Об основах метрического анализа плохо формализованных задач распознавания и классификации

И. Ю. Торшин, К.В.Рудаков

Московский Физико-Технический Институт

Вычислительный центр им. А.А. Дородницына РАН

*алгебраический подход, теория классификации **значений признаков**, комбинаторная теория разрешимости, метрический анализ данных, метрические сгущения*

Рамки алгебраического подхода к проблемам распознавания и классификации

- **Заданы**
  - матрица информации (набор признаковых описаний объектов)
  - информационная матрица (отнесение объектов к определенным классам)
- **Исследовать:**
  - Разрешимость/регулярность задач
  - Корректность/полноту моделей алгоритмов

Формализованные задачи – задано формальное описание

- ***Задача формализована*** - заданы матрица информации и информационная матрица

- ***Множество прецедентов*** - совокупность пар соответствующих строк матрицы информации и информационной матрицы

Построение алгоритмов распознавания

Плохо формализованные задачи

- Формальные описания существенно различны, даже могут быть получены существенно различными способами на основании некоторого «исходного описания» задачи в терминах проблемной области.
- На основе имеющегося «исходного описания» не существует однозначного метода определения
    1) **Объектов**
    2) **Признаковых описаний объектов**
    3) **Классов объектов**

Процедуре формализации той или иной прикладной задачи редко уделяется должное внимание…

- Адекватная формализация - существенное улучшение аккуратности и обобщающей способности алгоритмов распознавания

- Примеры из различных областей: анализ биомедицинских данных, биоинформатика, хемоинформатика, анализ текстов

