Карпов В.Э.

Интеллектуальный анализ данных

Методы ИАД

- Сети доверия (байесовские)
- Деревья решений
- Таблицы решений
- …

Основные понятия. Познавательные процедуры

- Дедукция – вывод частного заключения из общего (от общего к частному)
- Индукция – от частного к общему
- Абдукция - метод выдвижения гипотез.

Общая форма абдуктивного вывода:

1. Наблюдается аномальный, опровергающий некоторое убеждение факт С.
2. Если бы гипотеза А была истинна, факт С воспринимался бы как само собой разумеющийся.
3. Следовательно, имеется основание считать гипотезу А истинной.

- Задача ИВ заключается в выделении структурных закономерностей на основе анализа множества входных данных, т.е. движении от частного к общему.
- Обычно под ИВ понимается вывод из заданных данных объясняющего их общего правила.

Системы индуктивного вывода

- **КЛЕТКИ, Эдинбург**
  255 правил. Тележка.
  Перевернутый маятник

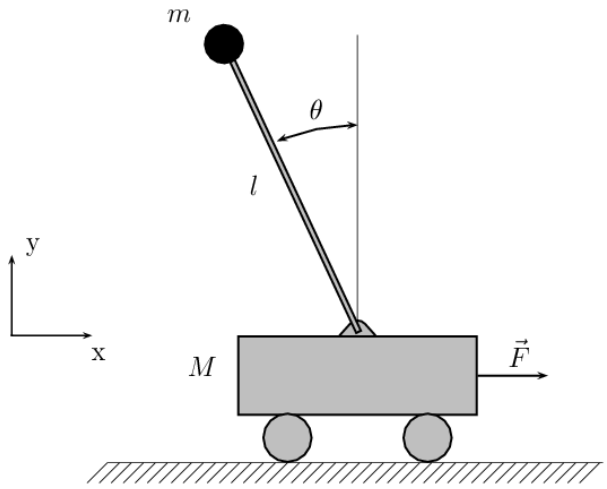


Рисунок из Википедии

Система AM/Eurisco,
Дуглас Ленат

- Lenat D. The nature of heuristics //Artificial Intelligence. 1983, N 19.
- Lenat D. The Ubiquity of Discovery, Artificial Intelligence (North-Holland), Vol.9, No.3, December 1977, p.274.

Системы индуктивного вывода

ПАСКАЛЬ, Михальски и Ларсон

«Поезда на запад и восток»

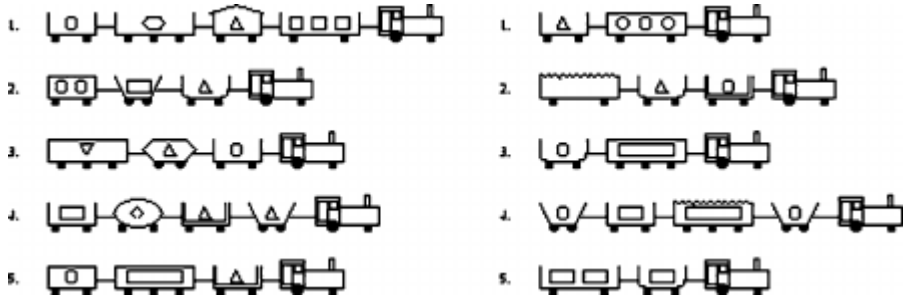
Michalski B. Theory and Methodology of Inductive Learning //Artificial Intelligence. 1983.- Vol.20, N 2.

Программа вывела два правила:

1. Если в составе есть вагон, груженный треугольником, а вагон за ним гружен многоугольником, то он идет на восток; в противном случае — на запад.
2. Если в составе ровно два вагона или если в составе есть вагон с гофрированным верхом, то он идет на запад; в противном случае — на восток.

Люди. Из общего числа 72 попыток в 44 было получено правило, а еще в 6 — вариант этого правила, основанный на подсчете осей, а не вагонов. Только в трех случаях испытуемые набрели на самое простое правило. Ни разу не было воспроизведено правило, но некоторые из полученных ответов по крайней мере не менее лаконичны. Например:

«Если состав перевозит более чем два разных типа грузов, то он идет на восток, в противном случае — на запад». Это правило было предложено двумя испытуемыми.



