«ШАМАНСТВО» В АНАЛИЗЕ ДАННЫХ

научно-популярная лекция

доцент ВМК МГУ имени М.В. Ломоносова, д.ф.-м.н. А.Г. Дьяконов
http://alexanderdyakonov.narod.ru/ **e-mail:** djakonov(собака)mail(точка)ru

ТЕКСТ НЕ ВЫВЕРЕН!!! ПРОСЬБА СООБЩАТЬ ОБ ОШИБКАХ!!!

ПРЕДИСЛОВИЕ

Иногда заметить феномен гораздо
ценнее, чем объяснить его.
(*Ю.И. Журавлёв*)

Когда в прошлом году лектор стал читать мини-спецкурс с таким названием, к нему подошёл известный специалист по спектральным методам в обработке сигналов профессор Ф.Ф. Дедус и спросил: «Я прочитал объявление о Вашем курсе. Скажите, а что Вы в нём будете рассказывать? Неужели про танцы с бубном?» Поэтому сразу хочется оговориться: **никакого шаманства в привычном смысле этого слова на лекции не будет**. На самом деле, это уже почти устоявшийся термин, историю которого полезно узнать.

В 60-х годах прошлого века, как тогда говорили, «партия и правительство» (а точнее Председатель Совета Министров СССР А.Н.Косыгин) поручили геологам **найти на территории СССР месторождение золота африканского типа.** Всего таких месторождений в мире было семь, они представляют собой «золотые блины» толщиной 10–15 сантиметров на глубине 2–3 километра (т.е. найти их «случайно» почти невозможно). Имелась полная статистика по ним: пробы грунта, снимки местности и т.д. Имелась также статистика по местам, в которых специалисты предполагали наличие золота, но тщательный анализ установил его отсутствие. Такую статистику в анализе данных часто собирают в таблицы «объект-признак», пример показан на рис. 1: по строкам записаны описания месторождений (объектов), а столбцы соответствуют некоторым признакам (мнение одного эксперта, особенность рельефа и т.д.).

Для консультации два молодых геолога (А.Н. Дмитриев[1] и Ф.П. Кренделев[2]) обратились в Институт математики новосибирского Академгородка. Проблемой заинтересовался математик Ю.И. Журавлёв[3]. На вид это была как раз математическая задача: интерполяция функции от многих переменных (функция равна единице на описаниях месторождений золота и нулю – на описаниях остальных мест, надо определить, где она ещё будет равна единице, кроме известных семи точек). Проблема была только в том, что переменных больше сотни, а точек, в которых известны значения функции, всего несколько. Задача казалась неразрешимой, по крайней мере, методами

[1] **Алексей Николаевич Дмитриев, доктор геолого-минералогических наук.**
[2] **Фёдор Петрович Кренделев, доктор геолого-минералогических наук, член-корреспондент АН с 1984 г.**
[3] **Юрий Иванович Журавлёв, доктор физико-математических наук, академик РАН с 1992 г.**

«классической математики», но помог случай. За год до этого Академгородок посетил американский профессор Э. Фейгенбаум, который в своих лекциях утверждал, в частности, что **ни одна по-настоящему сложная задача не может быть решена чисто математическим путём, необходимо использовать человеческий опыт, «подсматривать», как решают задачу специалисты и разрабатывать эвристические алгоритмы**. Ю.И. Журавлёв ухватился за эту идею, долго беседовал с геологами, выяснял, как вообще принято искать полезные ископаемые, и переводил эти «геологические» идеи в математическую форму.

1	1	да	…	7	1 (есть золото)
0	1	нет	…	5	1 (есть золото)
…	…	…	…	…	
0	1	нет	…	1	0 (нет золота)
1	0	нет	…	4	0 (нет золота)
0	0	да	…	2	0 (нет золота)

Рис.1. Таблица объект-признак.

В результате был изобретён алгоритм для поиска золота. С точки зрения «чистой математики» он был некорректен (сама постановка задачи была не совсем корректна), но он сработал! Описывать алгоритм в этой лекции мы не будем[4], но объясним один из основных принципов его работы. Рассмотрим рис. 1. Если подописание (1,1) (выделено в таблице) встречается только в первых двух признаках объектов первого класса и не встречается в объектах второго класса, то наличие такого подописания характерно для объектов первого класса. Это называют элементарным тестовым классификатором[5], их строят много, причём строят так, чтобы они были «неупрощаемы», т.е., например, если сократить это подописание, то значение 1 в первом признаке (или во втором) уже не будет характерно ни для одного из классов. Для нового объекта (потенциального месторождения), который надо классифицировать, смотрят, какие классификаторы «голосуют» за его вхождение в первый класс, а какие – за вхождение во второй. В простейшем случае «прислушиваются к большинству».

Описание алгоритма А.Н. Косыгин выслушал лично, а ведущий кибернетик того времени, **академик В.М. Глушков[6] назвал этот алгоритм «шаманским»**, что было справедливо: «математическим» его можно было назвать с натяжкой. Так появился этот термин. На самом деле, события, описанные выше, являются началом целого направления научных исследований. Но на этой лекции мы не будем уделять внимание этим

[4] **Описание есть в работе [ДЖК, 1968].**
[5] **Идея, кстати, основана на понятии «тест», введённым С.В. Яблонским [Я, 2001].**
[6] **Виктор Михайлович Глушков (1923 – 1982), выдающийся советский математик и кибернетик, решивший обобщённую пятую проблему Гильберта.**

исследованиям, а покажем, где и как возникают подобные «шаманские алгоритмы».

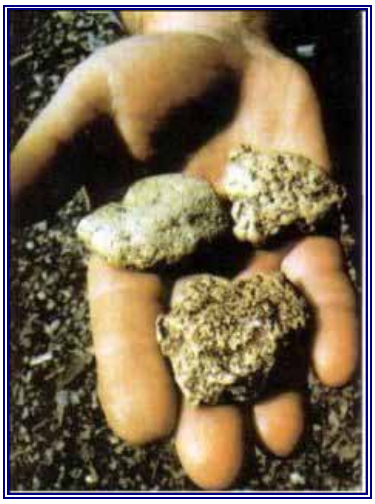

Рис. 2. Золото самородное [P0].

Первая задача этой лекции – показать, **какие задачи встречаются в анализе данных**. Конечно, мы не успеем охватить все задачи, даже привести примеры задач всех типов такой широкой области как анализ данных (на западе – data mining). Дело даже не в ограничениях по времени, а в том, что каждый день появляются всё новые и новые задачи. По сути, анализ данных – это всё, где можно применить математику, программирование и, конечно, здравый смысл для поиска закономерностей, интерпретации данных, принятия решения и т.д. Но главное, чтобы слушатели поняли, что задачи анализа данных есть «буквально повсюду», быть может, и их задачи (с которыми им приходится сталкиваться «по профессии») также являются задачами анализа данных.

Вторая задача – это как раз показать, а **что такое «здравый смысл» при решении задач анализа данных**. Ведь идеи, лежащие в основе решения таких задач, очень простые. Например, «все месторождения золота должны быть чем-то похожи друг на друга», осталось только выяснить чем! Часто (а это и будет показано) удаётся решать задачи «методом разглядывания данных». Это и есть «шаманство»… Мы не пишем сложных формул и больших программ, а смотрим, как выглядят данные нашей задачи.