Примеры плохо формализованных задач

Хемоинформатика

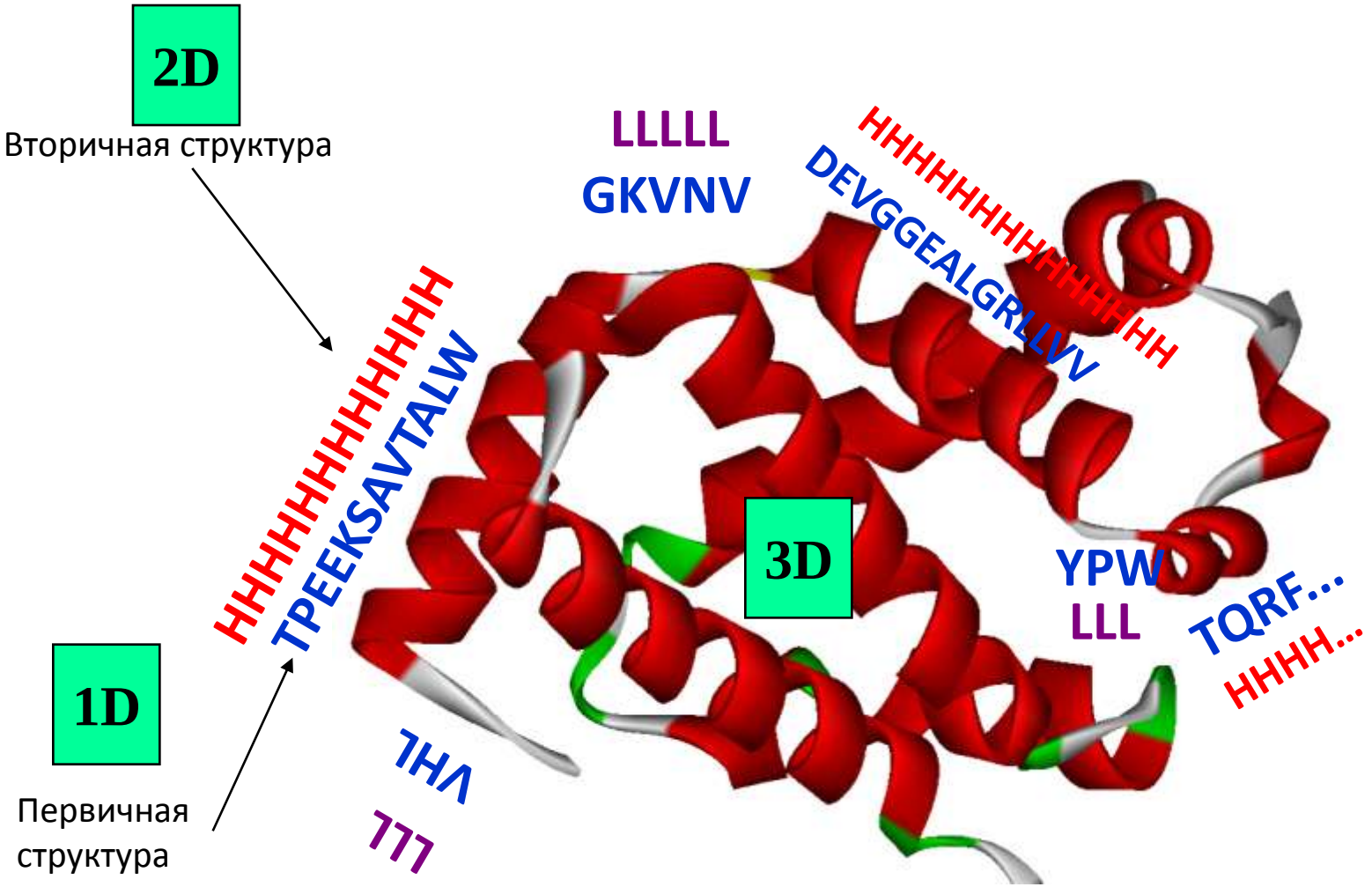
Биоинформатика

Задача	Однозначность формализации		
	Объекты	Классы	Признаки
Распознавание вторичной структуры белка	-	-	-
Аннотация генома	+	-	-
Поиск структурно схожих молекул	+	+	-
Поиск релевантных научных публикаций	+	+	-
Анализ биомедицинских данных	+	-	+-

Биомедицина

При адекватной формализации этих задач в рамках соответствующих проблемно-ориентированных теорий было достигнуто существенное улучшение качества распознавания.

Уровни структуры белка



Задача распознавания вторичной структуры белка

- распознавание = *перевод последовательности из 20-буквенного алфавита в 3-буквенный*

VHLTPEEKSA…

↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓

LLLHHHHHHH…

Алфавит «A»

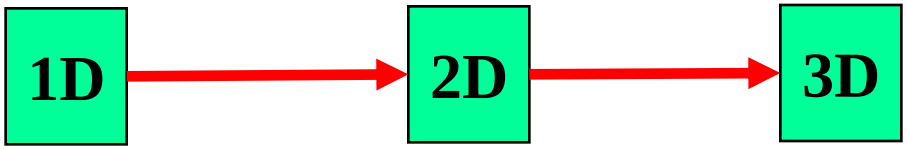
1D

A={A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y}

Алфавит «B»

