

# ВАРИАЦИОННЫЕ МЕТОДЫ КЛАССИФИКАЦИОННОГО АНАЛИЗ ДАННЫХ

Бауман Е.В.(ВВАТ,ИПУ),

Дорофеюк А.А.(ИПУ)

# Задачи структурного (классификационного) анализа данных

## 1. Классификация.

Разбить множество объектов на группы схожих.

## 2. Группировка параметров.

Набор параметров, описывающих систему, необходимо разбить на группы связанных и выделить из каждой группы наиболее существенный параметр.

### 3. Кусочная аппроксимация.

Требуется так разбить пространство входных параметров, чтобы сложная во всем пространстве зависимость выходного параметра от вектора входных была простой в пределах каждой области.

# Постановка задачи.

- 1). Классифицируемое множество объектов.
- 2). Класс допустимых классификаций.
- 3). Критерий качества классификации.

## 1). Классифицируемое множество объектов:

произвольное множество  $X$  с законом распределения  $P(A), A \subseteq X$ .

## 2). Класс допустимых классификаций.

Размытой классификацией множества  $X$  называется вектор-функция  $H(x) = (h_1(x), \dots, h_r(x))$  ( $h_i(x)$  - функция принадлежности к  $i$ -му классу) такая, для любого  $x \in X$  значение  $H(x)$  принадлежит некоторому множеству  $V$  т.е.  $H(x) \in V \subseteq R^n$ .

Класс допустимых классификаций:  $\Xi(V)$

### 3). Критерий качества классификации.

За критерий качества принимается произвольный выпуклый

функционал  $\Phi(H)$ , определенный на  $L_2(X, P)^r$

## Задача построения размытой классификации

$$\Phi(H) \rightarrow \max_{H \in \Xi(V)}$$

# Виды функционалов

## 1. Классификация евклидова пространства

$X = R^m$  с заданным законом распределения  $P(x)$

$$\Phi_1(H) = - \sum_{i=1}^r \int_X (x - \alpha_i)^2 h_i(x) dP(x),$$

где  $\alpha_i = \frac{\int_X x h_i(x) dP(x)}{\int_X h_i(x) dP(x)}$  - среднее  $i$ -го класса.

## 2. Экстремальная группировка параметров

$$X = \{x^{(1)}, \dots, x^{(m)}\} -$$

поведение  $n$  объектов.  $\rho(x^{(j)}, f^{(i)}) = \frac{r}{m}$  - набор параметров, описывающих

$$\Phi_2(H) = \sum_{i=1}^r \sum_{j=1}^m |\rho(x^{(j)}, f^{(i)})|^t h_i(x),$$

где  $\rho(x^{(j)}, f^{(i)})$  -

параметром  $x^{(j)}$  коэффициент корреляции между

$$f^{(i)} = \operatorname{argmax}_f \sum_{j=1}^m |\rho(x^{(j)}, f)|^t h_i(x).$$

и фактором  $f^{(i)}$  равным



### 3. Кусочно-линейная аппроксимация

$X = R^m$  - пространство входных параметров

с заданным законом распределения  $P(x)$

$y = y(x)$  - выходной параметр.

Для каждого  $i$ -го класса классификации  $H$  с помощью линейной регрессии строится линейная функция