УПРАВЛЯЕМ СИЛОЙ МЫСЛИ!

Не снабжайте детей готовыми формулами,
формулы - пустота, обогатите их образами,
на которых видны связующие нити.
(*А.Д. Сент-Экзюпери*)

Снова вынуждены предупредить, что название не следует понимать буквально: речь вовсе не пойдёт о каких-то сверхспособностях человека. В

начале этого века[7] стали очень популярными **исследования в области «Brain Computer Interface»** (Интерфейс «мозг–компьютер»), которые как раз **занимаются построением эффективных интерфейсов для управления ЭВМ с помощью… сигналов головного мозга**. Всё очень просто. Человек садится перед компьютером, а ему на голову надевается шапочка с электродами, которая подключается к компьютеру. Как известно, во время ментальных действий (по-простому «раздумий») меняется потенциальное и магнитное поле разных участков головного мозга, всё это фиксируется с помощью приспособления шапочка–провода–компьютер. Таким образом, компьютер знает, что там «происходит в голове у человека», правда, в терминах изменения потенциала. Осталось перевести это на более понятный язык, чтобы компьютер понимал, «о чём думает человек». Для того, кто считает это фантастикой, вот несколько иллюстраций, см. рис. 3–6.

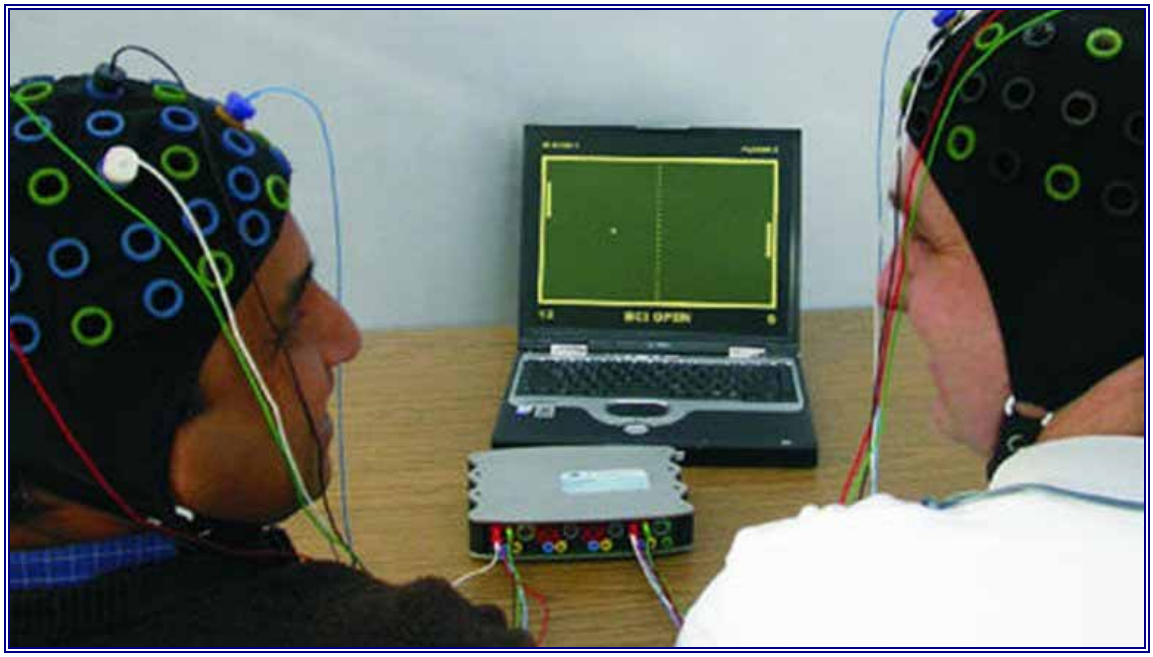

Рис. 3. Игра в теннис за компьютером «с помощью силы мысли» [P1].

Люди могут играть в теннис на компьютере, не касаясь клавиатуры, мыши и прочих приспособлений, которые задействуют их руки (рис. 3). Есть специальные системы ввода слов с помощью интерфейса «мозг–компьютер» (рис. 4). Естественно, в первую очередь, подобные разработки рассчитаны на людей с ограниченными возможностями. Есть даже специальные инвалидные кресла, которые приводятся в движение «силой мысли» (рис. 5). Чтобы это всё-таки не казалось фантастикой, отметим, что во всех случаях **речь идёт не о понимании мыслей человека компьютером, а о различении нескольких ментальных состояний** (это задача классификации, о которой мы поговорим ниже). Например, при игре в теннис человек хочет, чтобы ракетка переместилась вправо или влево, и компьютер должен отличить ментальное

[7] **Исследования начались ещё в 1970-х годах в Университете Лос-Анжелеса штат Калифорния (UCLA). В России этой тематикой занимается лаборатория нейрофизиологии и нейро-компьютерных интерфейсов на биологическом факультете МГУ (заведующий – д.б.н. А.Я. Каплан) [brain].**

состояние, вызванное желанием переместить ракетку вправо, от ментального состояния, соответствующего перемещению ракетки влево.

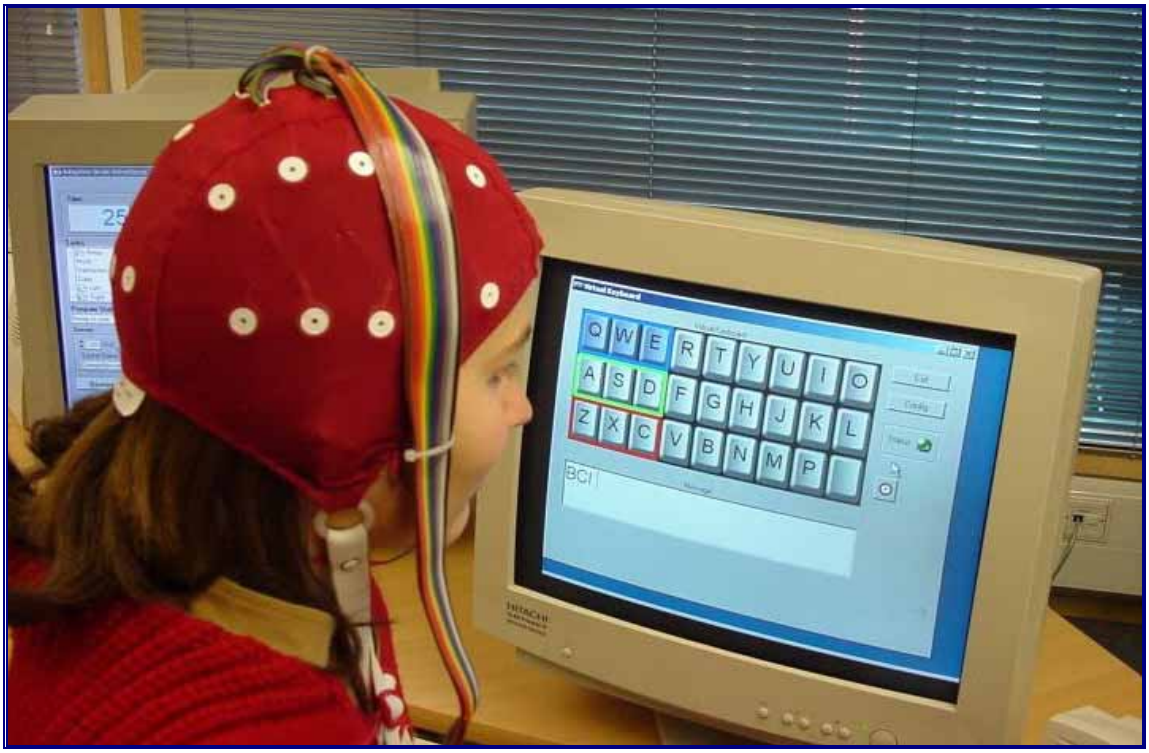

Рис.4. Ввод текста в компьютер без использования клавиатуры [P2].

