О критериях ветвления, используемых при синтезе решающих деревьев

Генрихов И. Е.

Задача распознавания по прецедентам

$M = \bigcup_{i=1}^{l} K_i$ – множество объектов

$K_1, \dots, K_l$ – непересекающиеся классы

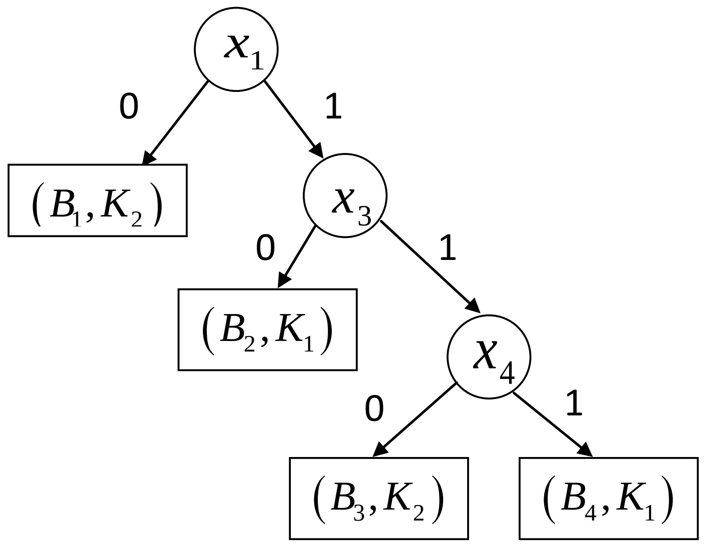
$\{x_1, \dots, x_n\}$ – система признаков

$\{S_1, \dots, S_m\}$ – множество обучающих объектов, где $S_j \in M$, $j = 1, \dots, m$

$S \in M$, S – распознаваемый объект

Известным инструментом для решения задачи распознавания по прецедентам являются **решающие деревья (РД)**.

Пример. Бинарное решающее дерево с бинарными признаками



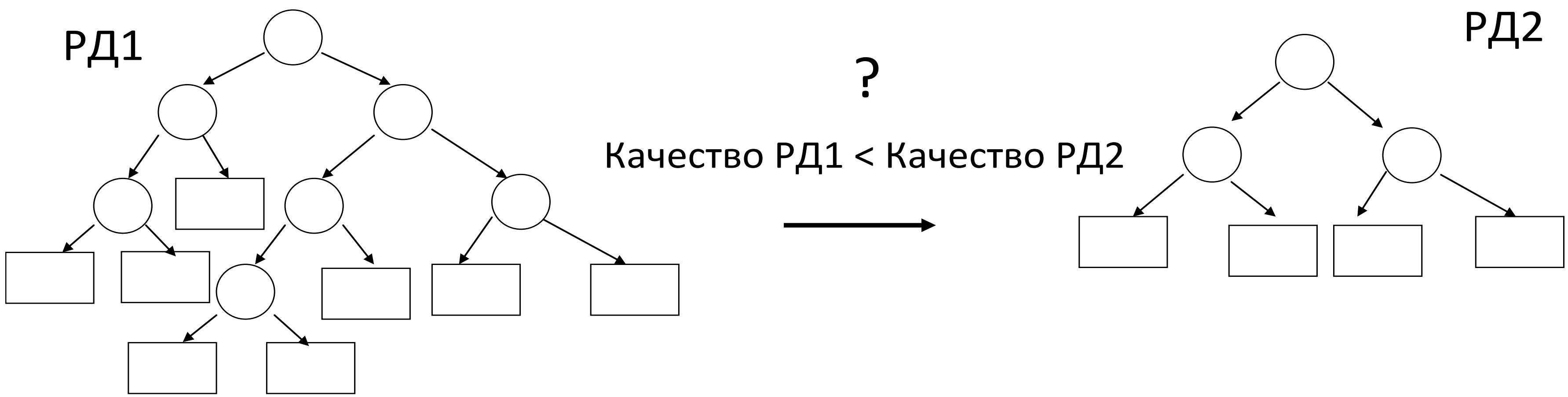
$B_1 = \overline{x_1}$, $B_2 = x_1 \wedge \overline{x_3}$, $B_3 = x_1 \wedge x_3 \wedge \overline{x_4}$, $B_4 = x_1 \wedge x_3 \wedge x_4$ - э.к.