$$(c_i, x) + d_i.$$

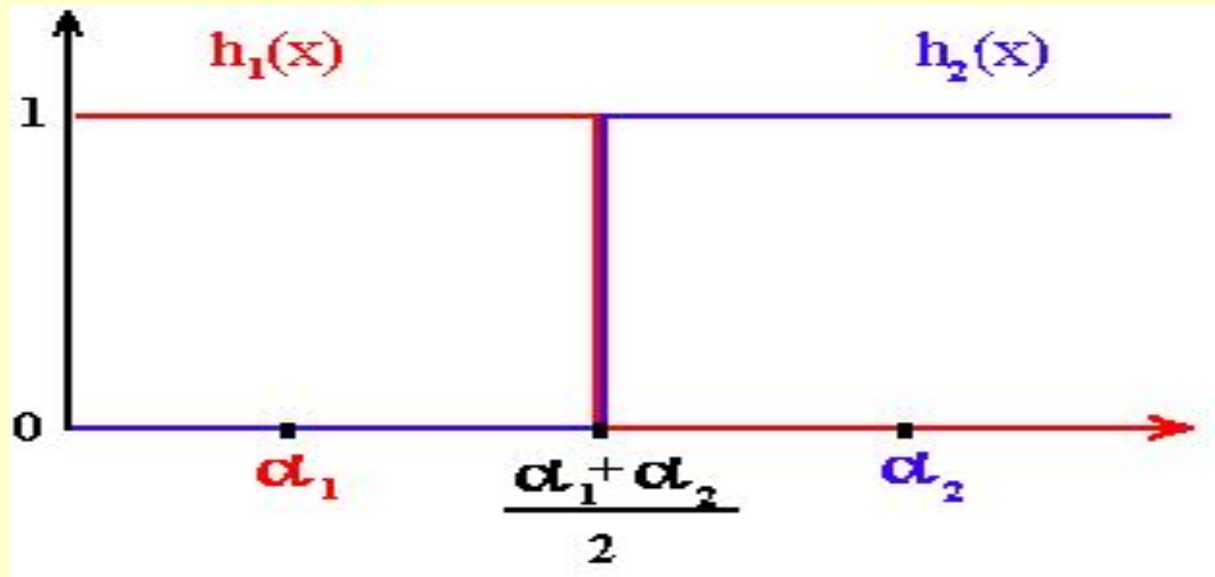
Функционал качества аппроксимации:

$$\Phi_3(H) = - \sum_{i=1}^r \int_X (y(x) - [(c_i, x) + d_i])^2 h_i(x) dP(x).$$

# Виды размытости классификации

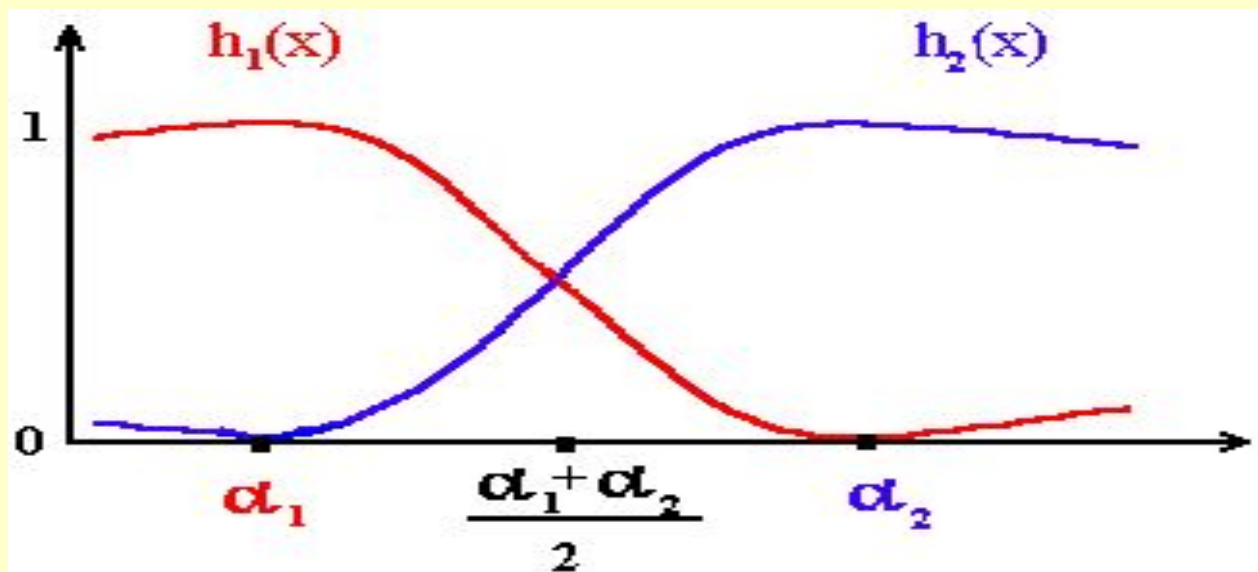
## 1. Четкая классификация

$$V_1 : h_i(x) \in \{0; 1\}, \sum_{i=1}^r h_i(x).$$



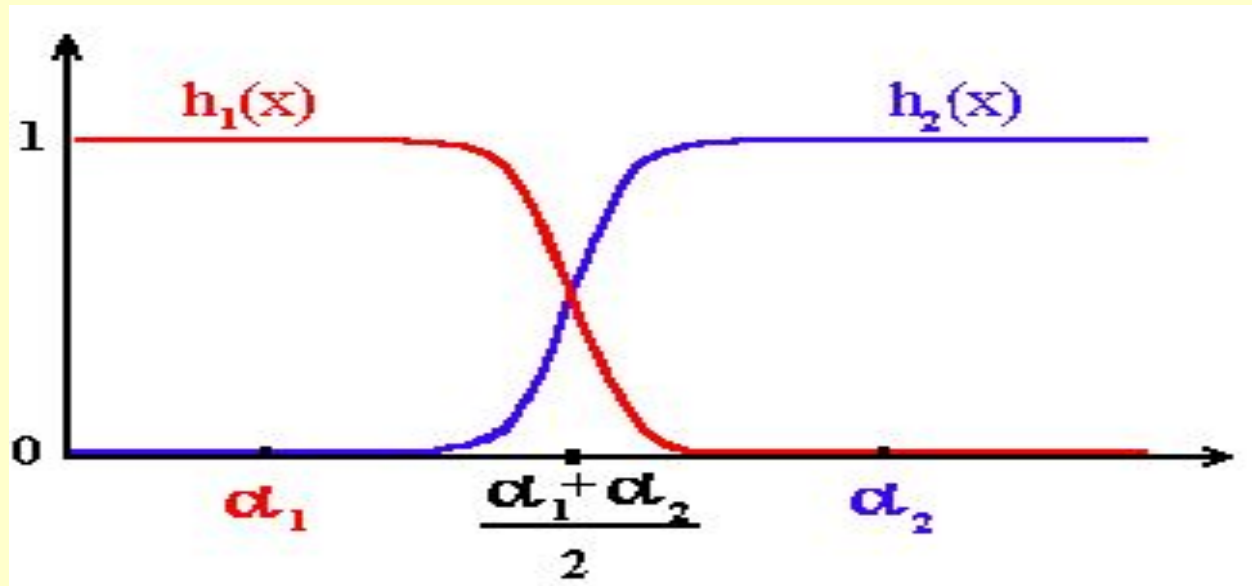
## 2. Размытая классификация по Беждеку

$$V_2 : h_i(x) \in [0; 1], \sum_{i=1}^r h_i^\lambda(x), (0 < \lambda < 1).$$