Рис.5. Управление роботом (протезами) при помощи интерфейса «человек-компьютер» [P3].

Рис.6. Шапочка с электродами [P4].

По тематике «Brain Computer Interface» было проведено даже несколько крупных международных соревнований. Например, участникам международного конкурса по классификации сигналов «BCI competition III» 2003 г. (данные Data set I [BCI 3]) была предложена следующая задача. Даны описания 278 сигналов, которые отражали два ментальных состояния, таким образом были разбиты на два класса. На самом деле, это многомерные сигналы, поскольку снимались с помощью ECoG-электродной[8] сетки размера 8x8 (электродов), т.е. одновременно измерялось 64 сигнала. Но все иллюстрации, которые у нас будут, соответствуют лишь одному из этих 64 сигналов, снятому, в некотором смысле, с «лучшего электрода», снимавшего показания с зоны мозга, в которой происходили «наиболее интенсивные» изменения. Каждый сигнал состоял из 3000 точек, поскольку отражал 3-секундную активность головного мозга во время некоторого ментального действия и снимался с частотой 1000Гц. **Обратите внимание, на размерности, которые возникают в такой простой реальной задаче: исходные данные записываются в трёхмерной матрице размера 278×64×3000** (278 64-мерных 3000-точечных сигнала). На соревновании требовалось построить алгоритм, который классифицирует сигналы. Качество классификации алгоритма проверялось на контрольной выборке из 100 сигналов (их верную классификацию знали лишь организаторы).

Попробуем решить описанную выше задачу. Оказывается, даже не зная методов обработки сигналов, теорию электромагнетизма и нейробиологию можно предложить достаточно неплохой метод решения, но придётся немного «пошаманить»… Сначала давайте посмотрим на наши данные. На рис. 7 показано несколько сигналов первого и второго класса, а также один из сигналов, которые нам надо классифицировать. **Самый естественный метод**

[8] **ECoG = для снятия электрокортикограмм.**

классификации – посмотреть, на сигналы какого класса похож классифицируемый (этот метод называется «ближайший сосед», NN - nearest neighbor [B, 2010]). Но наши классифицируемые сигналы вообще не похожи на те, о которых известна классификация[9]! Это следствие того, что сигналы были получены в разные дни, т.е. совершенно в разных условиях (могли чуть-чуть измениться сопротивления проводников в приборах, у испытуемого изменилось настроение, что повлияло на активность головного мозга и т.д.)

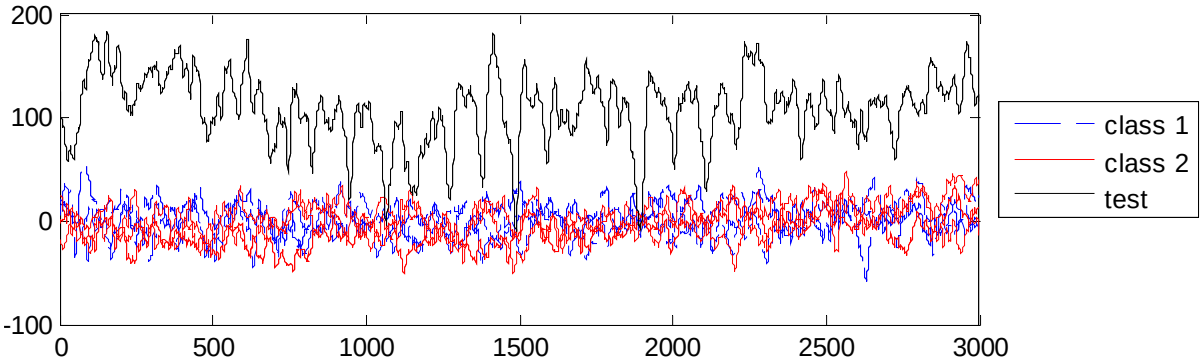

Рис. 7. Визуализация ECoG-сигналов головного мозга.

Вычислим для каждого сигнала его минимальное и максимальное значение. Тогда **сигнал представляется точкой в соответствующем пространстве** (первая координата – максимальное значение, второе - минимальное), см. рис. 8. Это так называемое двухмерное признаковое пространство.

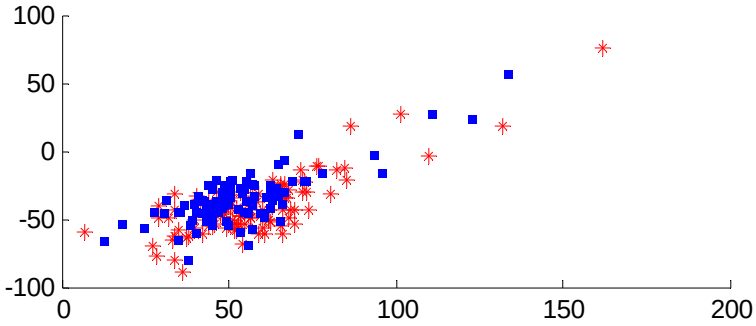

Рис. 8. Максимальные (по горизонтали) и минимальные (по вертикали) значения сигналов.

Вот ещё пример такого пространства (рис. 9). Здесь показаны средние значения сигналов в первые 1,5 секунды (первый признак) и в последние (второй признак). Видно, что значения коррелируют, т.е. по среднему значению в первые 1,5 секунды «угадывается» значение в последние 1,5 секунды: оно примерно такое же. **Чем больше похоже это облако точек на линию, тем точнее мы сможем «угадать» второе значение по первому**, а если облако точек размыто и на линию не похоже (рис. 8), то такого угадывания не получится (признаки некоррелированы, т.е. значение одного не определяет значение второго).

[9] **Или «похожесть сигналов», пригодная для решения этой задачи, отличается от визуальной схожести.**

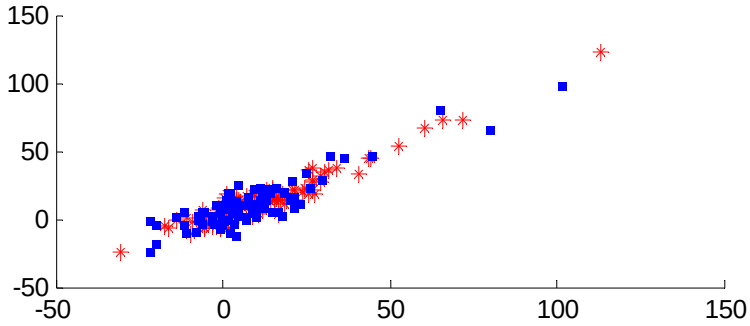

Рис. 9. Средние значения сигналов за первые 1,5 секунды (по горизонтали) и последние 1,5 секунды (по вертикали).