По построению распознаваемый объект может попасть только в один лист бинарного РД (БРД).

Распознаваемый объект $S = (1,1,1,1,1)$ попадает в лист (B_4, K_1), так как S принадлежит интервалу истинности N_{B_4} конъюнкции B_4. Следовательно, $S \in K_1$.

Основная проблема синтеза решающего дерева

Проблема: построение РД наиболее простого по структуре с хорошими распознающими качествами.



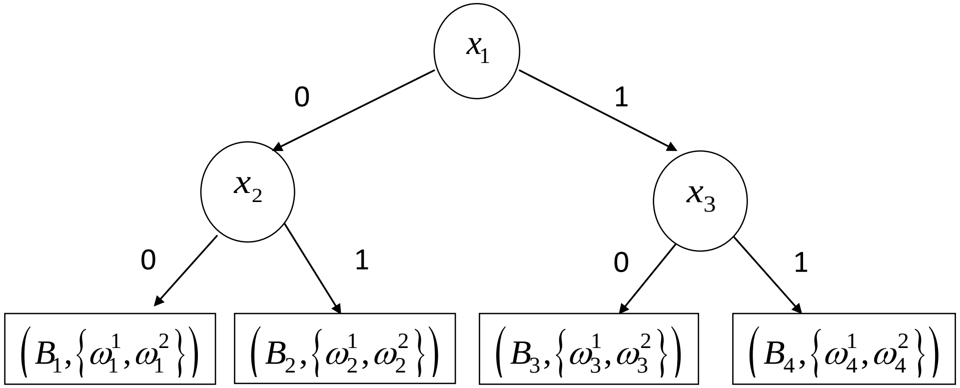
Основной способ решения проблемы: поиск баланса между «сложностью» РД и качеством с помощью методов редукции.

Наибольшее влияние на структуру и качество РД оказывает применяемый **критерий выбора признака** для построения очередной внутренней вершины дерева.

Основные задачи

1. Исследование особенностей разделения обучающих объектов при синтезе дерева в зависимости от применяемого критерия ветвления (Gain, GainRatio, Gini Index, Twoing и критерия равномерного разбиения (Dcrit)).
2. Исследование структурных и распознающих свойств РД в зависимости от применяемого критерия ветвления, а именно:
   - глубина РД;
   - число листьев РД;
   - средняя глубина листьев дерева;
   - «сбалансированность» дерева;
   - взвешенная глубина распределения описаний обучающих объектов по листья дерева;
   - «оптимальность» распределения обучающих объектов по листьям дерева;
   - оценка качества дерева с помощью метода LOO и анализа распределения отступов обучающих объектов.
3. Разработка критерия ветвления, позволяющего строить более «оптимальное» РД.

Пример бинарного РД с бинарными признаками и с двумя классами



$x_1,\ x_2,\ x_3$ – обычные вершины дерева

$\left\{\omega_i^1,\omega_i^2\right\}$ – вектор оценок за классы в i-ом листе

$B_1=\overline{x_1}\wedge\overline{x_2},\ B_2=\overline{x_1}\wedge x_2,\ B_3=x_1\wedge\overline{x_3},\ B_4=x_1\wedge x_3$

ω_i^j – оценка i-ого листа за принадлежность классу K_j

<u>Опр.</u> Лист $\left(B_v,\left\{\omega_v^1,\ldots,\omega_v^l\right\}\right)$ называется **голосующим** за объект S, если $S\in N_{B_v}$.

Если лист $\left(B_v,\left\{\omega_v^1,\ldots,\omega_v^l\right\}\right)$ – голосующий за S, то S получает оценку ω_v^i за класс K_i и

$$S\in K_i,\ \text{если}\ \ \omega_v^i=\max_{j\in\{1,\ldots,l\}}\omega_v^j,\ i\in\{1,\ldots,l\}.$$

Если классов с максимальной оценкой несколько, то среди них выбирается класс с максимальным числом обучающих объектов. Иначе происходит отказ.