Таблицы принятия решений

- Способ компактного представления модели со сложной логикой. Устанавливают связь между условиями и действиями.

Условия	Варианты выполнения условий
Действия	Необходимость действий

- Условия – список возможных условий,
- Варианты выполнения условий – комбинация из выполнения и/или невыполнения условий из этого списка.
- Действия – список возможных действий,
- Необходимость действий – указание надо или не надо выполнять соответствующее действие для каждой из комбинаций условий.

Ситуация «свет погас»

Свет в соседней комнате горит	Да	Нет	Нет
Свет у соседей горит	-	Да	Нет
Поменять лампочку	X		
Проверить пробки		X	
Позвонить электрику		X	X
Позвонить диспетчеру			X

Деревья принятия решений

Дано:

Соперник	Играем	Лидеры	Дождь	Победа
Выше	Дома	На месте	Да	Нет
Выше	Дома	На месте	Нет	Да
Выше	Дома	Пропускают	Нет	Нет
Ниже	Дома	Пропускают	Нет	Да
Ниже	В гостях	Пропускают	Нет	Нет
Ниже	Дома	Пропускают	Да	Да
Выше	В гостях	На месте	Да	Нет
Ниже	В гостях	На месте	Нет	?

Необходимо выяснить, каким будет исход следующей игры

Суть метода

- Дерево принятия решений — это дерево, на ребрах которого записаны атрибуты, от которых зависит целевая функция, в листьях записаны значения целевой функции, а в остальных узлах — атрибуты, по которым различаются случаи.
- Чтобы классифицировать новый случай, надо спуститься по дереву до листа и выдать соответствующее значение.

Общая схема построения дерева решений по тестовым примерам:

1. Выбираем очередной атрибут Q, помещаем его в корень.
2. Для всех его значений i:
   - Оставляем из тестовых примеров только те, у которых значение атрибута Q равно i
   - Рекурсивно строим дерево в этом потомке

Основной проблемой является **выбор очередного атрибута**.

Для ее решения существуют различные методы (ID3, CART, C4.5 и т.д.).

Алгоритм ID3. Концепция

Дж. Куинлан (John R. Quinlan), 1986

1. Взять все неиспользованные признаки и посчитать их **энтропию** относительно тестовых образцов.
2. Выбрать признак, для которого энтропия минимальна (а информационная выгода соответственно максимальна).
3. Создать узел дерева, содержащий этот признак.

Энтропия

Определение. Предположим, что имеется множество A из n элементов, m из которых обладают некоторым свойством S. Тогда энтропия множества A по отношению к свойству S:

$$H(A,S) = -\frac{m}{n} log_2 \frac{m}{n} - \frac{n-m}{n} log_2 \frac{n-m}{n}$$

Если свойство S не бинарное, а может принимать s различных значений, каждое из которых реализуется в m_i случаях, то:

$$H(A,S) = -\sum_{i=1}^{s} \frac{m_i}{n} log_2 \frac{m_i}{n}$$

Определение. Предположим, что множество A, некоторые из которых обладают свойством S, классифицировано посредством атрибута Q, имеющего q возможных значений. Тогда **прирост информации:**

$$Gain(A,Q) = H(A,S) - \sum_{i=1}^{q} \frac{|A_i|}{|A|} H(A_i,S)$$

где A_i – множество элементов A, на которых атрибут Q имеет значение i.

Алгоритм ID3(A, S, Q)

-- A - таблица примеров, S - целевой признак, Q – множество признаков

1. Создать корень дерева
2. if S выполняется на всех A, then -- все примеры положительны
   возвратить узел с меткой «+».
3. if S не выполняется ни на одном элементе A, then --все примеры отрицательны
   возвратить узел с меткой «-».
4. -- Если множество признаков Q пустое, то возвратить узел с меткой, которая
   -- больше других встречается в значениях целевого признака в примерах.