### 3. Классификация с размытыми границами

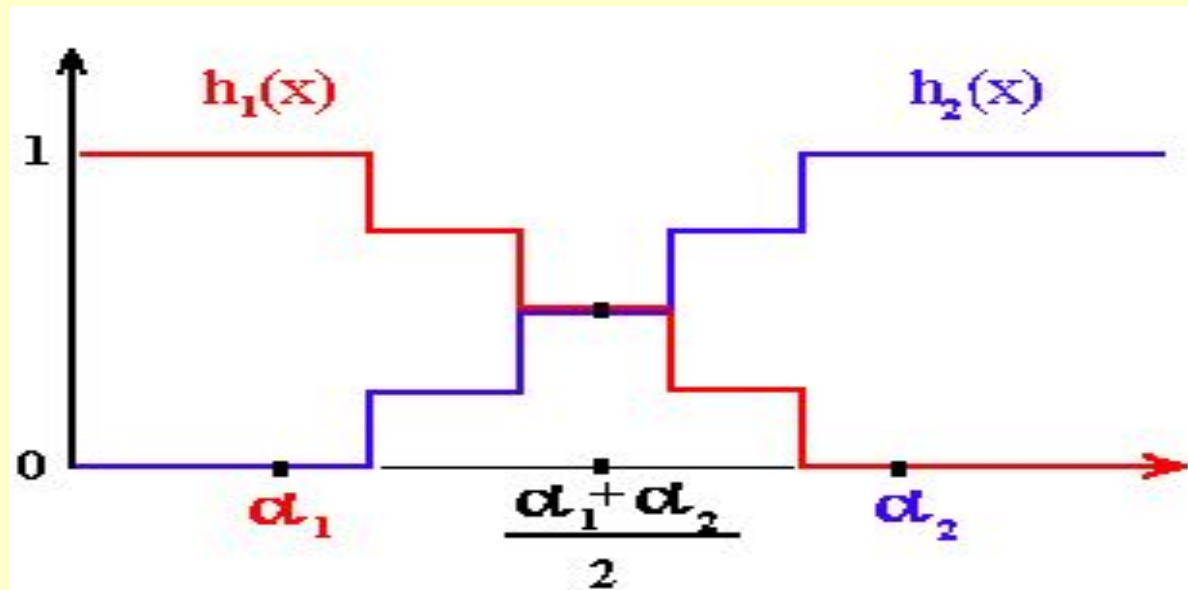
$$V_3 : h_i(x) \in [0; 1], \sum_{i=1}^r (a - h_i(x))^2 = (r - 1)a^2 + (a - 1)^2$$



## 4. Качественная размытая классификация

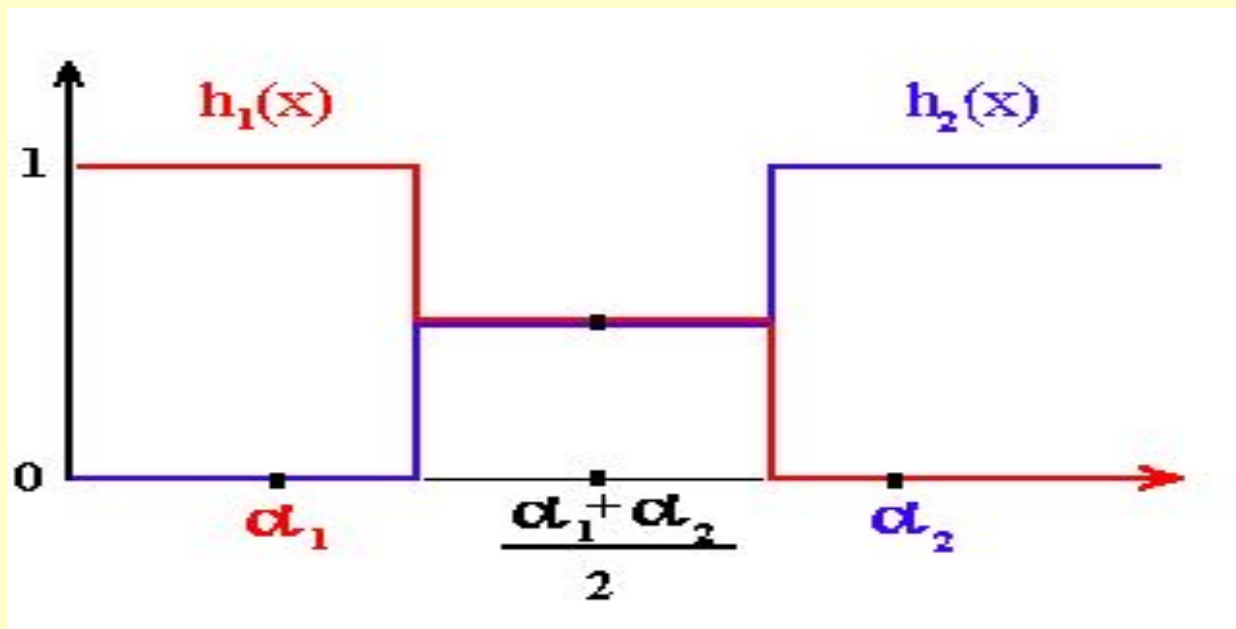
$$V_4 : h_i(x) \in [0; \sqrt[\lambda]{\frac{1}{k}}; \sqrt[\lambda]{\frac{2}{k}}; \dots; \sqrt[\lambda]{\frac{k-1}{k}}; 1],$$