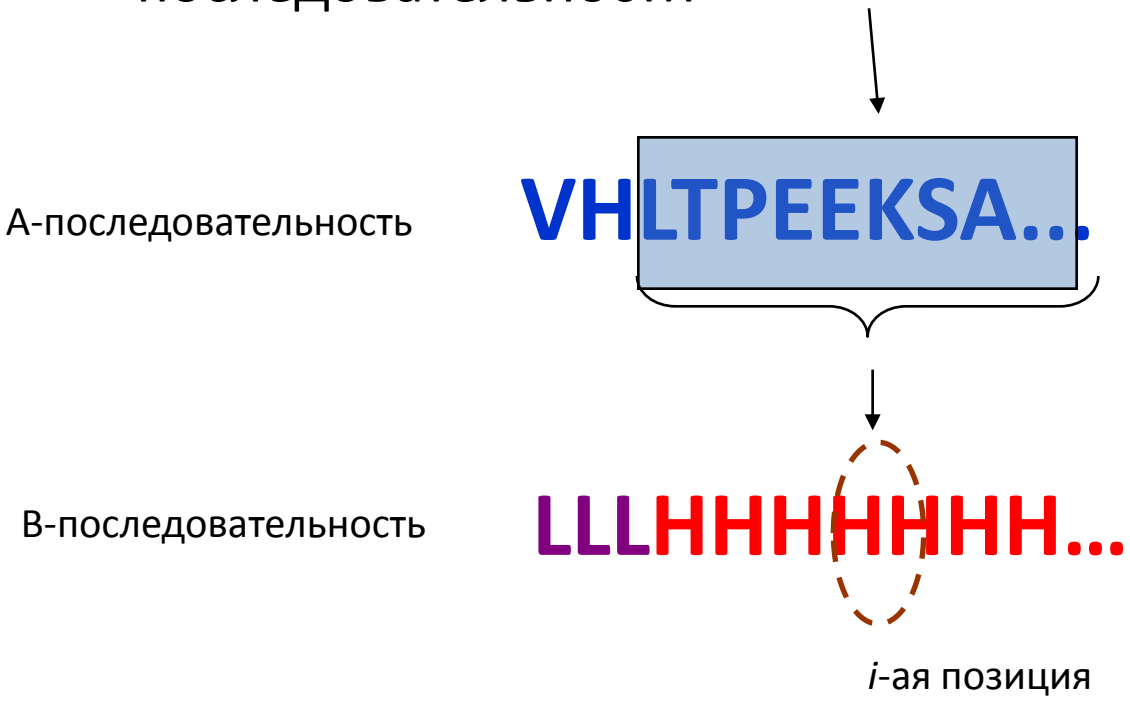
2D

B={H,S,L}



Гипотеза локальной разрешимости задачи позволяет ввести «объекты» задачи распознавания

- В *i*-ой позиции, состояние B-последовательности определяется локальным контекстом (**окрестностью**) в A-последовательности



Формальное описание локальности

слово $U = \{u_1, u_2, ..., u_n\}$

ведущая позиция i, $1 \le i \le n$

«маска» $\hat{m} = \{\mu_1, \mu_2, ..., \mu_m\}$

позиции маски $\mu_i \in Z$

$\mu_1 < \mu_2 < ... < \mu_m$

размерность маски

$|\hat{m}| = m$

протяженность маски

$[\hat{m}] = \mu_m - \mu_1 + 1$

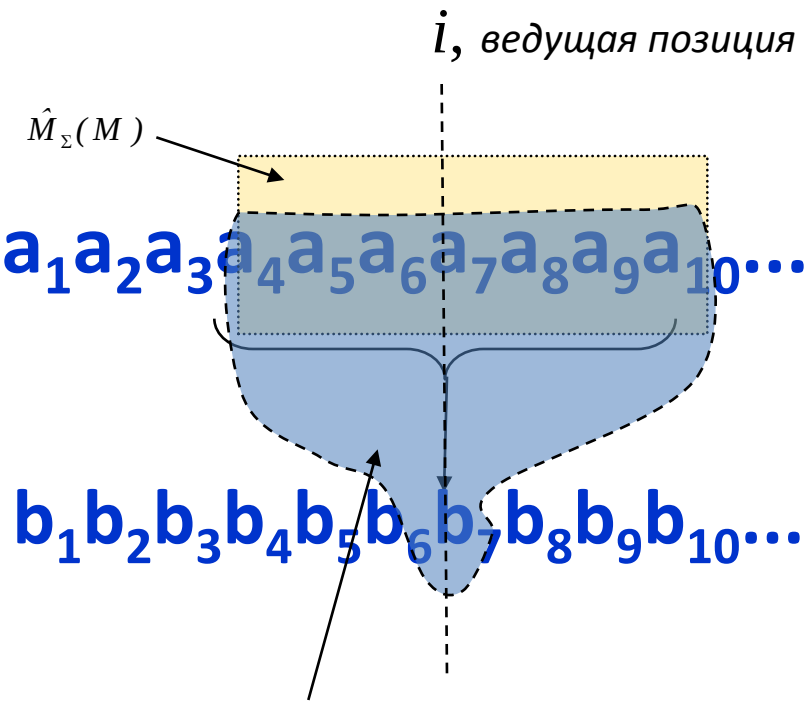
оператор выбора подслова $\eta(i, \hat{m}, U)$

$$\eta(i,\hat{m},U) = \begin{cases} u_{i+\mu_1} u_{i+\mu_2} .. u_{i+\mu_m}, & \text{если } i+\mu_1 \ge 1, i+\mu_m \le n, \\ \varnothing \quad \text{в противном случае.} \end{cases}$$

система масок $M = \{\hat{m}_1, \hat{m}_2, \ldots, \hat{m}_N\}$ $\mathrm{M}_{\mathrm{n}}^{\mathrm{m}}$