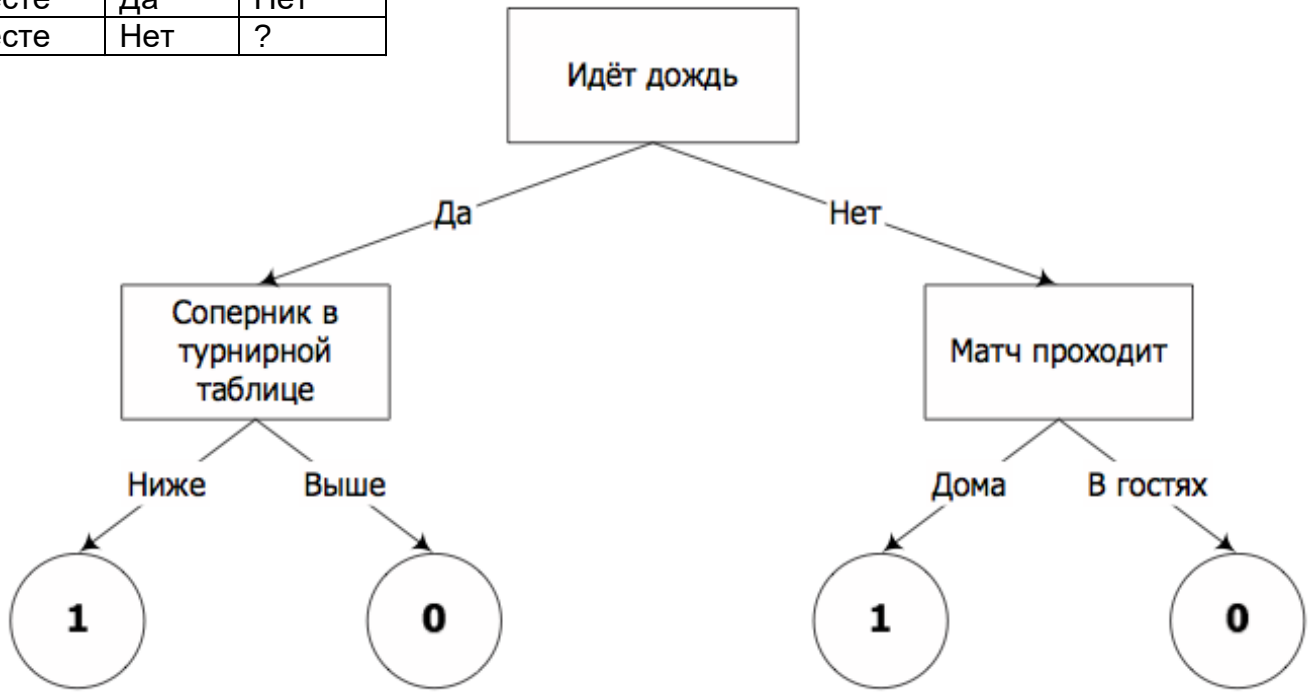
```
if Q==∅, then
    if S выполняется на большей части A, то
        поставить в корень метку «+» и выйти
    else
        поставить в корень метку «-» и выйти
 endif
endif
```

5. Выбрать $q \in Q$, для которого Gain(A,q) максимален
6. Поставить в корень метку q
7. Для каждого значения r атрибута q:

- добавить нового потомка корня и пометить соответствующее исходящее ребро меткой r
- if в A нет случаев, для которых q принимает значение r (т.е. $|A_r|=0$), then
  пометить потомка в зависимости от того, на какой части A выполняется S (аналогично п.4)
- else
- res := ID3(A_q, S, Q\{q})
- RESULT = RESULT + res

Пример

Соперник	Играем	Лидеры	Дождь	Победа
Выше	Дома	На месте	Да	Нет
Выше	Дома	На месте	Нет	Да
Выше	Дома	Пропускают	Нет	Нет
Ниже	Дома	Пропускают	Нет	Да
Ниже	В гостях	Пропускают	Нет	Нет
Ниже	Дома	Пропускают	Да	Да
Выше	В гостях	На месте	Да	Нет
Ниже	В гостях	На месте	Нет	?