Вычисление оценки ω_v^i в векторе оценок $\{\omega_v^1,...,\omega_v^l\}$

<u>Опр.</u> Обучающий объект, описание которого попадает в лист дерева $(B_v,\{\omega_v^1,...,\omega_v^l\})$, называется правильным для листа $(B_v,\{\omega_v^1,...,\omega_v^l\})$.

Пусть m_v^i – число правильных объектов класса K_i в листе $(B_v,\{\omega_v^1,...,\omega_v^l\})$, $m_v^* = \max\limits_{i=1,...,l} m_v^i$, $m_v = \sum\limits_{i=1}^{l} m_v^i$, где l – число классов.

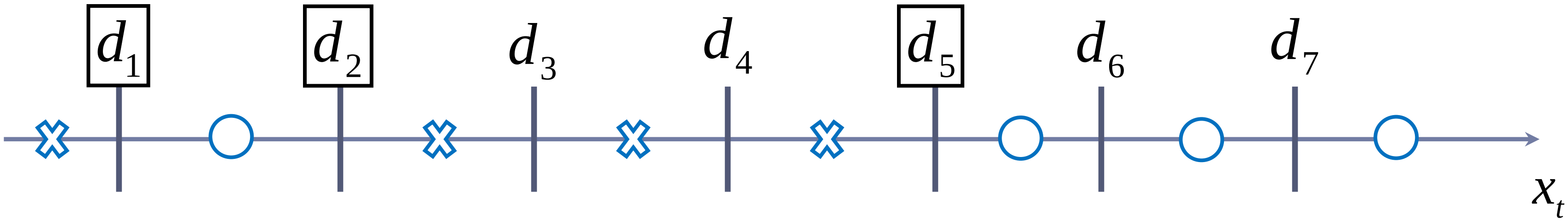
Оценка $\omega_v^i = (m_v^i + 1)\big/(m^i + l)$, $i = 1,...,l$, где m^i – число обучающих объектов класса K_i.

Выбор порога бинарной перекодировки в случае вещественнозначной информации

$\{c_1,\ldots,c_u\}$, $u \le m$, – множество текущих различных значений признака x_t, $c_{i+1} > c_i$, $1 \le i \le u-1$.

<u>С4.5.</u> Текущим порогом называется число $d = \left(c_i + c_{i+1}\right)/2$, $1 \le i \le u-1$.

<u>Построенные РД.</u> Текущим порогом называется число $d = \left(c_i + c_{i+1}\right)/2$, $1 \le i \le u-1$, если в текущей обучающей выборке можно указать два обучающих объекта $S_1 = \left(a_{11},\ldots,a_{1n}\right)$ и $S_2 = \left(a_{21},\ldots,a_{2n}\right)$, принадлежащих разным классам, таких, что $a_{1t} = c_i$ и $a_{2t} = c_{i+1}$.