$$\sum_{i=1}^r h_i^\lambda(x), \quad (0 < \lambda < 1).$$



## 5. Классификация с перекрывающимися классами

$$V_5 = \left\{ H = (h_1, \dots, h_r) : \forall (i_1, \dots, i_k) \quad h_j = \begin{cases} \lambda \sqrt{\frac{1}{k}}, & j \in (i_1, \dots, i_k) \\ 0, & j \notin (i_1, \dots, i_k) \end{cases} \right\}$$



# Вид оптимальной классификации

$H_F$  - опорная к  $F(x) = (f_1(x), \dots, f_r(x))$ , если

$$\forall x \quad H_F(x) = \operatorname{argmax}_{(h_1, \dots, h_r) \in V} \sum_{i=1}^r h_i f_i(x)$$

**Теорема 1.** Пусть  $\Phi(H^*) = \max \Phi(H)$   
 $F^* = \operatorname{grad} \Phi(H^*)$ . Тогда  $\Phi(H_{F^*}) = \Phi^{\text{И}}(H^*)$

# Алгоритм классификации при известном законе распределения (конечная выборка объектов)

$$H_0 \rightarrow \dots$$

$$\dots \rightarrow H_n \rightarrow F_n = \text{grad } \Phi(H_n) \rightarrow \\ \rightarrow H_{n+1} = H_{F_n} \rightarrow \dots$$

**Теорема 2.**  $\Phi$  - выпуклый, ограниченный  $\Rightarrow$   
в силу алгоритма  $\{H_n\}$   
стационарной точке функционала слабо сходится к



# Критерий качества классификации, зависящий от моментов классов

Пусть  $z(x)$  — отображение множества  $X$  в  $Z = R^k$ .

$Z$  — спрямляющее пространство.

$$p_i = \int_X h_i(x) dP(x), \quad M_i = \int_X z(x) h_i(x) dP(x)$$

$$\mu(H) = (p_1, M_1, \dots, p_r, M_r).$$

$$\Phi_4(H) = \phi(\mu(H)) \quad (1)$$

$\phi$  — непрерывная функция  $r(k+1)$ -мерного вектора.

# Вид оптимальной классификации функционала (1)

$$\pi = (d_1, c_1, \dots, d_r, c_r), d_i \in R^1, c_i \in R^k$$

$H_\pi$  - линейная с вектором  $\pi$ ,

$$H_\pi(x) = \underset{(h_1, \dots, h_r) \in V}{\operatorname{argmax}} \sum_{i=1}^r h_i [(c_i, z(x)) + d_i] \quad \text{если}$$

**Теорема 1.** Пусть  $\phi(\mu(H^*)) = \max \phi(\mu(H))$  и  
 $\pi^* = \operatorname{grad} \phi(\mu^*), (\mu^* = \mu(H^*))$ . Тогда  
 $\phi(\mu(H_{\pi^*})) = \phi(\mu(H^*))$ .

