ВАРИАЦИОННЫЕ МЕТОДЫ КЛАССИФИКАЦИОННОГО АНАЛИЗ ДАННЫХ

Бауман Е.В.(ВАВТ,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 на r классов называется вектор-функция $H(x) = (h_1(x), \ldots, h_r(x))$ ($h_i(x)$ - функция принадлежности к i-му классу) такая, что для любого x значение $H(x)$ принадлежит некоторому множеству V т.е. $H(x) \in V \subset R^r$.

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

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

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

$X = \{x^{(1)}, \ldots, x^{(m)}\}$ - набор параметров, описывающих поведение n объектов. $P(x^{(j)}) = \frac{1}{m}$.

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

где $\rho\left(x^{(j)}, f^{(i)}\right)$ - коэффициент корреляции между параметром $x^{(j)}$ и фактором $f^{(i)}$, равным

$$f^{(i)} = \arg\max_{f} \sum_{j=1}^{m} \left| \rho\left(x^{(j)}, f\right) \right|^t h_i(x).$$

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 $(c_i, x) + d_i$.

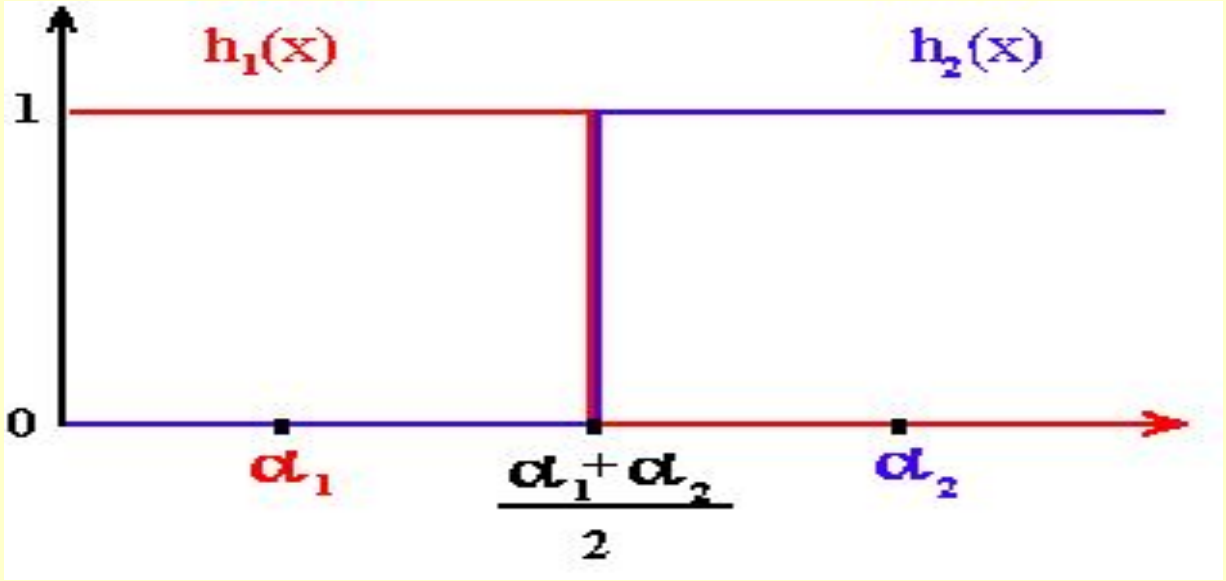
Функционал качества аппроксимации:

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

Виды размытости классификации

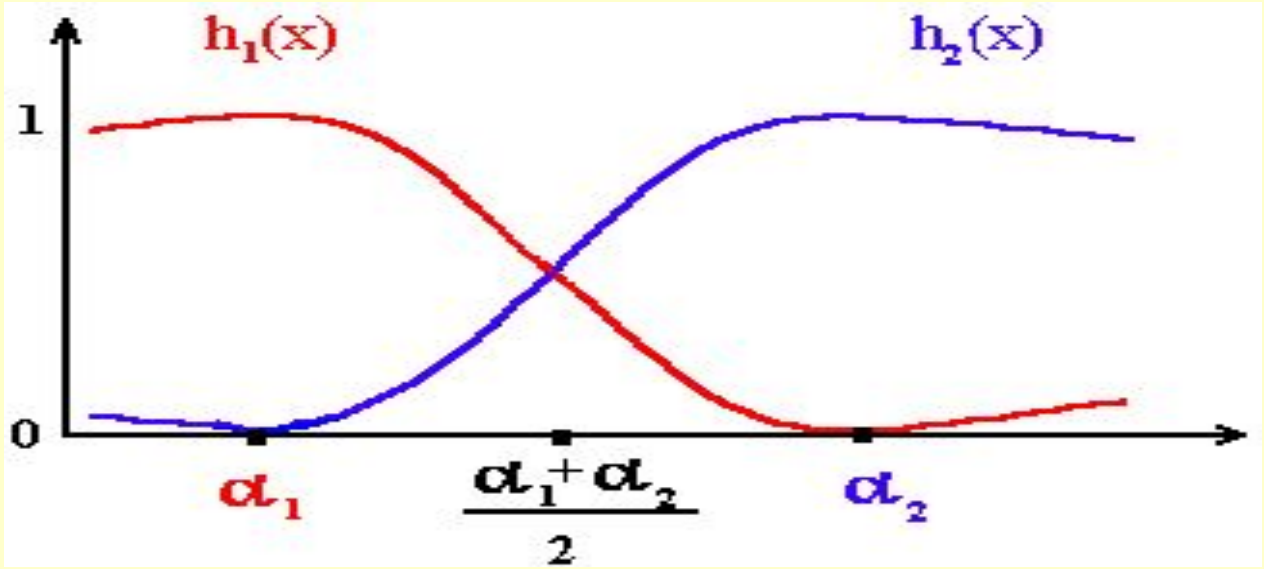
1. Четкая классификация

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



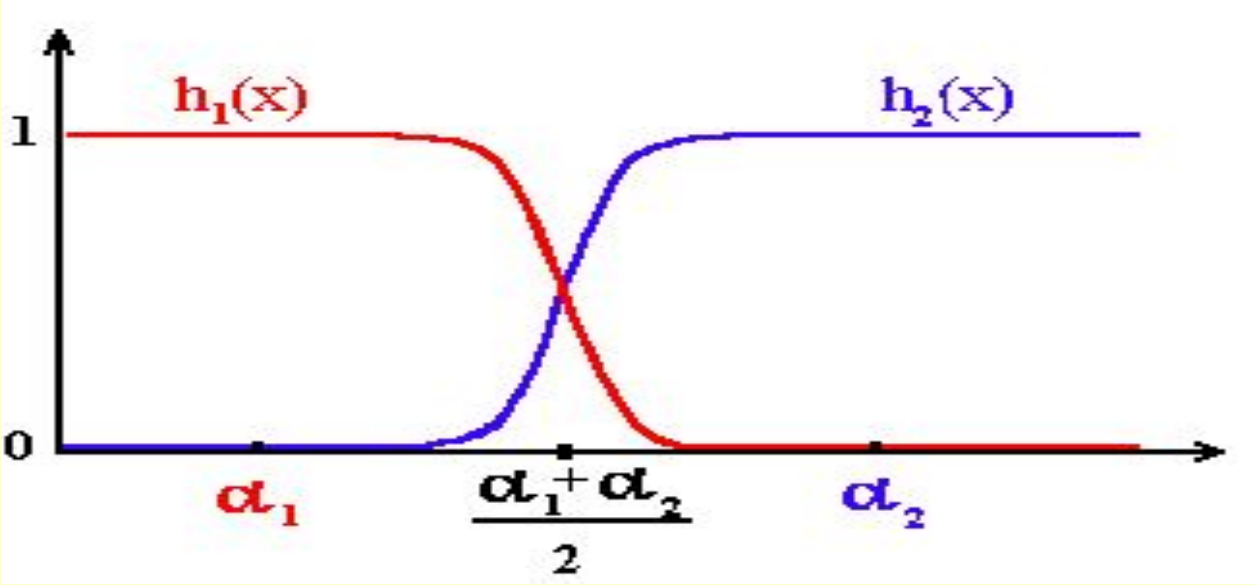