Дж.С.Милль

70-х гг.XIX в., Джон Стюарт Милль (1806-1873). Понятие о причине есть корень всей теории индукции. Индуктивные методы (модусы):

- **Метод сходства.** Если два или более случая подлежащего исследованию явления имеют общим лишь одно обстоятельство, то это обстоятельство, в котором только и согласуются все эти случаи, есть причина (или следствие) данного явления.
- **Метод различия**. Если случай, в котором исследуемое явление наступает, и случай, в котором оно не наступает, сходны во всех обстоятельствах, кроме одного, встречающегося лишь в первом случае, то это обстоятельство, в котором они только и разнятся, есть следствие, или причина, или необходимая часть причины.
- **Метод сопутствующих изменений**. Всякое явление, изменяющееся определенным образом всякий раз, когда некоторым особенным образом изменяется другое явление, есть либо причина, либо следствие этого явления, либо соединено с ним какою-либо причинною связью.
- **Метод остатков**. Если из явления вычесть ту его часть, которая, как известно из прежних индукций, есть следствие некоторых определенных предыдущих, то остаток данного явления должен быть следствием остальных предыдущих.
- **Соединенный метод сходства и различия**. Если два или более случая возникновения явления имеют общим лишь одно обстоятельство и два или более случая невозникновения того же явления имеют общим только отсутствие того же самого обстоятельства, то это обстоятельство, в котором только и разнятся оба ряда случаев, есть или следствие, или причина, или необходимая часть причины изучаемого явления.

Примеры формализации индукции

Метод сходства
a&x => y b&x => y
x => y

Метод различия
a&x => y b&¬x => ¬y
x => y

Метод остатков
z&x => y&w x => w
z => y

где a, b, x, y, z, w - любые формулы.

Правдоподобные правила

Два типа правил – дедуктивные и правдоподобные.

Дедуктивное правило: истинность посылок правила *обязательно* влечет истинность заключения.

Правдоподобное правило: истинность посылок правила *не обязательно* влечет истинность заключения (истинность посылок правдоподобного правила влечет заключение в *практически значимом количестве случаев)*.

- Пример дедуктивного правила - *modus ponens*

$$\frac{\Phi,\quad \Phi \to \Psi}{\Psi}$$

- Аристотель: «Каждый человек смертен. Сократ — человек. Следовательно, Сократ смертен» (на самом деле, это не modus ponens, а рассуждение по modus Barbara, т.к. здесь имеются кванторы всеобщности)

$$\frac{\Psi,\quad \Phi \to \Psi}{\Phi}$$

- Правдоподобное правило - правило абдукции Ч.С.Пирса.
- «Испорченный» *modus ponens.* Это правило не является достоверным.

- Пример абдуктивного рассуждения: «Каждый спортсмен имеет развитую мускулатуру. Джон имеет развитую мускулатуру. Следовательно, Джон — спортсмен».
- У того, что *Джон имеет развитую мускулатуру*, должна быть причина. Наиболее правдоподобная (очевидная) причина это то, что *Джон — спортсмен*.

ДСМ-рассуждение

- ДСМ-рассуждение – это т.н. *правдоподобный вывод* (из истинных посылок выводятся не обязательно истинные заключения).
- ДСМ-рассуждение является синтезом познавательных процедур:
  - индуктивное обобщение (порождение гипотез),
  - рассуждение по аналогии (предсказание),
  - абдукция (принятие гипотез).

В.К.Финн, 1970-гг. Фармакалогия, медицина, социология, робототехника.

Суть ДСМ

ДСМ-метод оперирует сущностями трёх сортов: объектами предметной области, свойствами этих объектов и возможными причинами этих свойств.

Предполагается, что объекты имеют структуру и причинами свойств объектов являются фрагменты этой структуры.

Пример:

Объект – лист растения.

Свойство объекта – зелёный цвет.

Причина свойства – хлорофилл.

