Интеллектуальный анализ данных, основанный на формальных уточнениях понятия сходства

Забежайло М.И.

ФИЦ ИУ РАН

Константин Владимирович **Рудаков**

(1954 – 2021)

Два эффекта
в задачах машинного обучения и интеллектуального анализа данных

- **Динамика**
  (изменения обучающей выборки, изменения объектов в обучающей выборке)

- **Устойчивость**
  (формируемых в динамической среде решений)

Компьютерный анализ данных и методы ИИ

- Особая роль ***эвристик*** в компьютерном анализе данных
- Задача выделения ***областей*** «*теоремно-доказуемой*» ***корректности*** таких эвристик. Проблема ***открытости*** анализируемой предметной области
- *Эвристика* ***сходства***:

  сходство *описаний* ***прецедентов*** *наличия целевого эффекта можно рассматривать как приближенное* ***описание причинных влияний****, вызывающих этот эффект*
- *Интегрируемы* ли ***АВО+*** и ***ДСМ-метод***? Если – да, то *как*?

Ю.И.Журавлев: неклассическая *интерполяция-экстраполяция* в компьютерном анализе данных

- Задачи Анализа Данных и Поддержки Принятия Решений, в которых *заключение* о *будущем* формируется на основе *обучения* на *прецедентах* из *прошлого* и *настоящего*
- Интерполяционно-экстраполяционные схемы АД и ППР:
  - Набор прецедентов (описанных тем или иным способом) – ***обучающая выборка***
  - ***Класс*** (вид) эмпирических ***зависимостей*** для описания прецедентов
  - ***Интерполяция*** обучающей выборки зависимостями выбранного класса
  - ***Прогнозирование*** свойств новых прецедентов проверкой ***экстраполируемости*** на них эмпирических зависимостей, найденных при интерполяции обучающей выборки

Классы И-Э-зависимостей на базе формальных уточнений понятия *сходство*

- Неформальная идея сходства. Отношение сходства.

рефлексивность $\dfrac{\mathbf{a}}{\mathbf{a}\ \boldsymbol{R}\ \mathbf{a}}$

симметричность $\dfrac{\mathbf{a}_1\ \boldsymbol{R}\ \mathbf{a}_2}{\mathbf{a}_2\ \boldsymbol{R}\ \mathbf{a}_1}$

- **Метрическое сходство**

  Метрика: числовая функция ρ(x,y), удовлетворяющая трем свойствам

  1. $0 \le \rho(x,y)$, причем $\rho(x,y)=0$ означает, что x=y
  2. $\rho(x,y) = \rho(y,x)$
  3. $\rho(x,z) \le \rho(x,y) + \rho(y,z)$

  Два объекта сходны, если их они **расположены близко** (расстояние между ними не превышает заданной величины)

- **Топологическое сходства**

  Два объекта сходны, если их область их различия имеет **малую меру**

- Сходство как **бинарная алгебраическая операция**

  (1) a * a = a
  (2) a * b = b * a
  (3) a * b * c = (a*b)*c = a*(b*c)

Отношение сходства
в *интерполяционно-экстраполяционной* схеме анализа данных

- Процедурная схема проверки *экстраполируемости* ЭЗ, сформированных в ходе *интерполяции* обучающей выборки
  - Отношение сходства, порождаемое в процессе формального уточнения понятия сходства
  - Классы сходства
  - Классы эквивалентности
  - Отнесение к классам эквивалентности
- Проблема идентификации и отсечения артефактов

Интеллектуальный Анализ Данных и Машинное Обучение (ML)

- Классическая схема ML:

  *обучение (**интерполяция**) => тестирование => прогноз (**экстраполяция**)*

- **ИАД**: необходимость перехода вдоль последовательности ***расширений*** стартовой ***обучающей выборки*** в (пример: задачи диагностического типа)
  - Обычная ситуация в экспериментальной лаборатории:

    начинать приходится с анализа выборок того размера, которые удалось собрать к данному моменту. => ***Пополнения***. К чему это ведет? (Риски)
  - Задачи с «***последствиями***» принимаемых ***управленческих решений***
  - На что (на какие факторы) ориентировать управляющие воздействия? Проблема ***наследуемости*** таких воздействий