Как видим, **подобные картинки позволяют много сказать о сигналах (нужно только «уметь их читать»)**. Например, что они однородные (с течением времени их признаки: среднее значение, максимальное и т.д. не сильно изменяются). Но мы пока не приблизились к решению задачи… Посмотрим, однако, на рис. 10.

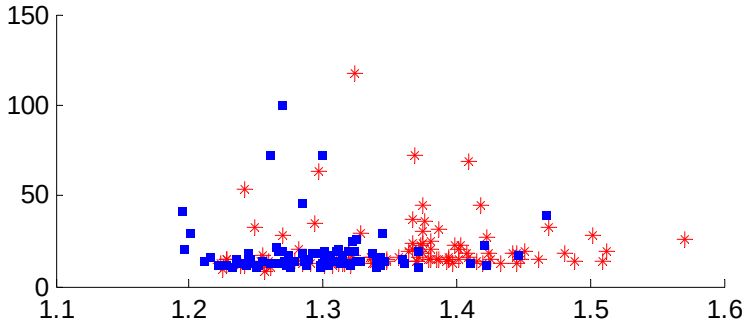

Рис. 10. Средний модуль разности последовательных значений сигнала (по горизонтали) и среднее значение сигнала (по вертикали).

На нём по вертикали отложено среднее значение сигнала, а по горизонтали – значение

$$\frac{1}{n-1}\sum_{i=1}^{n-1}|u_{i+1}-u_i|$$

(для сигнала $(u_1,\ldots,u_n)$). «Физический смысл» последнего признака понять нетрудно: он описывает **скорость изменения сигнала**. Удивительно, но **по этому признаку сигналы неплохо отличаются**: если значение признака маленькое, то сигнал, скорее всего, принадлежит первому классу (синему), а если большое – второму (красному).

Кстати, есть ещё один (очень известный!) признак, который также описывает «разнообразие значений сигналов»: дисперсия. На рис. 11 изображена пара этих признаков: наш «хороший» и дисперсия.

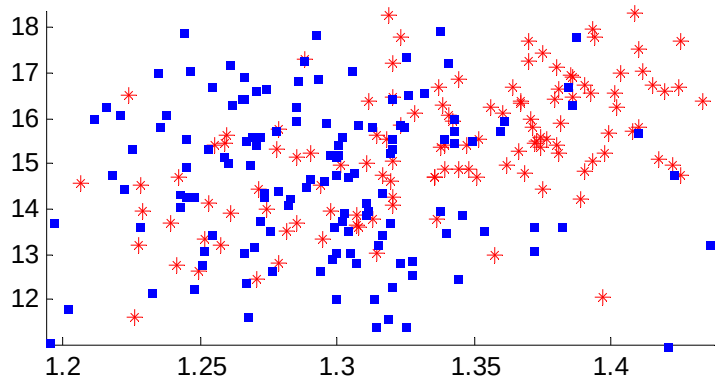

Рис. 11. Средний модуль разности последовательных значений сигнала (по горизонтали) и дисперсия значений сигнала (по вертикали).

Вот это очень типичная картина! Признаки не коррелируют, но **точки располагаются на плоскости не хаотично, а вдоль определённых линий**. Глядя на такую картинку, можно сделать много гипотез о природе данных. Например, что точки, расположенные последовательно вдоль одной линии, соответствуют данным, снятым в течение одного отрезка времени. А может, совсем наоборот, подобные расположения зависят не от времени, а от типа ментального действия. Осталось только проверить гипотезы!!!

Проверять гипотезы мы не будем, хотя именно здесь начинается «настоящая наука», а покажем **ещё один стандартный приём, применяемый при визуальном анализе данных: «чуть-чуть» изменить найденный признак**. Например, вместо модуля использовать квадрат:

$$\frac{1}{n-1}\sum_{i=1}^{n-1}(u_{i+1}-u_i)^2 .$$

Новый признак, как правило, сильно коррелирует с исходным (он ведь получен его незначительным изменением), но в проекции на эти два признака можно увидеть интересные закономерности, см. рис. 12.

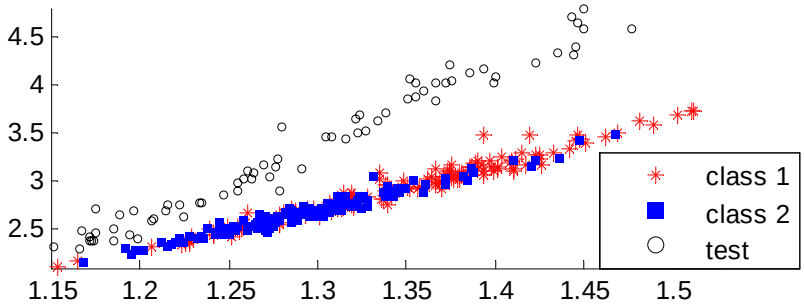

Рис.12. Средний модуль разности последовательных значений сигнала (по горизонтали) и средний квадрат (по вертикали).

На рис. 12 виден **небольшой зазор между объектами первого и второго классов**[10]. Самое интересное, что контрольная выборка тоже распадается на две

[10] **Точнее между двумя «облаками точек». В первом облаке преобладают точки первого класса, а во втором – второго.**

группы с зазором, хотя лежит в стороне от обучающей (в этом и состоит специфика нашей задачи: из-за разных условий формирования выборок они лежат в разных частях признакового пространства). Отсюда уже ясно, что надо использовать такое «распадение» контроля для его классификации.

Итак, **мы, собственно, решили задачу**! Конечно, не со 100%-й точностью (которая здесь и не достижима), но не прибегая к «высокой науке», просто просматривая картинки и фантазируя. На самом деле для аналитика в области анализа данных именно здесь и начинается наука. Необходимо убедиться, что в этой задаче лучше работают эвристики, которые оценивают интенсивность скачков сигнала, научиться их эффективно генерировать (а не перебирать вручную), найти среди них оптимальную. Кстати, подобным методом автором в 2003 году на соревновании «BCI competition III» удалось завоевать 3 место (притом, что участвовали в соревновании целые лаборатории, которые специализируются в решении подобных задач, а автор первый раз в жизни работал с сигналами). Полностью метод описывать не будем[11], отметим лишь, что точность его работы составила 86% верных ответов.

УДИВИТЕЛЬНЫЕ ЗАКОНОМЕРНОСТИ

Не отягощайте детей мёртвым грузом
фактов, обучите их приёмам и способам,
которые помогут им постягать.
(*А.Д.Сент-Экзюпери*)

Для того чтобы описанный выше метод не воспринимался как «чистое везение», отметим, что **весь процесс можно автоматизировать**, т.е. не самим разглядывать картинки в придуманных признаковых пространствах, а доверить это **ЭВМ, которая будет генерировать признаки и оценивать качество получаемых признаковых пространств** с помощью некоторого функционала (вот тут начинается математика). А мы покажем, как вручную была решена задача классификации сигналов уже другой природы (т.е. «шаманство» работает на разных данных).