объединенная маска

$$M_\Sigma(M) = \bigcup_{k=1}^{|M|} \hat{m}_k$$



Элементарный объект q

$$q = \left(\eta\ (i, \hat{M}_\Sigma(M), V),\ W_i\right) \in Q$$

Q=Q(Pr,M) - множество объектов

Мотив - пара из подслова и соответствующей маски

$$\kappa_\alpha = (\hat{m}_\alpha, \vec{v}_\alpha)$$

(Рудаков, Торшин, 2011)

Условие *локальной* разрешимости

$$\forall_{\Pr}(V^1,W^1),(V^2,W^2):(V^1=V^2)\Rightarrow(W^1=W^2)$$

Теорема 1

$\hat{m}=\{\mu_1,\mu_2,\ldots,\mu_m\}$

$\eta(i,\hat{m},U)$

$M=\{\hat{m}_1,\hat{m}_2,\ldots,\hat{m}_N\}$

$Q=Q(Pr,M)$

$\kappa_\alpha=(\hat{m}_\alpha,\vec{v}_\alpha)$

$$\forall_{Q(Pr,M)}(i,j):w_i\neq w_j\Rightarrow\exists_{\alpha=1}^{|K_1|}\alpha:(\kappa_\alpha\in\bar{v}_i)\neq(\kappa_\alpha\in\bar{v}_j)$$

…

Примеры плохо формализованных задач

Хемоинформатика

Задача	Однозначность формализации		
	Объекты	Классы	Признаки
Распознавание вторичной структуры белка	-	-	-
Аннотация генома	+	-	-
Поиск структурно схожих молекул	+	+	-
Поиск релевантных научных публикаций	+	+	-
Анализ биомедицинских данных	+	-	+-

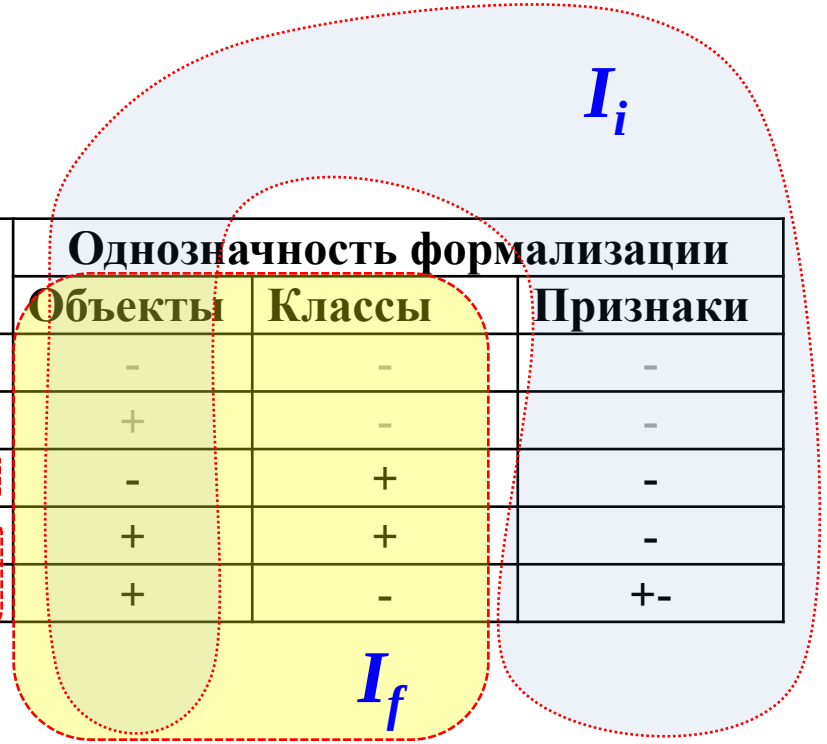
Биомедицина

процедуры преобразования признаковых описаний

О множествах начальных и конечных информаций в плохо формализованных задачах

Хемоинформатика

Биомедицина



Задача	Однозначность формализации		
	Объекты	Классы	Признаки
Распознавание вторичной структуры белка	-	-	-
Аннотация генома	+	-	-
Поиск структурно схожих молекул	-	+	-
Поиск релевантных научных публикаций	+	+	-
Анализ биомедицинских данных	+	-	+-