ДСМ-метод автоматического порождения гипотез

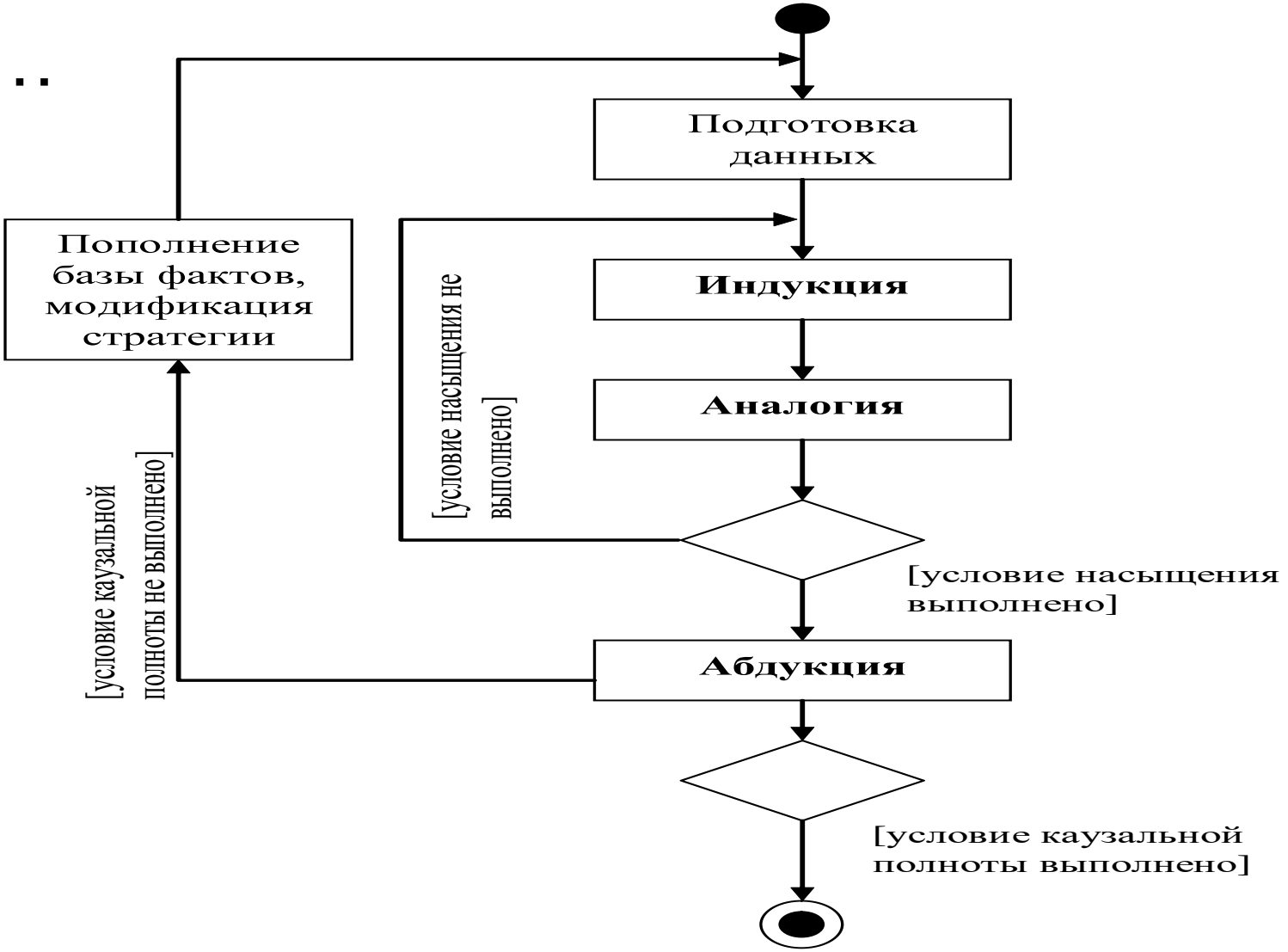
Вход: некоторое множество изучаемых объектов и сведения об их структуре, о наличии или отсутствии у них определенных свойств, а также, в некоторых случаях, о связи между структурой объектов и их свойств.