2. Размытая классификация по Беждеку

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



3. Классификация с размытыми границами

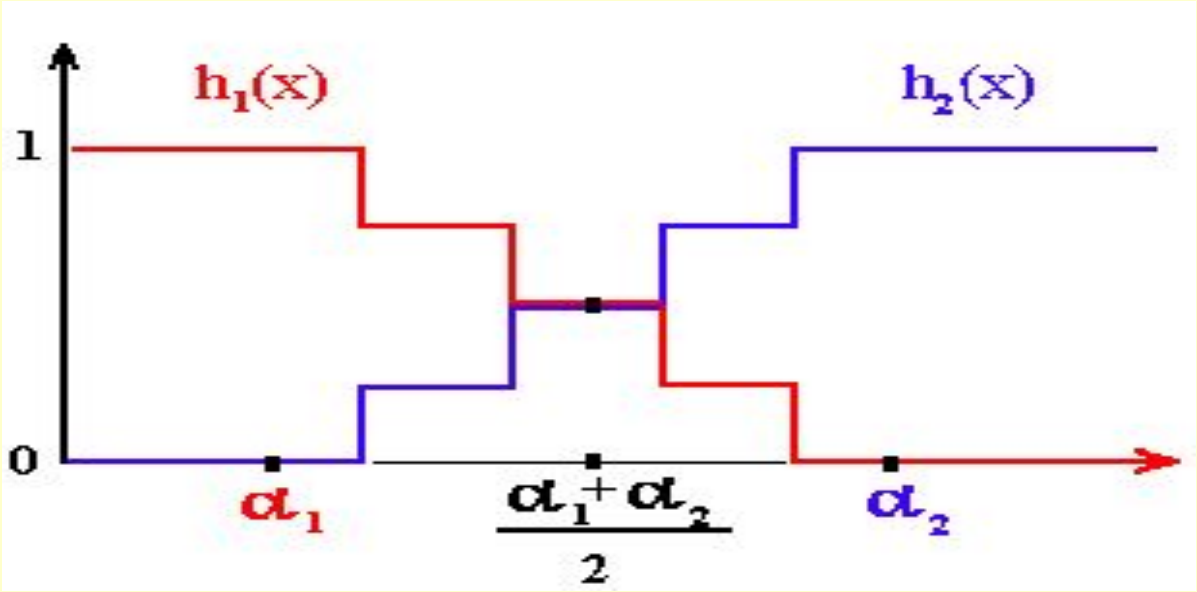
$$V_3 : h_i(x) \in [0;1], \sum_{i=1}^{r} (a - h_i(x))^2 = (r-1)a^2 + (a-1)^2$$



