3.4. В этом режиме есть Help. Этот Help мы можем посмотреть.

Вот у нас решение задач. Первая задача, которую мы рассматриваем – это задача идентификации. Вот у нас решение задач. Первая задача, которую мы рассматриваем – это задача идентификации. Ну можно просто говорить идентификация, потому что то, что здесь добавлено системное, это просто акцентирует внимание на то, что объект описан большим числом признаков, параметров. Будем эти задачи решать. Потом перейдём к задаче прогнозирования и принятия решений, которые тесно взаимосвязаны. А сейчас посмотрим, э-э, вот что. Для того, чтобы решить задачу идентификации, нам нужно выбрать текущую модель. Это особенность именно при решении задачи идентификации существует. Во всех остальных режимах мы можем, э-э, то есть задача решается для всех моделей сразу, кроме задачи идентификации. Почему? Потому что эта задача очень трудоёмкая в вычислительном отношении. Ну представьте себе, что у нас есть 700.000 объектов обучающей выборки, а в модели 400 классов есть, и объекты описаны 11.000 признаков. Каждые, то есть модель 11.000 признаков, 400 классов – это обычное дело, это обычная модель, не какая-то там гигантская, а такие, которых очень много в приложениях с такими моделями получается. И вот что мы делаем? Для того, чтобы решить задачу идентификации, мы организуем цикл по объектам обучающей выборки от одного до 700.000. Потом внутри него организуем цикл по классам от одного до 400. А внутри него организуем цикл по признакам объекта от одного до 11.000, да? И э-э, узнав, какое суммарное количество информации во всей системе признаков объекта о принадлежности его к этому классу, который в цикле сейчас рассматривается как текущий, мы заносим соответствующую информацию в базу данных: код объекта обучающей выборки, код класса и количество информации суммарное о принадлежности к этому классу этого объекта. И всё, ребята, понимаете? Запоминаем это всё, потом сортируем и вот так вот визуализируем. Вот у нас объект, вот суммарное количество информации больше всего о том, что это элемент компьютера, мышка, и вот об этих классах тоже у нас положительное количество информации, об этих отрицательное. В системе признаков этого объекта о принадлежности к житому классу. Как его можно посчитать? Нужно взять, сделать цикл по всем признакам объекта и суммировать те признаки, количество информации о принадлежности к житому классу только по тем признакам, которые есть у этого объекта. А если его нет, тогда не суммировать. То есть делать цикл, а внутри него сделать if then else там или просто if. И суммировать только в том случае, если признак есть у объекта, суммировать. Нету признака, не суммировать. А как узнать, есть он или нет? Берём объект обучающей выборки. И смотрим. Есть у этого объекта этот признак или нет? Двенадцатый, семнадцатый, двадцать третий есть. А вот одиннадцатого там с первого по одиннадцатый нету. А сорок седьмого тоже нету. То есть можно сделать массив, в котором будут элементы либо один, либо ноль. Если признак есть, то единица, если признака нет, то ноль. Массив, в котором столько же элементов, сколько всего признаков в модели. Вот в этой модели. Вот у нас 50 признаков суммарно во всех шкалах описательных. Поэтому создаётся массив из пятидесяти элементов. И элементы, каждый элемент, номер каждого элемента соответствует коду признака вот в этой матрице, коду. Прямо вот элемент первый, первому, второй, второму соответствует, понимаете? И вот если у нас какой-то признак у объекта есть, то в этот элемент тогда равен единице этого массива. Если нету, тогда ноль. В результате мы, если э-э, просуммируем, э-э, вот так сделаем цикл по элементам от первого до последнего признака, да, М-того, ну это я так обозначал последний признак, номер последнего признака. Мы можем сделать там if и сравнивать элемент массива L-итый равен нулю или единице. Если он равен единице, тогда суммировать вот это количество информации к некоторому сумматору, который перед циклом обнуляется. В результате в конце цикла получаем сумму количества информации в признаках объекта о принадлежности к житому классу. Эту информацию, сумму, суммарное количество информации заносим в базу данных: код объекта обучающей выборки, код класса и количество информации суммарное о принадлежности к этому классу этого объекта. И всё, ребята, понимаете?

Это колоссальный объём вычислений, ребят, колоссальный, просто огромнейший. Вот. И для того, чтобы в разумные сроки он осуществлялся, я решил осуществлять его не во всех моделях, а в какой-то определённой модели. Ну сразу возникает вопрос: а в какой модели осуществлять нам решение задачи идентификации? Ну, я думаю, что можно догадаться с трёх раз, что, конечно, в той, которая наиболее достоверна. Вот в той модели, которая оказалась наиболее достоверной, в ней надо решать все эти задачи.

6. Математические Основы: Количество Информации и Скалярное Произведение

Вот это вот выражение, оно представляет собой скалярное произведение в координатной форме двух векторов: вектор класса из матрицы модели и вектор, описывающий объект. Вектор класса, ребята, это просто колоночка из модели. Вот. Колонка, соответствующая классу. Это вот координаты этого вектора. Это их название этих координат. Можно F3 посмотреть. Вот берём любую колонку – это координаты вектора. Это название этих координат. Так что скалярное произведение. Чем оно интересно? Тем, что скалярное произведение, оно, как говорится, и в Африке скалярное произведение. То есть в случае, если пространство не ортонормированное, применять меру Евклида некорректно. И в случае, если оно искривлено или закручено, или если у него топология не плоскости, ну то есть оно не Евклидово, короче, пространство, то меру Пифагора, Евклидово расстояние применять нельзя. Ну её все применяют, но при этом получается некая погрешность. А вот межвекторное расстояние можно применять и корректно.

7. Заключение

Вот, собственно говоря, причины, по которым мы должны выбрать наиболее достоверную модель, в которой будем решать задачу идентификации..