Динамические изменения обучающей выборки

- Проблема ***устойчивости*** интерполяционных ЭЗ при ***расширении*** обучающей выборки

  - ***ABO+*** (метрические представления о сходстве)
  - ***ДСМ-метод*** (сходство как бинарная алгебраическая операция)

- Метрические уточнения сходства

  Поиск ***наследуемых*** при расширении обучающей выборки ***корректных алгоритмов***

- Сходство как алгебраическая операция

  Математическая техника ***до-определения*** значений ***частично-определенных*** (на прецедентах обучающей выборки) ***характеристических функций***

Некоторые математические задачи (*ABO*+)

Дано:

- **FB** (множество прецедентов)
- последовательность $\Delta\mathbf{FB}_1$, $\Delta\mathbf{FB}_2$, … , $\Delta\mathbf{FB}_n$ заданных расширений для **FB** (заданных подмножеств новых прецедентов)

Найти:

- ***Емкость*** (число элементов) ***множества корректных алгоритмов***, точно интерполирующих исходную **FB**
- ***Корректный алгоритм***, ***наследуемый*** при заданном расширении исходной выборки прецедентов.
- ***Необходимые*** и ***достаточные*** условия ***существования*** корректного алгоритма (Теорема существования):
  - для *заданной* последовательности **FB**, $\Delta\mathbf{FB}_1$, $\Delta\mathbf{FB}_2$, … , $\Delta\mathbf{FB}_n$
  - в общем случае
- ***Емкость множества*** таких ***устойчивых*** к заданному расширению ***корректных алгоритмов***
- …

Характеристические Функции (*ХФ*) на последовательностях расширяющихся Баз Фактов

- Характеристическая Функция (***ХФ***), принимает
  - значение «*истина*» на всех фактах **ф** (примерах) текущей базы фактов **FB**, характеризуемых наличием анализируемого целевого свойства и
  - значение «*ложь*» на всех фактах **ф** (контрпримерах) текущей базы фактов **FB**, характеризуемых наличием анализируемого целевого свойства

- Характеристические Функции как (*каузально-ориентированный*) подкласс семейства частичных функций, интерполирующих выборки прецедентов – примеров и контрпримеров диагностируемого явления. Пополнение текущей **FB** описаниями новых прецедентов обеспечивает до-определение некоторых ранее не определенных значений ***ХФ***

- Представление о наследуемости ***ХФ*** при расширении **FB** описаниями новых прецедентов. Особая роль контрпримеров

Некоторые математические задачи (*ДСМ-метод*)

Дано:

- **FB** (множество прецедентов)
- последовательность $\Delta\mathbf{FB}_1$, $\Delta\mathbf{FB}_2$, … , $\Delta\mathbf{FB}_n$ заданных расширений для **FB** (заданных подмножеств новых прецедентов)
- новый прецедент ϕ

Найти:

1. ***Репрезентативна*** ли исходная **FB**: существует ли ***характеристическая функция***, точно интерполирующая данную **FB**
2. ***Емкость*** (число элементов) ***множества характеристических функций***, точно интерполирующих исходную **FB**
3. ***Существует ли характеристическая функция***, ***наследуемая*** при заданном расширении исходной выборки прецедентов
4. ***Характеристическую функцию***, ***наследуемую*** при заданном расширении исходной выборки прецедентов
5. ***Необходимые*** и ***достаточные*** условия ***существования характеристической функции, наследуемой*** при заданном расширении исходной выборки прецедентов (*Теорема существования*):
   - для *заданной* последовательности **FB**, $\Delta\mathbf{FB}_1$, $\Delta\mathbf{FB}_2$, … , $\Delta\mathbf{FB}_n$
   - в *общем* случае

Некоторые математические задачи (*ДСМ-метод*) - 2

Дано:

- **FB** (множество прецедентов)
- последовательность $\Delta\mathbf{FB}_1, \Delta\mathbf{FB}_2, \ldots, \Delta\mathbf{FB}_n$ заданных расширений для **FB** (заданных подмножеств новых прецедентов)
- новый прецедент ϕ

Найти:

6. ***Емкость множества*** таких ***устойчивых*** к заданному расширению ***характеристических функций***
7. ***Существует ли характеристическая функция***, экстраполируемая на заданный новый прецедент ϕ, ***наследуемая*** при заданном расширении исходной выборки прецедентов и ***сохраняющая экстраполируемость*** на этот прецедент ϕ
8. ***Емкость множества*** таких ***устойчивых*** к заданному расширению и ***сохраняющих экстраполироуемость*** на заданный прецедент ϕ ***характеристических функций***
9. …

Некоторые комментарии (*ДСМ-метод*)

- Может быть попробуем перебрать все характеристические функции, интерполирующие обучающую выборку **FB**?