4. Качественная размытая классификация

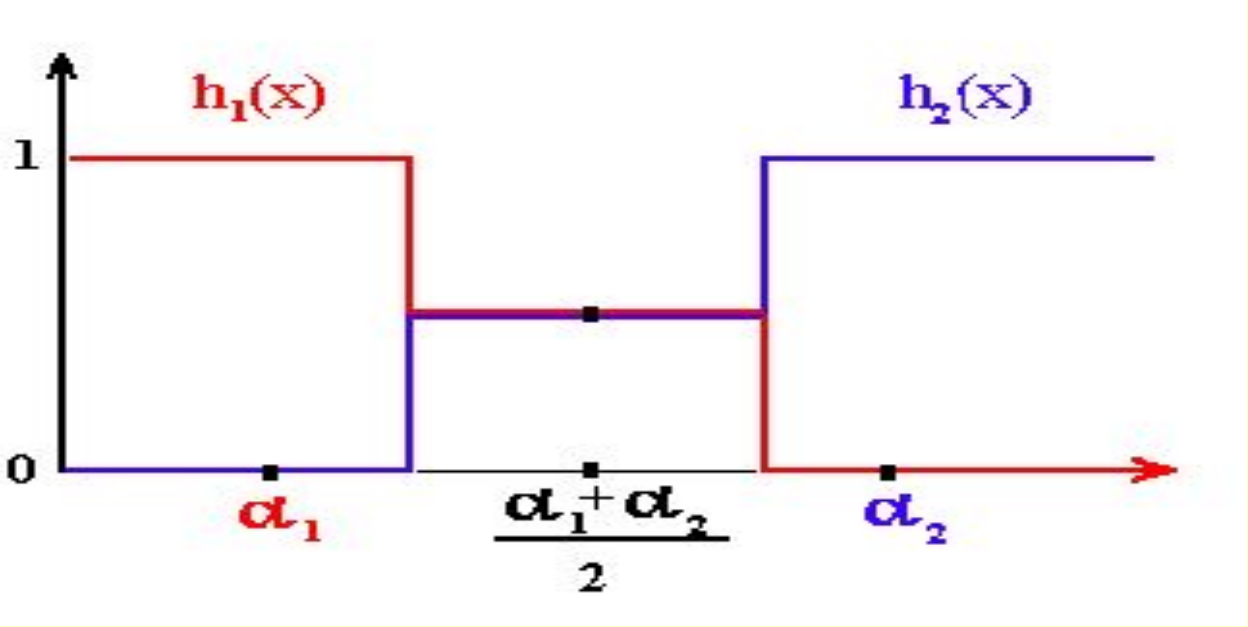
$$V_4 : h_i(x) \in \left[0; \sqrt[\lambda]{\frac{1}{k}}; \sqrt[\lambda]{\frac{2}{k}}; \ldots; \sqrt[\lambda]{\frac{k-1}{k}}; 1\right],$$

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



5. Классификация с перекрывающимися классами

$$V_5 = \left\{ H = (h_1,\ldots,h_r) : \forall (i_1,\ldots,i_k) \quad h_i = \begin{cases} \sqrt[\lambda]{\frac{1}{k}}, i \in (i_1,\ldots,i_k) \\ 0, i \notin (i_1,\ldots,i_k) \end{cases} \right\}$$



Вид оптимальной классификации

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

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

Теорема 1. Пусть $\Phi(H^*) = \max \Phi(H)$

$F^* = grad\,\Phi(H^*)$. Тогда $\Phi(H_{F^*}) = \Phi^{И}(H^*)$

Алгоритм классификации при известном законе распределения (конечная выборка объектов)

$$H_0 \to \dots$$

$$\dots \to H_n \to F_n = grad\,\Phi(H_n) \to$$

$$\to H_{n+1} = H_{F_n} \to \dots$$

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

Критерий качества классификации, зависящий от моментов классов

Пусть $z(x)$ X в $Z = R^k$.

Z - спрямляющее пространство.

отображение множества

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

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

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

ϕ - r(k+1)-мерного вектора.

выпуклая функция

Вид оптимальной классификации функционала (1)

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

H_π - линейная с вектором π,

если

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

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

Классификация по бесконечной выборке объектов

$S = \{x_1, \ldots, x_n, \ldots\}$ по $P(A)$.

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

Ограничения на закон распределения:

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

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

Алгоритм

$$\nu_i^n = (1-\tfrac{1}{n})\nu_i^{n-1} + \tfrac{1}{n} h_i^{n-1}(x_n),$$
$$m_i^n = (1-\tfrac{1}{n})m_i^{n-1} + \tfrac{1}{n} z(x_n) h_i^{n-1}(x_n),$$
$$\psi^n = \left(\nu_1^n, m_1^n, \ldots, \nu_r^n, m_r^n\right),$$
$$\pi^n = grad\left(\phi\left(\psi^n\right)\right),$$
$$H_n = H_{\pi^n}.$$

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

Сходимость алгоритма

Теорема 3. ϕ - дважды непрерывно дифференцируема и сильно выпукла $\Gamma \to \Gamma$ и-

$$\underline{\lim_{n\to\infty}}\,\phi(\mu(H_n)) \overset{п.н.}{\geq} \lim_{n\to\infty}\phi(\psi_n) \overset{п.н.}{=} C(S)$$

$C(S)$ - онарное значение функции ϕ ;