# Классификация по бесконечной выборке объектов

$$S = \{x_1, \dots, x_n, \dots\} \quad \text{по } P(A).$$

**Задача.** По  $S$  выборка  $\phi(\mu(H))$ .

максимизировать

Ограничения на закон распределения:

$$1). \exists A: P(|z(x)| > A) = 0,$$

$$2). \forall c, d \quad P\{(c, z(x)) + d = 0\} = 0.$$

# Алгоритм

$$\begin{aligned}
 v_i^n &= \left(1 - \frac{1}{n}\right)v_i^{n-1} + \frac{1}{n}h_i^{n-1}(x_n), \\
 m_i^n &= \left(1 - \frac{1}{n}\right)m_i^{n-1} + \frac{1}{n}z(x_n)h_i^{n-1}(x_n), \\
 \psi^n &= (v_1^n, m_1^n, \dots, v_r^n, m_r^n), \\
 \pi^n &= \text{grad}(\phi(\psi^n)), \\
 H_n &= H_{\pi^n}.
 \end{aligned}$$

Здесь  $v_i^n$  - оценка  $p_i^n$ ,  $m_i^n$  - оценка  $M_i^n$ ,  
 $\psi^n$  - оценка  $\mu(H^n)$ .  
 оценка

# Сходимость алгоритма

**Теорема 3.**  $\phi$  -

И-

двумя непрерывно дифференцируема и сильно выпукла  $\Rightarrow$

$$\lim_{n \rightarrow \infty} \phi(\mu(H_n)) \geq \lim_{n \rightarrow \infty} \phi(\psi_n) = C(S)$$

$C(S)$  - минимальное значение функции  $\phi$  ;