На международном соревновании «Ford Classification Challenge» 2008 г. [Ford] требовалось разработать **алгоритм, который различает сигналы датчиков в автомобиле, соответствующие нормальной работе двигателя и неисправной работе**. Такой алгоритм может использоваться как детектор

[11] **В нём сначала происходит сглаживание сигнала (это такое преобразование, после которого график сигнала из хаотичного становится «более гладким»), а затем вычисляется «обобщённая скорость изменения сигнала» (учитывается не изменение значения сигнала в соседних точках, т.е. в близкие моменты времени, а изменение за небольшой промежуток времени).**

неисправности. Несколько сигналов обучающей выборки соревнования изображено на рис. 13.

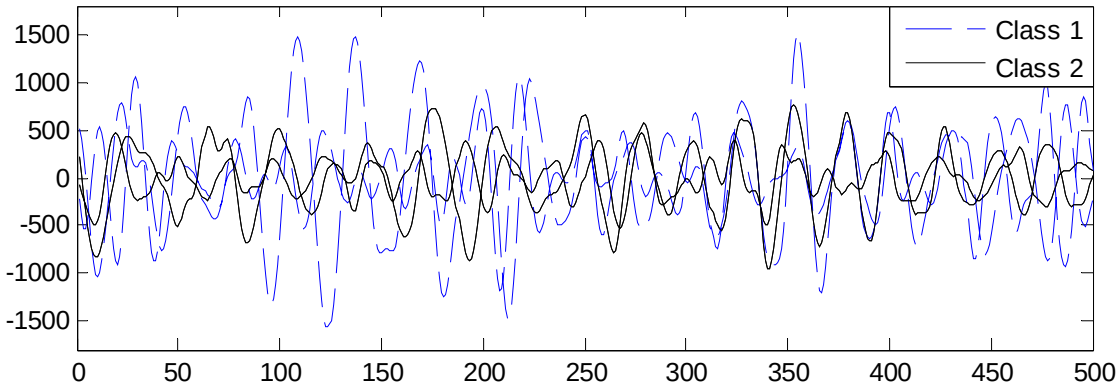

Рис. 13. Сигналы датчиков двигателя (в задаче [Ford]).

Интересно, что **в этой задаче сигналы уже «существенно неоднороны»**: среднее значение второй половины сигнала не зависит от среднего значения первой половины, см. рис. 14.

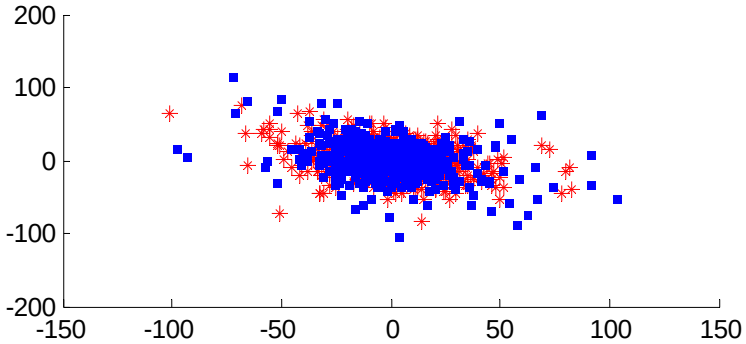

Рис. 14. Средние значения первой половины сигнала (по горизонтали) и последней (по вертикали).

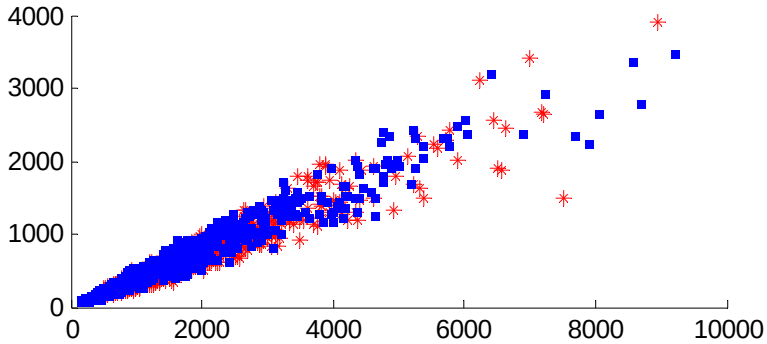

Рис. 15. Максимальные значения сигналов (по горизонтали) и дисперсии (по вертикали).

Интересна также зависимость (и даже корреляция) между выборочной дисперсией и максимальным значением сигнала, см. рис. 15.

Хотя более явно коррелируют максимальные и минимальные значения, см. рис. 16 (что, кстати, бывает достаточно часто на реальных данных).

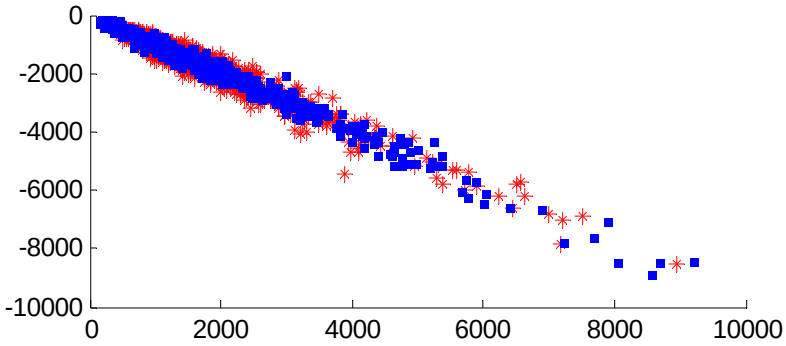

Рис. 16. Максимальные значения сигналов (по горизонтали) и минимальные (по вертикали).

На первый взгляд подобные картинки не раскрывают никаких закономерностей, но если внимательно посмотреть рис. 16, то видно, что **точки одного класса слегка «окружают» точки другого**, а если её увеличить (см. рис. 17), то видно, что **часть точек одного из классов образует плотный сгусток**. Эвристика «если максимальное значение сигнала меньше 350, то это сигнал первого класса» безошибочно относит к первому классу 622 сигнала обучения (из 3271).

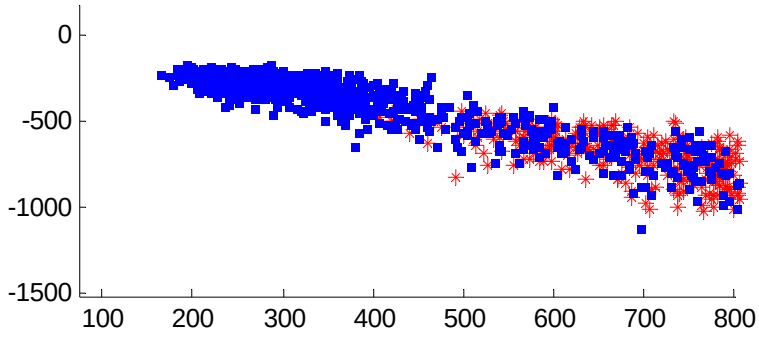

Рис. 17. Увеличенный рис. 16.