Вряд ли:

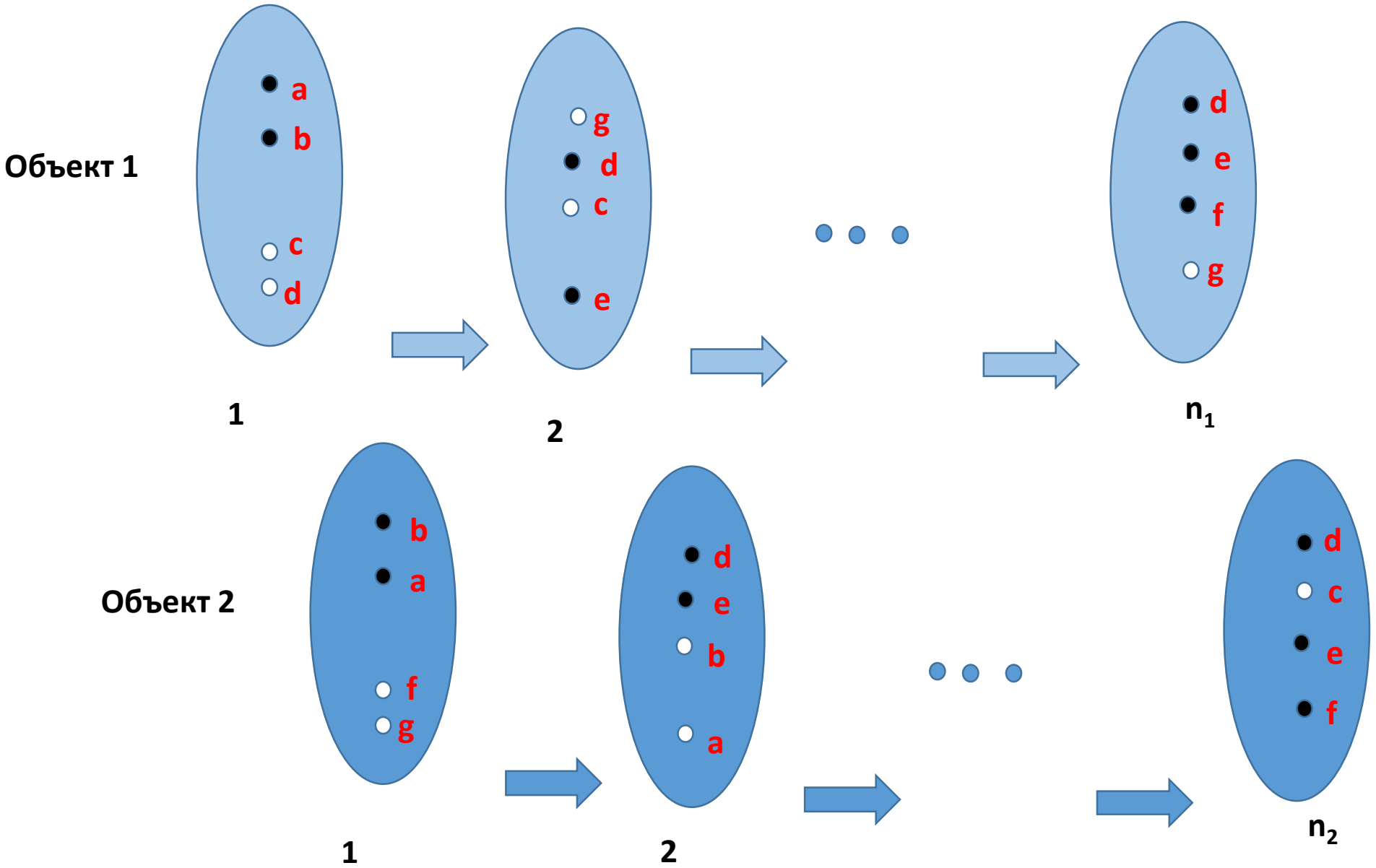
- Емкость множества характеристических функций, точно интерполирующих обучающую выборку прецедентов **FB**, в общем случае растет ***экспоненциально быстро*** при линейном росте размеров **FB**
- Емкость множества характеристических функций, точно интерполирующих обучающую выборку прецедентов **FB** и при этом экстраполируемых на новый прецедент ϕ , в общем случае растет ***экспоненциально быстро*** при линейном росте размеров **FB**
- Существуют такие **FB** и **FB**$\cup\Delta$**FB**, что множества характеристических функций, интерполирующих каждую из них, растут ***экспоненциально быстро*** и ***не имеют общих элементов***
- Существуют такие **FB** и **FB**$\cup\Delta$**FB**, что множества характеристических функций, интерполирующих каждую из них и экстраполируемых на заданный новый прецедент ϕ , растут ***экспоненциально быстро*** и ***не имеют общих элементов***