Выход: гипотезы двух типов:

- гипотезы о связи определенных структурных фрагментов изучаемых объектов со свойствами, которыми они обладают,
- гипотезы о наличии или отсутствии целевых признаков у объектов, для которых изначально это было неизвестно, формируемые на основании установленной взаимосвязи между свойствами объектов и их структурными компонентами.

ДСМ-рассуждение

- …



Пример

Исходные данные

Объект	Тип	Шерсть	Перья	Несет яйца	Дает молоко	Летающее	Водоплавающее	Хищное	Есть зубы	Позвоночное	Дышащее	Ядовитое	Есть плавники	Ноги	Хвост
Лягушка	амфибия	0	0	1	0	0	1	1	1	1	1	1	0	1	0
Тритон	амфибия	0	0	1	0	0	1	1	1	1	1	0	0	1	1
Жаба	амфибия	0	0	1	0	0	1	0	1	1	1	0	0	1	0
Курица	птица	0	1	1	0	0	0	0	0	1	1	0	0	1	1
Ворона	птица	0	1	1	0	1	0	1	0	1	1	0	0	1	1
Голубь	птица	0	1	1	0	1	0	0	0	1	1	0	0	1	1
Пингвин	**?**	0	1	1	0	0	1	1	0	1	1	0	0	1	1
Лебедь	**?**	0	1	1	0	1	1	0	0	1	1	0	0	1	1

Причины свойств

- В качестве возможных причин наличия/отсутствия свойства **p** у объектов будем рассматривать некоторые непустые подмножества множества структурных фрагментов C.

Фрагмент	**Свойства фрагмента**	**Класс**
C_1: O={Лягушка, Тритон}	C={Несет яйца, Водоплавающее, Хищное, Есть зубы, Позвоночное, Дышащее, Ноги}	Амфибия
C_2: O={Лягушка, Жаба}	C={Несет яйца, Водоплавающее, Есть зубы, Позвоночное, Дышащее, Ноги}	Амфибия
C_3: O={Тритон, Жаба}	C={Несет яйца, Водоплавающее, Есть зубы, Позвоночное, Дышащее, Ноги}	Амфибия
C_4: O={Лягушка, Тритон, Жаба}	C={Несет яйца, Водоплавающее, Есть зубы, Позвоночное, Дышащее, Ноги}	Амфибия
C_5: O={Курица, Ворона}	C={Перья, Несет яйца, Позвоночное, Дышащее, Ноги, Хвост}	Птица
C_6: O={Курица, Голубь}	C={Перья, Несет яйца, Позвоночное, Дышащее, Ноги, Хвост}	Птица
C_7: O={Ворона, Голубь}	C={Перья, Несет яйца, Летающее, Позвоночное, Дышащее, Ноги, Хвост}	Птица
C_8: O={Курица, Ворона, Голубь}	C={Перья, Несет яйца, Позвоночное, Дышащее, Ноги, Хвост}	Птица

Шаг 2. Индукция. Применение правил первого рода

Определяем функцию **H**. Множество $C_i \subseteq$ **C** будем доопределять так:

- C_i относится к (+)-гипотезе, если C_i вкладывается как подмножество в два и более (+)-примера и при этом не вкладывается ни в один (-)-пример;
- C_i относится к (-)-гипотезе, если C_i вкладывается как подмножество в два и более (-)-примера и при этом не вкладывается ни в один (+)-пример;
- C_i относится к (0)-гипотезе (противоречивой гипотезе), если C_i вкладывается как в (+)-пример, так и в (-)-пример.

Исходя из этого C1 и C2 становятся (+)-гипотезами, а C5 и C7 – (–)-гипотезами:

(+)-гипотезы (Амфибия):

- H1: {Несет яйца, Водоплавающее, Хищное, Есть зубы, Позвоночное, Дышащее, Ноги}
- H2: {Несет яйца, Водоплавающее, Есть зубы, Позвоночное, Дышащее, Ноги}

(-)-гипотезы (Птица):

- H3: {Перья, Несет яйца, Позвоночное, Дышащее, Ноги, Хвост}
- H4: {Перья, Несет яйца, Летающее, Позвоночное, Дышащее, Ноги, Хвост}

3. Аналогия. Применение правил второго рода

- Определяются τ-примеры, применяя к ним полученные (+)- и (-)- гипотезы.