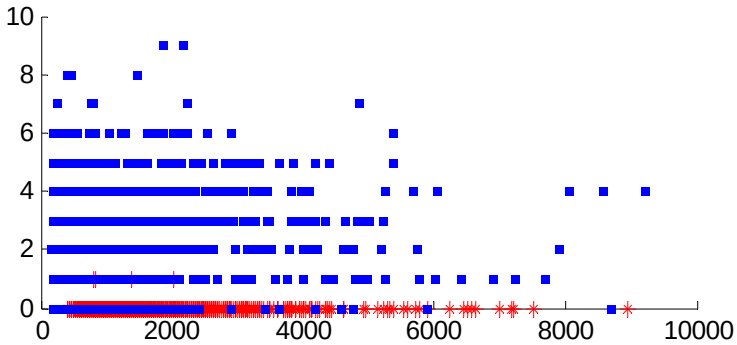

Рис. 18. Максимальные значения сигналов и количества повторов соседних значения в сигналах.

На рис. 18 видно, что также неплохим признаком оказывается

$$|\{i \in \{1,2,\ldots,n-1\} \,|\, u_i = u_{i+1}\}|$$

для сигнала $\widetilde{u} = (u_1, \ldots, u_n)$, т.е. число точек, в которых соседние значения u_i и u_{i+1} совпадают. **Раз уж мы «нащупали» такой неплохой признак, попробуем его обобщить**. Первое естественное обобщение – число незначительно отличающихся соседних точек. Второе – число повторов значений в сигнале[12]. Как видно на рис. 19, второе обобщение «работает», причём на 100%! Алгоритм, который разделяет по этому признаку, единственный из всех участников соревнования «Ford Classification Challenge» [Ford] показал стопроцентный результат верной классификации. Заметим, что этот алгоритм реализуется в математической системе MatLab[13] всего одной командой:

`2*(sum(diff(sort(X'))==0)<20)'-1`.

Это и есть «шаманство в анализе данных», когда ответ задачи кроется в 33 символах (т.е. не надо писать очень большие и сложные программы). Надо просто посмотреть на данные…

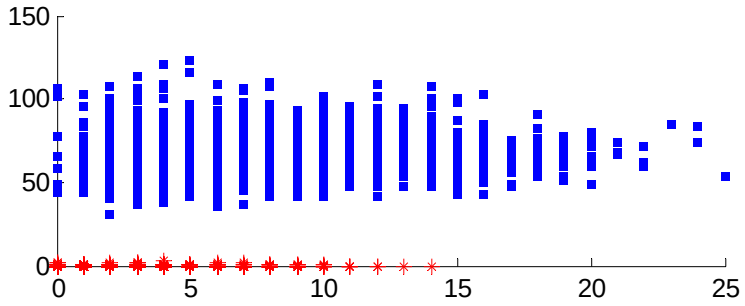

Рис. 19. Первое обобщение признака (по вертикали) и второе (по горизонтали).

ПРАВИЛА НАСТОЯЩЕГО ШАМАНА

Дорога к истине вымощена парадоксами (*О. Уайльд*).

1. Сначала надо «посмотреть на задачу».

До того как применять проверенные (и не очень) алгоритмы надо взглянуть, а что из себя представляют реальные данные. Возможно, в них есть ошибки, видные «невооружённым взглядом». Мой сокурсник, работавший в институте океанологии, рассказал мне следующую историю. Он полгода провёл на корабле в составе группы, исследовавшей дно Индийского океана. Исследователи плавали по океану и замеряли различные показатели, которые записывались в статистические таблицы. Мой сокурсник как раз писал ПО для автоматической записи в таблицы и обработки данных. После возвращения в Москву он ушёл в отпуск, а за время его отсутствие файлы с данными повредили: просто-напросто вставили лишнее значение, а остальные при этом «съехали» на чужие места, т.е. показатели, которые соответствовали одним географическим координатам, теперь были приписаны другим. Эту ошибку можно было увидеть, но группа так спешила выполнить проект и отчитаться по гранту, что даже не посмотрела на «расформатированные» таблицы. Кстати,

---

[12] **Для простоты опустим математическую формализацию этого понятия. Заметим только, что в системе MATLAB значение признака вычисляется простой командой `sum(diff(sort(X'))==0)`.**

[13] **MATLAB – система для математических вычислений компании MathWork.**

эти неправильные (!) данные были даже использованы в одной из диссертаций (до того как заметили ошибку, провели столько вычислений, что лень было получать верный результат). Поэтому **важно просто посмотреть на данные, понять, какое значение чему соответствует, изобразить это на графиках** (как было показано выше).

Ещё одна история из моего опыта. Решалась задача обработки траектории зрачка: необходимо по траектории зрачка при прочтении текста определить некоторые характеристики этого текста, например релевантность некоторой теме. Задача актуальна в проблеме анализа пользователя ЭВМ: по его действиям определить, что на экране монитора (например, при просмотре Web-страницы) он считает актуальным/устаревшим/назойливым и т.д. Одним из признаков, описывающих данные, был диаметр зрачка. Некоторые его значения были явно «неестественными»: почти сантиметр. Оказалось, что все они соответствуют одному временному интервалу. Видимо, произошёл некоторый сбой, из-за которого данные в таблице были испорчены. Поэтому при решении их лучше не учитывать.

Всё прекрасное так же трудно,
как и редко (*Б. Спиноза*).

2. У реальной задачи есть очень простое и эффектное решение.

Это изречение нельзя подтвердить (для этого надо перерешать все прикладные задачи) и опровергнуть (всегда можно сказать, что если мы знаем лишь «сложные» решения, то мы просто не нашли простое). На примерах мы видели, что действительно некоторые задачи удаётся решить «в одну строчку кода», все представленные решения легко интерпретируются, но мы рассмотрели лишь несколько реальных задач. Скорее это девиз шамана, **если он не верит в наличие «красивого решения», то задачу решать будет очень скучно**. Кстати, даже в «классической математике» некоторые учёные следовали этому правилу[14]. Правда, отыскать такое красивое решение порой очень сложно, возможно потому, что таких решений мало (может, всего одно, см. эпиграф).

Время не щадит то, что сделано без
затраты времени. (*Э. Делакруа*)

3. Решение прикладных задач требует практики.

Во многом, **анализ данных** – это действительно **не наука, а ремесло**, потому что приходится много программировать, причём эффективно программировать! Например, в задаче анализа социальной сети приходится «возиться» с огромным графом[15] (см. рис. 20). Ведь социальная сеть это граф: пользователи – это вершины, а отношения дружбы – рёбра. Число вершин может быть

---

[14] **Замечательным примером является высказывания Пауля Эрдёша о Книге, в которую Бог включает совершенные доказательства математических теорем. Попыткой представить, как могла бы выглядеть эта книга, является издание [АЦ, 2006]. Очень рекомендуем эту книгу всем любителям математики. В ней собраны доказательства очень непростых математических фактов, которые уместились буквально на одной странице.**

[15] **Те, кто не знают, что такое граф, могут представить множество точек на плоскости, некоторые пары которых соединены отрезками (см. рис. 20). Точки называются вершинами графа, отрезки – рёбрами.**

больше миллиона, а число рёбер – несколько миллионов. Алгоритм, который анализирует этот граф, не может работать вечно! Он должен работать, как это принято говорить, «за приемлемое время». Кстати, вот ещё пример задачи анализа данных – **предсказывание связности графа**. Необходимо спрогнозировать, какие рёбра в динамическом (т.е. постоянно меняющемся) графе появятся в ближайшее время. В терминах социальной сети – это предложить пользователю «потенциальных друзей», т.е. людей, с которыми скорее всего он знаком, но ещё не «зафрендил»[16].