Однако:

- Задача о *репрезентативности* произвольной обучающей выборки **FB** эффективно разрешима
- Задача о существовании наследуемой при переходе от **FB** и **FB**$\cup\Delta$**FB** характеристической функции, сохраняющей экстраполируемость на заданный новый прецедент ϕ, эффективно разрешима

Учет динамики изменений объектов в обучающей выборке

Описания объектов в виде последовательностей состояний

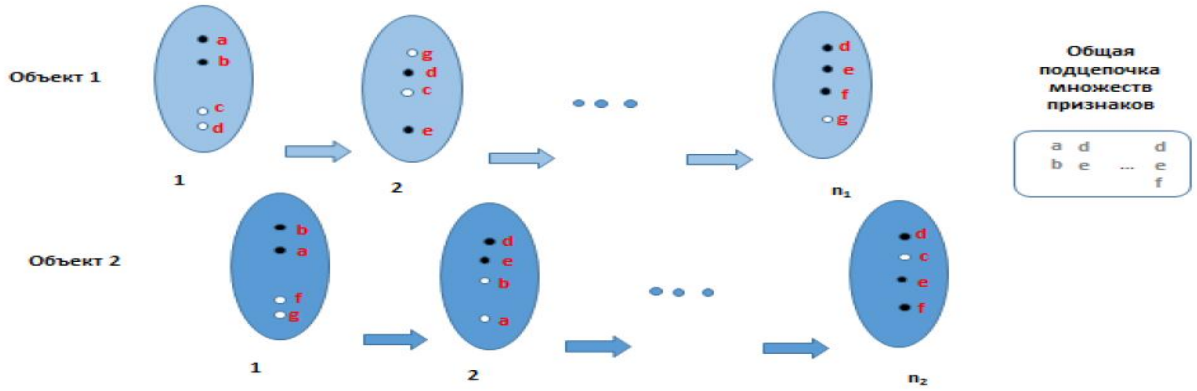


Общая подцепочка множеств признаков

a	d		d
b	e	...	e
			f

Динамические описания объектов (прецедентов) в обучающей выборке

- Описание объекта как ***последовательность*** (цепочка) его ***состояний***



- **Состояние**:
  - множество признаков,
  - множество признаков с отношения на его элементах,
  - граф с числовыми метками на ребрах и вершинах,
  …
- Цепочки состояний: ***сходство*** на цепочках – множество ***общих подцепочек*** (в цепочках состояний, описывающих динамику поведения объектов)

Динамические изменения объектов-прецедентов в обучающей выборке

Проблема ***устойчивости***

порождаемых ***эмпирических зависимостей***

при расширении обучающей выборки

Заключение

Что в этом (базирующемся на уточнении понятия сходства) подходе особенного с точки зрения приложений?

- Содержательная ***интерпретируемость*** и
- Неформальная ***объясняемость*** –

  ответы не только на вопрос ***КАК***?, но и на вопрос ***ПОЧЕМУ***?

результатов компьютерного анализа данных

Спасибо за внимание

m.zabezhailo@yandex.ru