I_f

Формализация задачи поиска «оптимальных» I_f

$$\mathrm{Pr} \subseteq \mathrm{I_i} \times \mathrm{I_f}, \quad \mathrm{Pr} \neq \varnothing, \quad \mathrm{Pr} = \{(\alpha, \beta)\}$$

разрешимость задачи

$$\forall_{Pr} (\alpha_1, \beta_1), (\alpha_2, \beta_2) : \alpha_1 = \alpha_2 \Rightarrow \beta_1 = \beta_2$$

множество признаковых описаний прецедентов

$$Q(\mathrm{Pr}) = \{\alpha_1, \alpha_2, ..., \alpha_N\}, \quad Q(\mathrm{Pr}) \subseteq \mathrm{I_i}$$

множество описаний классов

$$B(\mathrm{Pr}) = \{\beta \mid \exists \alpha, (\alpha, \beta) \in \mathrm{Pr}\} \cup \Delta$$

Теорема 1. *Задача разрешима тогда и только тогда, когда совокупность β-подмножеств признаковых описаний над B(Pr) является разбиением Q(Pr).*

оператор выделения β-подмножества признаковых описаний

$$\hat{\alpha}(\beta, \mathrm{Pr}) = \{\alpha_i \mid (\alpha_i, \beta_i) \in Pr, \beta_i = \beta\}$$

оператор покрытия

$$\breve{\alpha}(\mathrm{B}, \mathrm{Pr}) = \bigcup_{k=1}^{|B|} \hat{\alpha}(\beta_k, \mathrm{Pr})$$

(ξ,γ)-разбиение – учет противоречивости реальных множеств прецедентов.

$$B = \{\beta_1, \beta_2, ..., \beta_m\}$$

$$\begin{cases} \left|Q(\mathrm{Pr}) \cap \breve{\alpha}(\mathrm{B}, \mathrm{Pr})\right| \geq \xi \cdot \left|Q(\mathrm{Pr})\right| \\ \forall_{k=1}^{|B|} \left|\hat{\alpha}(\beta_k, \mathrm{Pr}) \cap \breve{\alpha}(\mathrm{B} \setminus \beta_k, \mathrm{Pr})\right| \leq \gamma \cdot \left|Q(\mathrm{Pr})\right| \end{cases}$$

Принцип максимального покрытия

Принципом максимального покрытия назовём утверждение о максимальном вкладе α-го элемента словаря $B_j \subset I_{j,\delta(j)}(\tilde{B})$, α > 1, в покрытие *Q(Pr')*

$$\forall_{i=j}^{j+\delta(j)} \beta_i, \quad \left|\breve{\alpha}(I(B_j,1..\alpha),Pr')\right| \geq \left|\breve{\alpha}(I(B_j,1..\alpha-1),Pr') \cup \hat{\alpha}(\beta_i,Pr')\right|$$

Lm 4. $\sum_{l=1}^{|B|} \left|\hat{\alpha}(\beta_l,Pr')\right| \geq \left|Q(Pr')\right|$

- необходимое условие покрытия

Th. 5. *Принцип максимального покрытия ведущих позиций – необходимое условие оптимальности словаря.*

Упорядоченный список $I(\tilde{B}(n))=\{\beta_1, \beta_2, \ldots, \beta_j, \ldots\}$

$\left|\hat{\alpha}(\beta_1,Pr')\right| \geq \left|\hat{\alpha}(\beta_2,Pr')\right| \geq \left|\hat{\alpha}(\beta_j,Pr')\right| \geq \ldots$

δ(ξ)-покрытие $I_{j,\delta}(\tilde{B}(n)) \subseteq I(\tilde{B}(n))=\{\beta_j, \beta_{j+1}, \ldots, \beta_{j+\delta-1}\}$:

(7) $\left|\breve{\alpha}(I_{j,\delta}(\tilde{B}),Pr')\right| \geq \xi \cdot \left|Q(Pr')\right|$

ξ-тупиковое покрытие B^T - (7) нарушено для любого подмножества B^T.

Th. 6. *δ(ξ)-покрытия содержат ξ-тупиковые покрытия.*

Th. 7. *Принцип максимального покрытия – необходимое, а при* $\xi'(B_j,Pr') \leq \xi \quad \gamma'(B_j,Pr') \leq \gamma$ *достаточное условие ξ,γ-разбиения*

→ **Теорема 1**

Th. 8. *Множество оптимальных словарей есть подмножество множества Г(ξ), каждый элемент которого удовлетворяет принципу максимального покрытия.*

Еще одна форма записи критерия разрешимости…

I_i?