стации

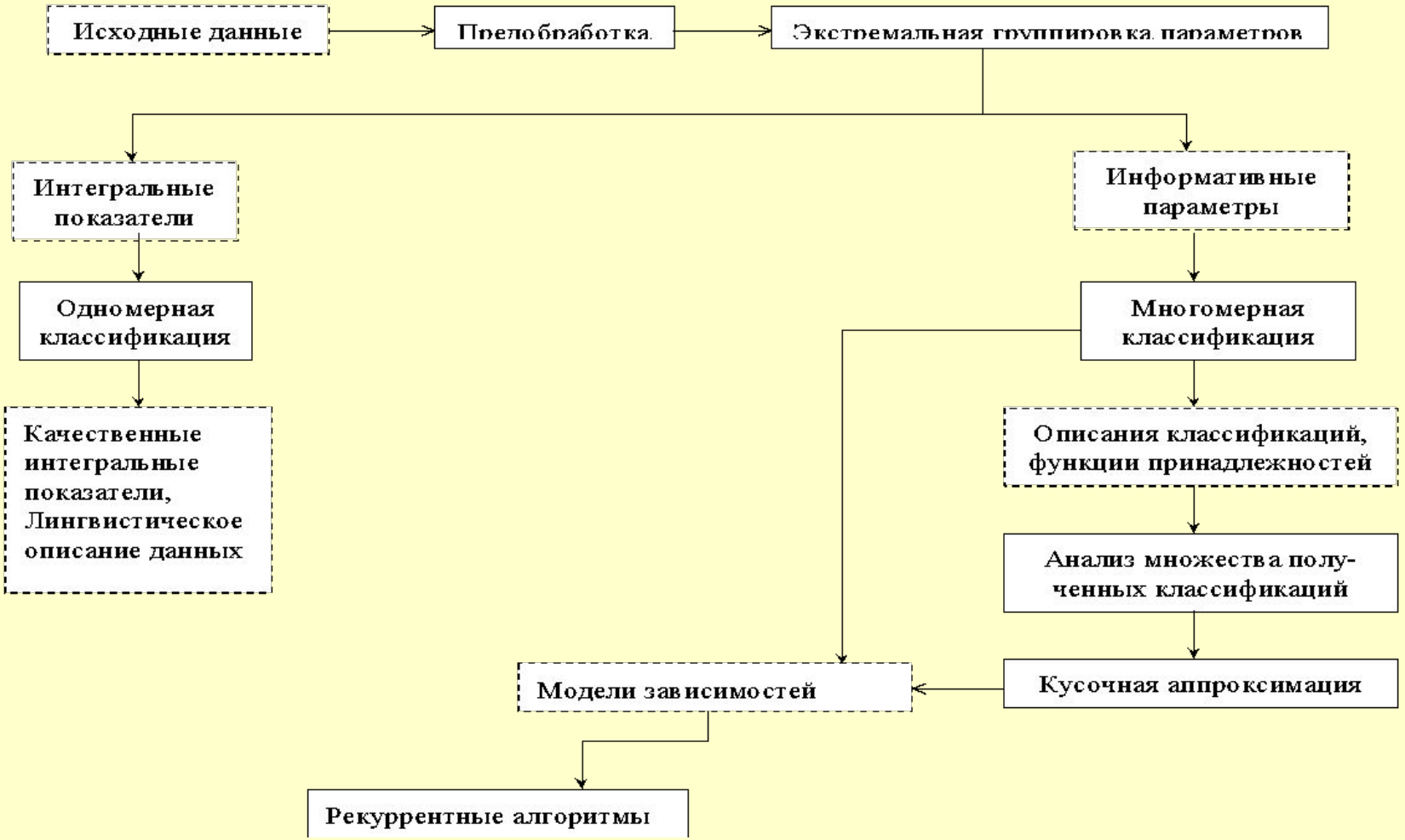
если при этом число классов равно 2, то

$$\lim_{n\to\infty}\left(\mu(H_n)-\psi^n\right)^2 \overset{п.н}{=} 0$$

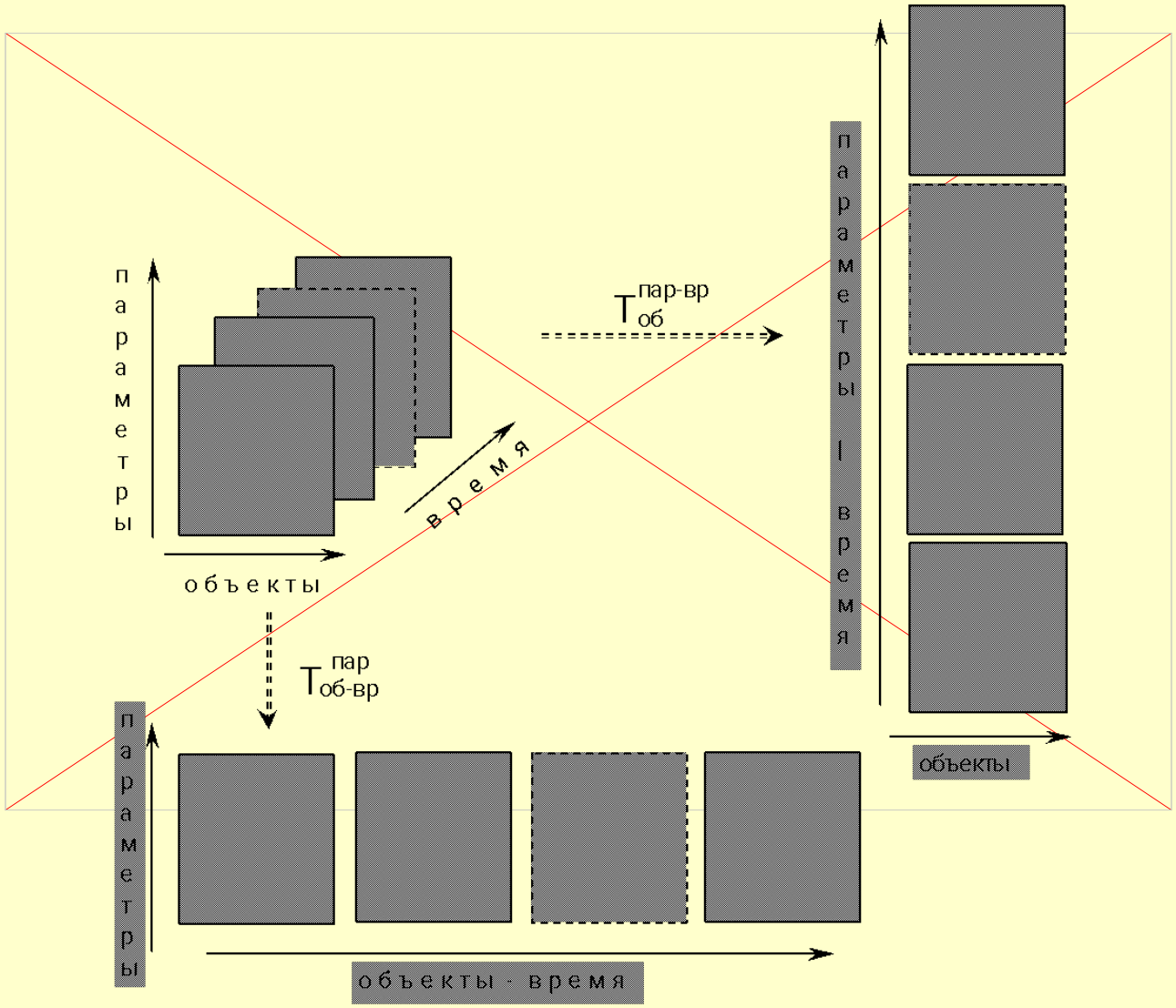
Система анализа данных «АНАЛИТИК»

- 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

$Т^{пар}_{об\text{-}вр}$- выявление структуры набора параметров вне зависимости от времени.

$Т^{пар\text{-}вр}_{об}$ учитывает сдвиг времени параметров друг относительно друга.

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

Классификация объектов

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

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

Результаты: функции принадлежностей объектов к классам + центры классов.

Кусочная аппроксимация

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

Начальное разбиение входов – результаты классификации.

Результаты: функции принадлежностей 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.