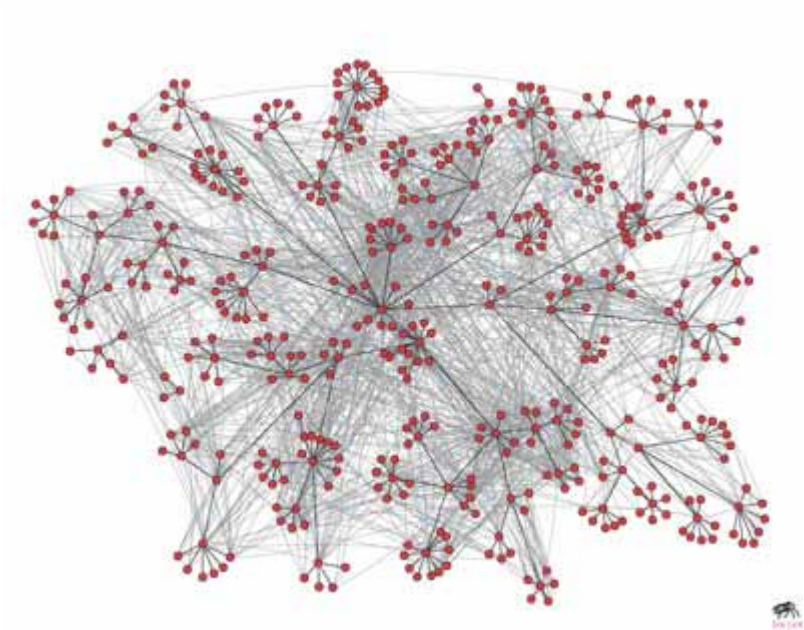

Рис. 20. Граф, соответствующий телефонным переговорам [P5].

Кстати, с точки зрения программирования очень похожи совсем разные задачи. Как-то пришлось решать **задачу иерархической классификации текстов**. Есть новостной ресурс, все новости которого хранятся в иерархической форме, как файлы в операционной системе (см. рис. 21). Есть разделы «спорт», «наука», «политика», «искусство» и т.д. (представьте, что это каталоги), в каждом из них есть подразделы (подкаталоги). Например, раздел «спорт» имеет подразделы «футбол», «хоккей», «баскетбол» и т.д. Подраздел футбол также может иметь подразделы «немецкая бундеслига», «английская премьер-лига» и т.д. На ресурс поступают новости, которые надо по этим разделам раскладывать, т.е. должен быть алгоритм, который анализирует содержание новости (например, наличие слов и словосочетаний «Аршавин», «гол», «отборочный матч» и т.д. должно ему подсказать, куда определить новость) и помещает её в нужный каталог. Даже если алгоритм будет ошибаться, лучше чтобы он делал это на нижнем уровне иерархии, например, новость о российском футбольном чемпионате поместил в новости об

[16] **Задача имеет массу коммерческих приложений. Например, к такой задаче сводится разработка алгоритма, который предлагает новые услуги клиентам банка, оператора сотовой связи и т.д. Причём к алгоритму предъявляют следующее требование: он должен «угадывать» тех, кому действительно нужна эта услуга, т.е. чтобы его предложение не сочли спамом (иначе клиент может вообще отказаться от всех услуг, обидевшись на «назойливость» компании).**

отборочном турнире чемпионата мира, поскольку если она попадёт в раздел «классическая музыка», то туда перестанут заглядывать посетители (зачем им такой «неадекватный» информационный ресурс?!). Так вот, **в этой задаче исходная информация представлялась в виде гигантской разреженной**[17] **матрицы**: по строкам были перечислены тексты, а по столбцам слова, *ij*-й элемент равнялся количеству вхождений *j*-го слова в *i*-й текст. Очень много времени ушло на написание алгоритмов такой иерархической классификации, которые бы работали быстро. Зато был приобретен очень ценный опыт: была уверенность, что быстрее уже нельзя. Через несколько лет пришлось решать задачу с социальной сетью. К удивлению, в этой задаче, которая совсем не похожа на классификацию текстов, очень пригодился опыт иерархической классификации. Ведь **здесь тоже была огромная разреженная матрица: матрица смежности графа**, в ней *ij*-й элемент равнялся единице, если *i*-й пользователь «дружил» с *j*-м, и нулю в противном случае. Причём алгоритмы в обеих задачах делают похожие операции с этими матрицами.

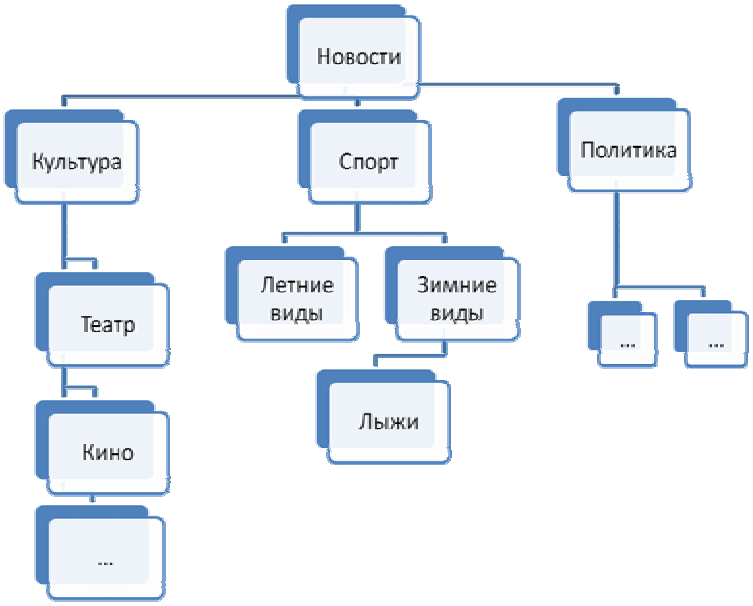

Рис. 21. Каталоги иерархической классификации [Р6].