Объект	**Признаки**	**Гипотезы**	**Результат**
Пингвин (?)	{Перья, Несет яйца, Водоплавающее, Хищное, Позвоночное, Дышащее, Ноги, Хвост}	Н3:{Перья, Несет яйца, Позвоночное, Дышащее, Ноги, Хвост}	(–) (Птица)
Лебедь (?)	{Перья, Несет яйца, Летающее, Водоплавающее, Позвоночное, Дышащее, Ноги, Хвост}	Н3: {Перья, Несет яйца, Позвоночное, Дышащее, Ноги, Хвост}, Н4: {Перья, Несет яйца, Летающее, Позвоночное, Дышащее, Ноги, Хвост}	(–) (Птица)

4. Абдукция. Проверка каузальной полноты

Полученные гипотезы применяются ко всем примерам. Результаты совпадают с тем, что было определено в начале, т.е. каждый исходный положительный и отрицательный пример является объясненным.

Объект	**Свойства**	**Гипотеза**	**Результат**
Лягушка (+, Амфибия)	{Несет яйца, Водоплавающее, Хищное, Есть зубы, Позвоночное, Дышащее, Ядовитое, Ноги}	{Несет яйца, Водоплавающее, Хищное, Есть зубы, Позвоночное, Дышащее, Ноги}	+ (Амфибия)
Тритон (+, Амфибия)	{Несет яйца, Водоплавающее, Хищное, Есть зубы, Позвоночное, Дышащее, Ноги, Хвост}	{Несет яйца, Водоплавающее, Хищное, Есть зубы, Позвоночное, Дышащее, Ноги}	+ (Амфибия)
Жаба (+, Амфибия)	{Несет яйца, Водоплавающее, Есть зубы, Позвоночное, Дышащее, Ноги}	{Несет яйца, Водоплавающее, Есть зубы, Позвоночное, Дышащее, Ноги}	+ (Амфибия)
Курица (–, Птица)	{Перья, Несет яйца, Позвоночное, Дышащее, Ноги, Хвост}	{Перья, Несет яйца, Позвоночное, Дышащее, Ноги, Хвост}	– (Птица)
Ворона (–, Птица)	{Перья, Несет яйца, Летающее, Хищное, Позвоночное, Дышащее, Ноги, Хвост}	{Перья, Несет яйца, Позвоночное, Дышащее, Ноги, Хвост}	– (Птица)
Голубь (–, Птица)	{Перья, Несет яйца, Летающее, Позвоночное, Дышащее, Ноги, Хвост}	{Перья, Несет яйца, Позвоночное, Дышащее, Ноги, Хвост}	– (Птица)
Пингвин (–, Птица)	{Перья, Несет яйца, Водоплавающее, Хищное, Позвоночное, Дышащее, Ноги, Хвост}	{Перья, Несет яйца, Позвоночное, Дышащее, Ноги, Хвост}	– (Птица)
Лебедь (–, Птица)	{Перья, Несет яйца, Летающее, Водоплавающее, Позвоночное, Дышащее, Ноги, Хвост}	{Перья, Несет яйца, Позвоночное, Дышащее, Ноги, Хвост}	– (Птица)

Примеры

Список параметров:

1. Высокая
2. Фигуристая
3. Красивое лицо
4. Светлые волосы
5. Скромная
6. Хорошо готовит
7. Имеет свою квартиру
8. Общительная
9. Умная

Имя	Симпатия	Номера параметров								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Света	Нравится	1	1	1	1	0	0	1	0	0
Аня	Не нравится	1	1	1	0	1	0	0	1	1
Лена	Не нравится	1	0	0	0	1	1	0	1	1
Лиля	Нравится	0	1	0	1	0	1	0	1	0
Ангелина	Не нравится	0	1	1	0	0	1	0	1	1
Даша	Нравится	0	1	1	0	0	1	1	0	1
Клава	?	1	1	0	1	1	0	1	0	0

ДСМ-метод

4 гипотезы принадлежности классу «нравится»:

1. (Света, Лиля, Даша): [Фигуристая],
2. (Лиля, Даша): [Фигуристая, Хорошо готовит],
3. (Света, Лиля): [Фигуристая, Светлые волосы],
4. (Света, Даша): [Фигуристая, Красивое лицо, Своя квартира]

4 гипотезы принадлежности классу «не нравится»:

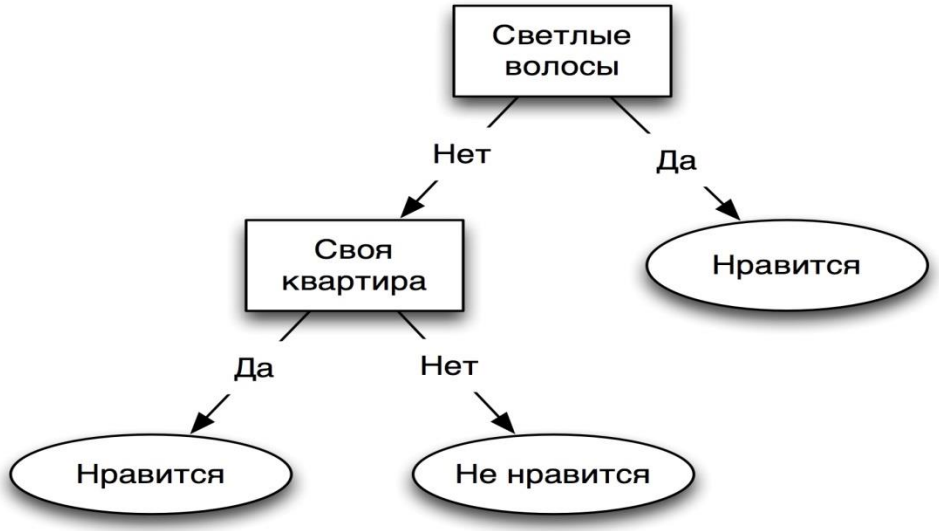
1. (Лена, Ангелина): [Хорошо готовит, Общительная, Умная],
2. (Аня, Лена): [Высокая, Скромная, Общительная, Умная],
3. (Аня, Лена, Ангелина): [Общительная, Умная],
4. (Аня, Ангелина): [Фигуристая, Красивое лицо, Общительная, Умная]

К объекту «Клава» применима одна гипотеза из первой группы: **[Фигуристая, Светлые волосы]** и ни одной гипотезы из второй группы. => Клава нравится.

Деревья решений ID3

Определение весов параметров, чтобы выбрать те из них, которые будут составлять дерево.

- На первом шаге максимальный вес (0.4591) набрали 5 параметров: «Светлые волосы», «Скромная», «Своя квартира», «Общительная», «Умная».
  Поскольку они являются равнозначными, и мы не знаем как выбор повлияет на результат, возьмем первый вариант по списку — «Светлые волосы». Это - корень нашего дерева.
- На втором шаге наибольший вес (0.8112) набрали параметры «Своя квартира» и «Общительная». Снова берем первый вариант по списку



=> Клава нравится

Выбор подходящих методов

Метод	Тип	Тип вх. данных	Скорость	Устойчивость к шумам	Сложность реализации
ДСМ-метод	По подобию	Булевы	Низкая	Средняя	Средне
ID3	Иерархический	Булевы	Высокая	Высокая	Средне, готовые решения
ДПР-2	Иерархический, вероятностный	Булевы	Средняя	Высокая	Просто
Нейросеть	Вероятностный	Числовые	Низкая	Высокая	Сложно, готовые решения
kNN	Вероятностный, по подобию	Числовые	Средняя	Средняя	Просто
Таблица подобия	Вероятностный, по подобию	Булевы	Средняя	Низкая	Просто

Тестовые примеры

- Коллекции тестовых наборов данных Калифорнийского Университета UCI Machine Learning Repository

Заключение

Прочие методы

- Нейронные сети
- Эволюционные модели
- Статистические (байесовские) методы

…

- Классификация?
- Распознавание?
- Индуктивный вывод?
- Приобретение знаний?