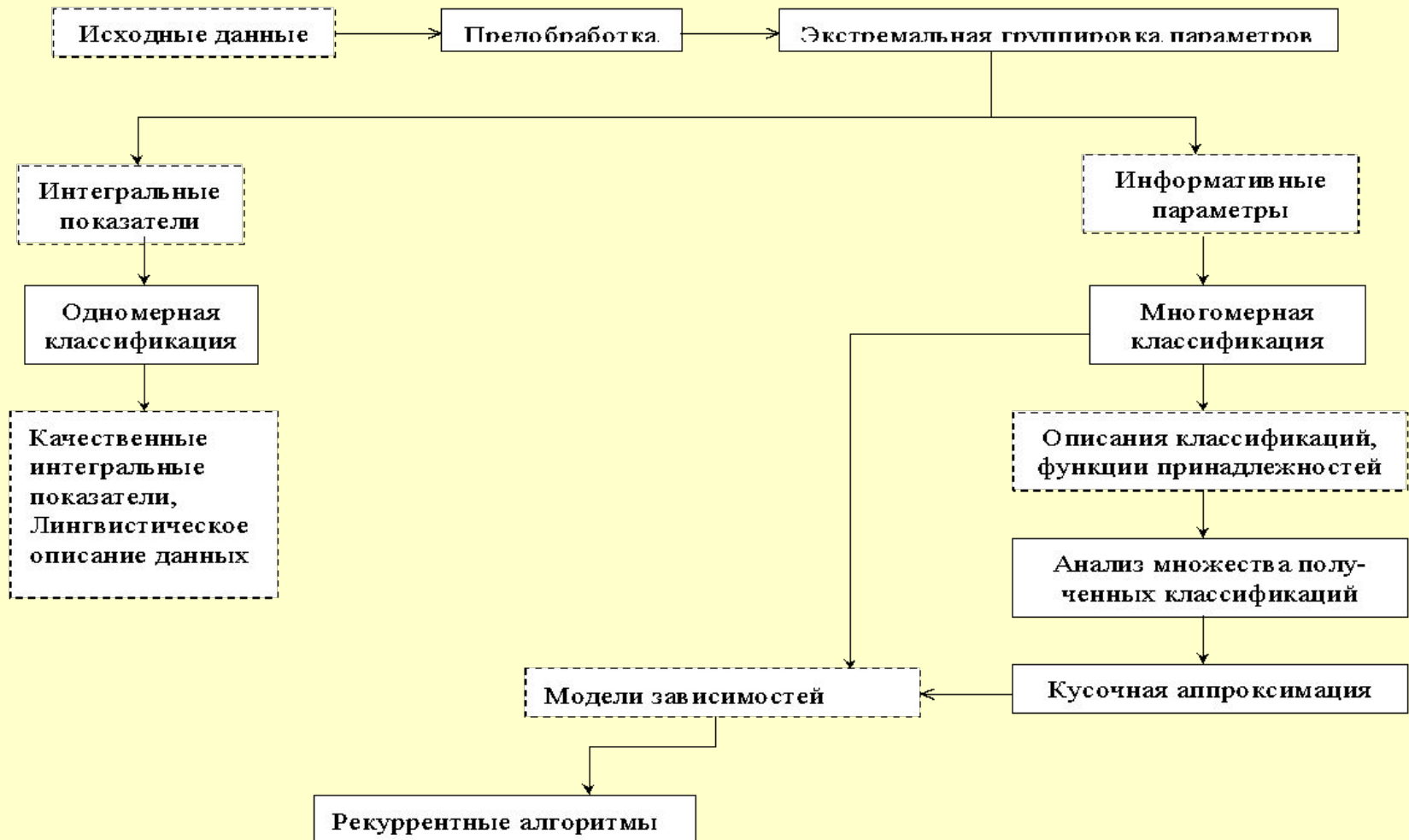
если при этом число классов равно 2, то

$$\lim_{n \rightarrow \infty} (\mu(H_n) - \psi^n)^2 = 0$$

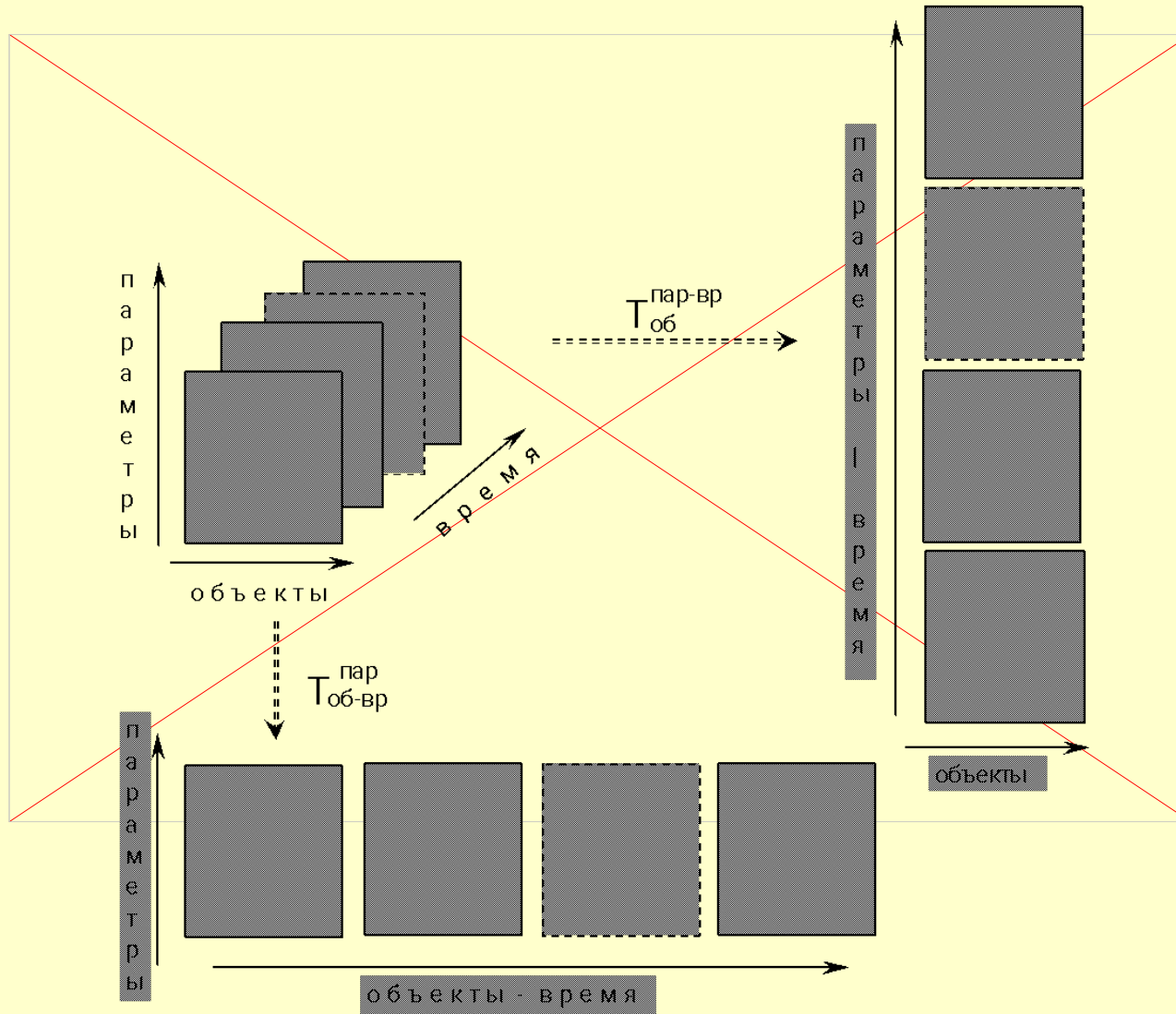
# Система анализа данных «АНАЛИТИК»