Формализм для проведения метрического анализа плохо формализованных задач

- Аксиома соответствия
- Топологии и решётки над множеством исходных описаний
- Метрические пространства
- Приложения:
  - *введение метрик на множествах признаковых описаний,*
  - *метрики на множествах объектов*
  - *анализ «взаимодействий» разнородных признаковых описаний.*

Об аксиоме (однозначного) соответствия

- $X=\{x_i\}$ – конечное множество исходных описаний объектов в проблемной области
- $\{m_i\}$ матрица информации, $\{\iota_i\}$ - информационная матрица
- $Q = \{q_i | q_i = (m_i, \iota_i)\}$ - множество «объектов ИАД»
- Формализация задачи - нахождение функции $\varphi : X \to Q$
- Аксиома соответствия – ограничения на ф:
  - Слабая форма: ф – инъективна
  - **Сильная форма: ф – биекция**

О топологиях над множеством исходных описаний

- Процесс нахождения ф:
  - От множества X к некоторой топологии τ(X),
  - От τ(X) к τ-пространству T(X)=(X,τ(X))
  - От T(X) к решётке L(X)=L(T(X))
  - От L(X) к метрическому пространству M(τ(X))=(τ(X),$\rho_{L(X)}$).
  - От M(τ(X)) к Q(X)
- По сильной форме аксиомы соответствия - существуют биекции между всеми этими представлениями X

Пример. X - множество описаний пациентов

- Разбиения на неупорядоченные подмножества
  - описаний мужчин и женщин,
  - подмножества, соответствующие тому или иному симптому,
  - числовые и «категориальные» признаки: подмножества, соответствующие *значениям признаков*
  - …

Формализуем…

- Над множеством исходных описаний X задается π(X) - система подмножеств X $\pi(X) = \{\varnothing, a_1, a_2, \ldots, a_n \mid a_i \subseteq X\}$
  - объединение элементов π(X) $\breve{\pi}(X) = \{\bigcup_{i=1..n} a_i\}$
  - пересечение элементов π(X) $\hat{\pi}(X) = \{\bigcap_{i=1..n} a_i\}$
- Считая все подмножества π(X) «открытыми», формируем множества $\underline{\tau}(X) = \{\hat{a} \mid a \subseteq \pi(X)\}$ $\tau(X) = \{\breve{b} \mid b \subseteq \underline{\tau}(X)\}$
- **Теорема 1**. *Система множеств* τ(X) - *топология над* X, *а* π(X) - *предбаза* τ(X) *тогда и только тогда, когда* $\breve{\pi}(X) = X$
  - *Следствие.* $\breve{\pi}(X) = X$ *- необходимое условие выполнения сильной формы аксиомы соответствия.*
  - *Следствие.* $\breve{\pi}(X) = X$ *- необходимое условие разрешимости и регулярности задач распознавания (при выполнении сильной формы аксиомы).*

Выполнимость условия $\breve{\pi}(X) = X$ - критерий оценки набора исходных признаковых описаний

- При $\breve{\pi}(X) \subset X$ ни одна задача распознавания, которая могла бы быть сформулирована для элементов множества X, *не является формализуемой.*
- *Условие* $\breve{\pi}(X) = X$*, наряду с разрешимостью и регулярностью, является важным конструктивным критерием качества постановки задачи.*

BigDATA

Анализ аксиом отделимости в топологическом пространстве T(X)=(X,τ(X))

- T_0 (аксиома Колмогорова)

$$\underset{X}{\forall} x, y \underset{\tau(X)}{\exists} u : (x \in u) \neq (y \in u)$$

- T_1

$$\underset{X}{\forall} x, y \underset{\tau(X)}{\exists} u, v : (x \in u) \wedge (y \in v) \wedge (x \notin v) \wedge (y \notin u)$$

- T_2 (аксиома Хаусдорфа):

$$\underset{X}{\forall} x, y \underset{\tau(X)}{\exists} u, v : (x \in u) \wedge (y \in v) \wedge (u \cap v = \varnothing)$$