Конечно, для того **чтобы стать хорошим специалистом в анализе данных, нужна постоянная практика**. Все примеры задач в этой лекции были взяты с Международных соревнований, которые популярны в последнее время. Участие в соревновании не только способ решить настоящую и нужную прикладную задачу, но и узнать, насколько ты хорошо её решаешь, как её решают другие специалисты и любители, сравнить разные подходы и обсудить решения на форумах. Кроме того, это шанс выиграть какой-нибудь приз, многие организаторы предлагают неплохое денежное вознаграждение тому, кто предложит лучший алгоритм. Самый яркий пример последних лет –

[17] **Матрица называется разреженной, если большинство элементов у неё нулевые. Здесь понятно, что в текст не входят все слова (поэтому в строке матрицы в основном нули) и нет слов, которые входят во все тексты (поэтому в столбце матрицы в основном нули). Конечно, есть, например, союз «и», который входит почти во все тексты, но союзы и предлоги относятся к так называемым стоп-словам, которые не рассматриваются в подобных задачах, поскольку их наличие ничего не говорит о содержании текста.**

соревнование компании Netflix [Netflix], в котором 1 миллион долларов достался команде, разработавший лучший алгоритм по рекомендации фильмов для просмотра. Сейчас появляются специальные ресурсы, на которых организуются соревнования. Очень успешный проект – [KAGGLE]. Так что, **если есть желание, дерзайте!**

ЛИТЕРАТУРА И ССЫЛКИ
слушателям лекции *«Шаманство» в анализе данных*

[ML] www.MachineLearning.ru Вики-ресурс, посвященный машинному обучению и интеллектуальному анализу данных.

[ДС, 2001] Дюк В., Самойленко А. Data Mining: учебный курс (+CD).. — СПб: Изд. Питер, 2001. — 368 с.

[В, 2010] Константин Воронцов. Курс лекций Математические методы обучения по прецедентам, МФТИ, 2004-2008. см. [ML].

[brain] http://brain.bio.msu.ru/
Сайт лаборатории нейрофизиологии и нейро-компьютерных интерфейсов биологического факультета МГУ. Собраны видео-лекции и видео-демонстрации работы интерфейса «мозг-компьютер».

http://www.univertv.ru/video/psihologiya/psihofiziologiya/shkola_v_buduwee_nauk_o_mozge_i_intellekte_1/interfejs_mozgkompyuter_teoreticheskaya_kolliziya_eksperimentalnaya_paradigma_prakticheskaya_neobhod/
Лекция А.Я.Каплана Интерфейс мозг-компьютер: теоретическая коллизия, экспериментальная парадигма, практическая необходимость или бизнес-проект? 1-я Всероссийская научная школа "В будущее наук о мозге и интеллекте". Съемка 6 ноября 2009 г. Видео, интегрированное со слайдами.

[neurofuture] http://neurofuture.ru/
Сайт научной школы «В будущее наук о мозге и интеллекте»

[BCI 3] http://www.bbci.de/competition/iii/
Сайт Международного соревнования по классификации сигналов головного мозга.

[Ford] http://home.comcast.net/~nn_classification/
Сайт соревнования по классификации сигналов работы двигателя.

[АЦ, 2006] М. Айгнер, Г. Циглер Доказательства из книги. Лучшие доказательства со времен Евклида до наших дней. – М.: Мир, 2006.

[Netflix] http://www.netflixprize.com/ Соревнование по анализу данных компании Netflix.

[KAGGLE] http://www.kaggle.com
Платформа для проведения соревнований по интеллектуальному анализу данных.

МАТЕРИАЛЫ, ИСПОЛЬЗОВАННЫЕ ПРИ ПОДГОТОВКЕ
лекции *«Шаманство» в анализе данных*

[Ж, 1998] *Журавлёв Ю.И.* Избранные научные труды. – М.: «Магистр», 1998. – 420 с.

[ДЖК, 1968] *Дмитриев А.Н., Журавлёв Ю.И., Кренделев Ф.П.* Об одном принципе классификации и прогноза геологических объектов и явлений. Известия Сиб. Отд. АН СССР, Геология и геофизика, 5, 1968, 50 – 64.

[Я, 2001] *Яблонский С.В.* Введение в дискретную математику: Учебное пособие для вузов. / Под ред. В.А. Садовничего. – 3-е изд., стер. – М.: Высш. шк.; 2001. – 384 с.

В лекции представлены данные соревнования **[BCI 3]**, которые подробно описаны в статье
Thomas Lal, Thilo Hinterberger, Guido Widman, Michael Schröder, Jeremy Hill, Wolfgang Rosenstiel, Christian Elger, Bernhard Schölkopf, Niels Birbaumer. Methods Towards Invasive Human Brain Computer Interfaces. Advances in Neural Information Processing Systems (NIPS), 2004.

[P0] Иллюстрация взята с сайта http://mk-altai.narod2.ru/turizm_na_altae/zoloto_altaiskogo_kraya/
[P1] Иллюстрация взята с сайта http://www.gizmag.com/
[P2] Иллюстрация взята с сайта http://www.lce.hut.fi/research/css/bci/
[P3] Иллюстрация взята с сайта http://www.gizmag.com/go/2084/picture/7791/
[P4] На рисунке изображена Kei Utsugi. Иллюстрация взята с сайта http://dsc.discovery.com/news/2007/06/22/gallery/brainmachine_zoom.jpg
[P5] Граф «Corporate E-Mail Communication» взят из работы
Lada A. Adamic and Eytan Adar How to search a social network. Social Networks, 27(3):187–203, 2005.
[P6] Иллюстрация взята из презентации дипломной работы выпускницы ВМК МГУ Токаревой Е.И. (работа выполнена под руководством лектора).
Также использована иллюстрация с сайта http://trans-legion.ru/
Все графики подготовлены лектором в математическом пакете MATLAB.

Также использованы материалы из раздела «ЛИТЕРАТУРА И ССЫЛКИ слушателям».

ГЛОССАРИЙ
к лекции *«Шаманство» в анализе данных*

Интеллектуальный анализ данных (data mining) – наука об обнаружении в данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности, а также процесс такого обнаружения. Подразделяется на задачи классификации, прогнозирования и другие.
(из Википедии http://ru.wikipedia.org/wiki/Data_mining)

Классификация – отнесение объекта к одному из заранее известных классов.

Эвристика (от греч. heurésko — отыскиваю, открываю),

1) специальные методы решения задач (эвристические методы), которые обычно противопоставляются формальным методам решения, опирающимся на точные математические модели. Использование эвристических методов (эвристик) сокращает время решения задачи по сравнению с методом полного ненаправленного перебора возможных альтернатив.
2) наука, изучающая эвристическую деятельность, её основной объект — творческая деятельность, развивается на стыке психологии, теории искусственного интеллекта, структурной лингвистики, теории информации.

(информация из Большой советской энциклопедии)

Интерфейс «мозг-компьютер» (Brain Computer Interface) – система, созданная для обмена информацией между мозгом и электронным устройством (например, компьютером). Здесь рассматриваются только однонаправленная система мозг-компьютер, в которой внешние устройства принимают сигналы мозга. См. также
http://ru.wikipedia.org/wiki/%D0%9D%D0%B5%D0%B9%D1%80%D0%BE%D0%B8%D0%BD%D1%82%D0%B5%D1%80%D1%84%D0%B5%D0%B9%D1%81

Граф – это совокупность непустого множества вершин и множества пар вершин (множества рёбер). Многие структуры, представляющие практический интерес в математике и информатике, могут быть представлены графами. Например, любое сообщество моделируется графом, в котором вершины – это люди, а рёбра означает отношение дружбы (или общения, наличия общего имущества, одинаковые интересы и т.д.) между двумя индивидами: если дружат, то ребро есть.