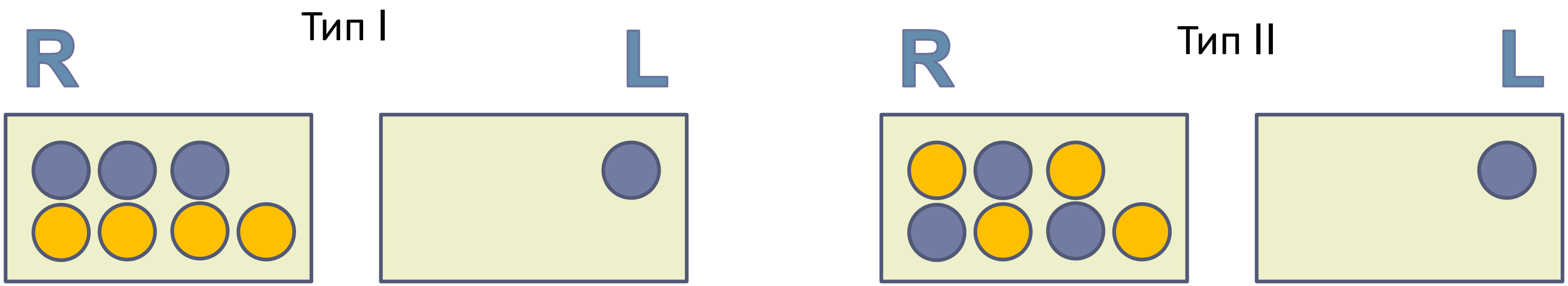


Исследуемые критерии выбора признака для построения внутренней вершины РД

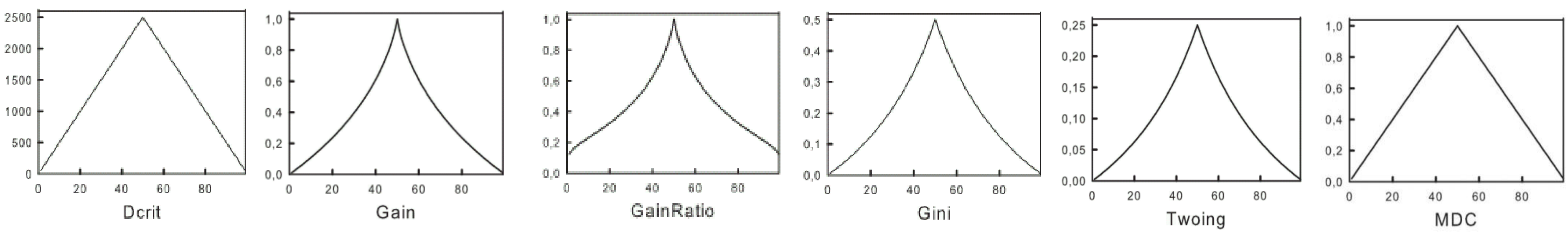
Пусть T – текущее множество, $T_d^{(1)}$ $\left(T_d^{(2)}\right)$ – множество объектов для левой (правой) ветви, R_t – множество объектов из T для которых значение признака x_t не определено. Обозначим через $P_t^i(T)=f(K_i,T\setminus R_t)\big/\left|T\setminus R_t\right|$, $f(K_i,T\setminus R_t)$ – число объектов из $T\setminus R_t$, принадлежащих классу K_i, $i\in I=\{1,\ldots,l\}$, $c_d^{(i)}(t)=\dfrac{T_d^{(i)}}{\left|T\setminus R_t\right|}$. Оптимальным порогом для признака x_t считается порог d, для которого:

1. $\mathrm{Gain}(x_t)_d=\mathrm{Info}(T)_t-\mathrm{Info}(x_t)_d\to\max$,
   $\mathrm{Info}(T)_t=-\sum\limits_{i\in\{1,2\}}P_t^i(T)\log P_t^i(T)$, $\mathrm{Info}(x_t)_d=-\sum\limits_{i\in\{1,2\}}c_d^{(i)}(t)\mathrm{Info}(T_d^{(i)})_t$.
2. $\mathrm{GainRatio}(x_t)_d=\mathrm{Gain}(x_t)_d\big/\mathrm{SplitInfo}(x_t)_d\to\max$, $\mathrm{SplitInfo}(x_t)_d=-\sum\limits_{i\in\{1,2\}}c_d^{(i)}(t)\log c_d^{(i)}(t)$
3. $\mathrm{Gini}(x_t)_d=\mathrm{Gini}(T)_t-\sum\limits_{i\in\{1,2\}}c_d^{(i)}(t)\mathrm{Gini}(T_d^{(i)})_t\to\max$, где $\mathrm{Gini}(T)_t=1-\sum\limits_{i\in I}\left(P_t^i(T)\right)^2$.
4. $\mathrm{Twoing}(x_t)_d=0.25c_d^{(1)}(t)c_d^{(2)}(t)\left(\sum\limits_{i\in I}\left|P_t^i(T_k^{(1)})-P_t^i(T_k^{(2)})\right|\right)^2\to\max$.
5. $\mathrm{Dcrit}(x_t)_d=\sum\limits_{i\in I}\prod\limits_{j\in I\setminus\{i\}}f(K_i,T_k^{(1)})f(K_j,T_k^{(2)})\to\max$.
6. $\mathrm{MDC}(x_t)_d=\sum\limits_{i\in I}c_d^{(i)}(t)\mathrm{MDC}(T_d^{(i)})\to\max$, $\mathrm{MDC}(T_d^{(k)})=\sum\limits_{i\in I}\sum\limits_{j\in I\setminus\{i\}}\dfrac{f(K_i,T_d^{(k)})}{f(K_i,T\setminus R_t)}-\dfrac{f(K_j,T_d^{(k)})}{f(K_j,T\setminus R_t)}$.