- T_3 (аксиома Тихонова):

$$\underset{X}{\forall} x \underset{\tau(X)}{\forall} v \,|\, x \in v \underset{\tau(X)}{\exists} u : (x \in u) \wedge (u \subset v)$$

- T_4 (аксиома нормальности):

$$\underset{\tau(X)}{\forall} u, v : u \cap v = \varnothing \Rightarrow \underset{\tau(X)}{\exists} u_1, v_1 : (u \subset u_1, v \subset v_1) \wedge (u_1 \cap v_1 = \varnothing)$$

Регулярность (по Журавлёву) – достаточное условие разрешимости задачи

- *множество прецедентов* $\mathrm{Q} \subseteq I_i \times I_f,\ \mathrm{Q} \neq \varnothing,\ \mathrm{Q} = \{(\alpha_j, \beta_j)\}$
- *разрешимость задачи – существует* $f : I_i \to I_f$
  $\underset{\mathbf{Q}}{\forall} (\alpha_1, \beta_1), (\alpha_2, \beta_2) : \alpha_1 = \alpha_2 \Rightarrow \beta_1 = \beta_2$
- *регулярность задачи*
  $\underset{Q}{\forall} (\alpha_1, \beta_1), (\alpha_2, \beta_2) : \alpha_1 \neq \alpha_2$
- **Теорема 2.** *Пусть выполнена аксиома соответствия, так что задано* Q=ϕ(X). *Аксиома отделимости* T_1 *выполнена для* τ(X) *тогда и только тогда, когда выполнено условие регулярности множества прецедентов* Q.
  - *Следствие. Регулярность следует из аксиомы* T_2.
  - *Следствие. При выполнении условия регулярности* T(X)=(X,τ(X)) - *дискретное топологическое пространство.*

Решётки над множеством исходных описаний

- Формально, топология τ(X) - неупорядоченное множество. В то же время, элементы τ(X) могут являться подмножествами друг друга, так что между ними возникает естественное упорядочение по включению
- **Теорема 3.** *Решётка* L(X) *над* L(X) *является булевой при выполнении аксиомы соответствия и условия регулярности для* X*.*
  - *Следствие*. L(X) *– дистрибутивна*

Метрическое пространство M(τ(X))

- *Оценка на решётке* L(X) – функция $v : L(X) \to R^{+}$, для которой выполнено условие оценки $\forall_{L(X)} a, b : v[a] + v[b] = v[a \wedge b] + v[a \vee b]$
- *Изотонная оценка: v- монотонна*
- **Лемма 5 (Биркгофа)**. *Пусть* L - *дистрибутивная решётка,* v - *оценка на* L, *а* $d(x, y) = v[x \vee y] - v[x \wedge y]$. *Тогда, для произвольных выполнено* $d(a \vee x, a \vee y) + d(a \wedge x, a \wedge y) = d(x, y)$
- **Теорема 6**. *В решётке L(X) с изотонной оценкой v[] функция* $\rho(x, y) = v[x \vee y] - v[x \wedge y]$ *- метрика, образующая метрическое пространство* $M(\tau(X)) = (\tau(X), \rho(x, y))$
  - *Следствие. При* $v[x] = h[x]$ $\rho(x, y) = |x \Delta y|$
  - *Следствие. В конечной L(X) с заданной изотонной оценкой v[] определено* $\varepsilon_{min} > 0$ *- минимальное, не равное нулю значение расстояния ρ(x,y) между элементами (x,y).*
  - *Следствие. Произвольное подпространство* $\tilde{m} \subseteq M(\tau(X))$ *также является метрическим пространством.*

Выводы

- Разработан концептуальный аппарат для анализа математической структуры исследуемых наборов *разнородных признаковых описаний* в произвольной задаче распознавания или классификации.

- Регулярность (по Журавлеву) множества объектов позволяет переносить результаты анализа метрического пространства на элементы множества X, т.е. в изучаемую проблемную область.

Выводы

- Условие $\breve{\pi}(\mathrm{X}) = \mathrm{X}$ - конструктивный критерием качества постановки задачи, допускающий комбинаторное тестирование

- Разрабатываются приложения аппарата к
  - введению метрик на множествах признаков
  - анализу «значимости взаимодействий» между разнородными признаковыми описаниями.
  - ведению метрик на множествах объектов