- Вид обрабатываемых данных. Куб данных - таблица «объекты-параметры», развернутая во времени.
- Основные модули: предобработки, экстремальной группировки параметров, классификации объектов, анализа множества полученных классификаций, кусочной аппроксимации и рекуррентных алгоритмов.
- Выдача результатов: в том числе на карту.

# Схема обработки данных в системе «АНАЛИТИК»



# Развертка куба данных





# Предобработка

- Выбор текущего подкуба данных
- Создание производных показателей
- Описательная статистика
- Выявление выбросов в данных
- Заполнение пропусков в данных
- Нормирование данных

# Группировка параметров

$T_{об-вр}^{пар}$  - выявление структуры набора параметров вне зависимости от времени.

$T_{об}^{пар-вр}$  учитывает сдвиг времени параметров друг относительно друга.

Результаты: группы параметров + интегральные показатели (информативные для классификации).

# Классификация объектов

$T_{\text{об-вр}}^{\text{пар}}$  - выявление режимов работы объектов, не зависящих от времени. В результате один объект в разные моменты времени может попасть в разные классы.

$T_{\text{об}}^{\text{пар-вр}}$  - в один класс попадают объекты, с одинаковой динамикой изменения показателей работы.

Результаты: функции принадлежности объектов к классам + центры классов.

# Кусочная аппроксимация

Используется только  $T_{\text{об-вр}}^{\text{пар}}$ .

Начальное разбиение входов – результаты классификации.

Результаты: функции принадлежности объектов к классам + регрессионные модели зависимостей внутри классов.

# Анализ множества полученных классификаций

За счет большого числа свободных параметров алгоритмов получается много результирующих классификаций.

С помощью классификационных методов можно структурировать это множество.

Результат: набор информативных классификаций

# Рекуррентные алгоритмы

Если данные об исследуемой системе поступают последовательно во времени (например, статистические данные о деятельности предприятий), то используются рекуррентные алгоритмы классификации и кусочной аппроксимации, позволяющие корректировать решающие правила и локальные модели в соответствии с новой информацией.