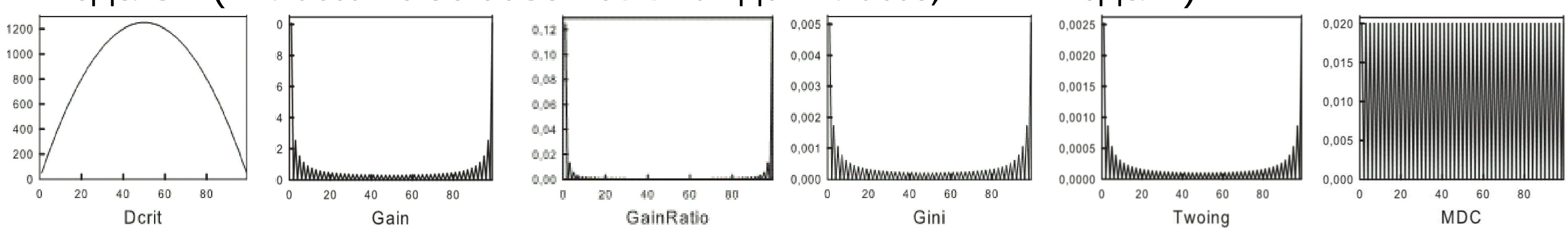
Численные эксперименты. Модельные данные (1-3)



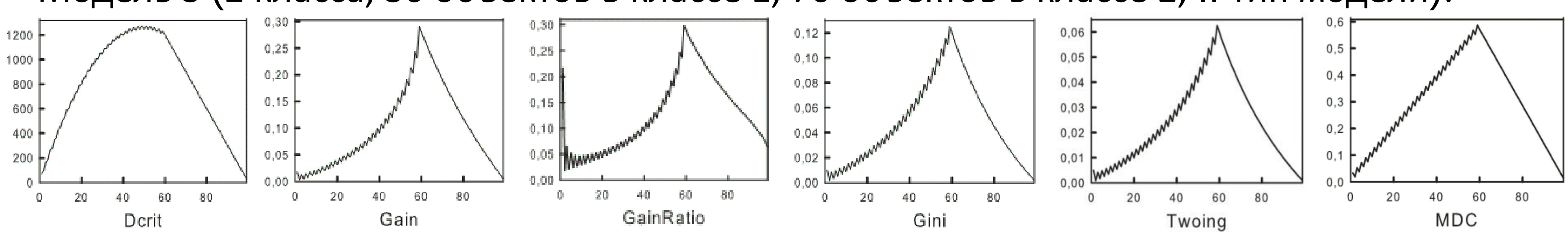
Модель 1 (2 класса по 50 объектов в каждом классе, I тип модели):



Модель 2 (2 класса по 50 объектов в каждом классе, II тип модели):

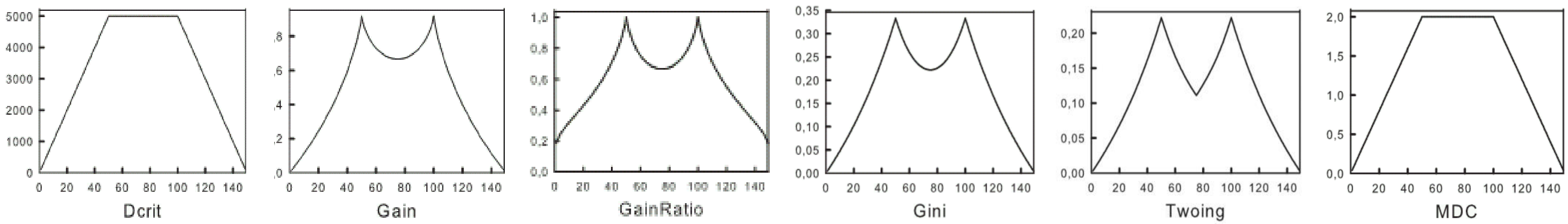


Модель 3 (2 класса, 30 объектов в классе 1, 70 объектов в классе 2, II тип модели):

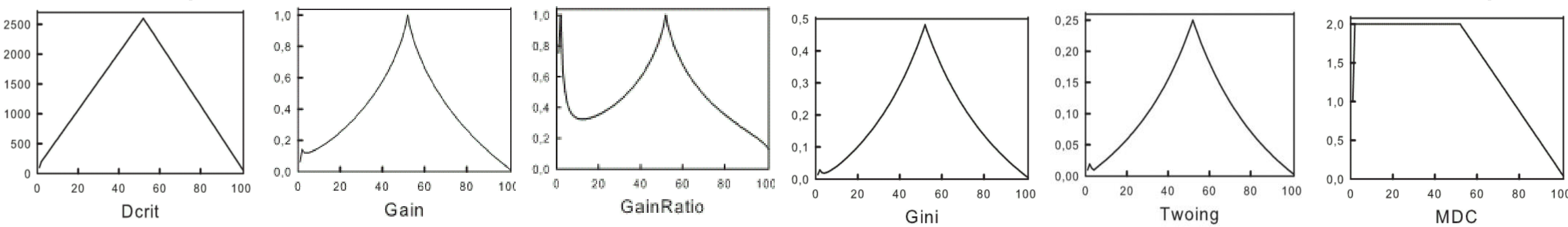


Численные эксперименты. Модельные данные (4-7)

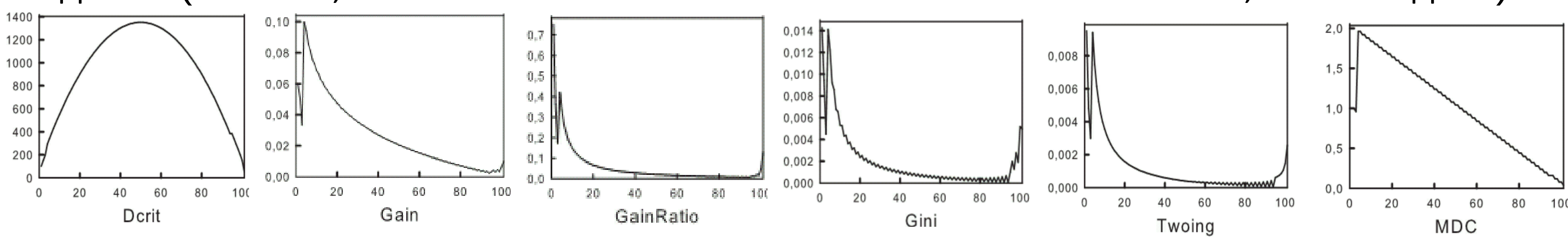
Модель 4 (3 класса по 50 объектов в каждом классе, I тип модели):



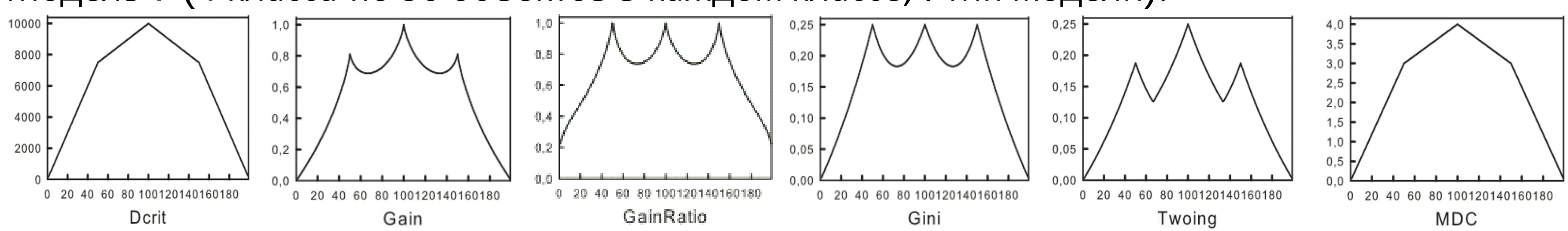
Модель 5 (3 класса, 2 объекта в классе 1 и по 50 объектов в классе 2 и 3, I тип модели):



Модель 6 (3 класса, 2 объекта в классе 1 и по 50 объектов в классе 2 и 3, II тип модели):

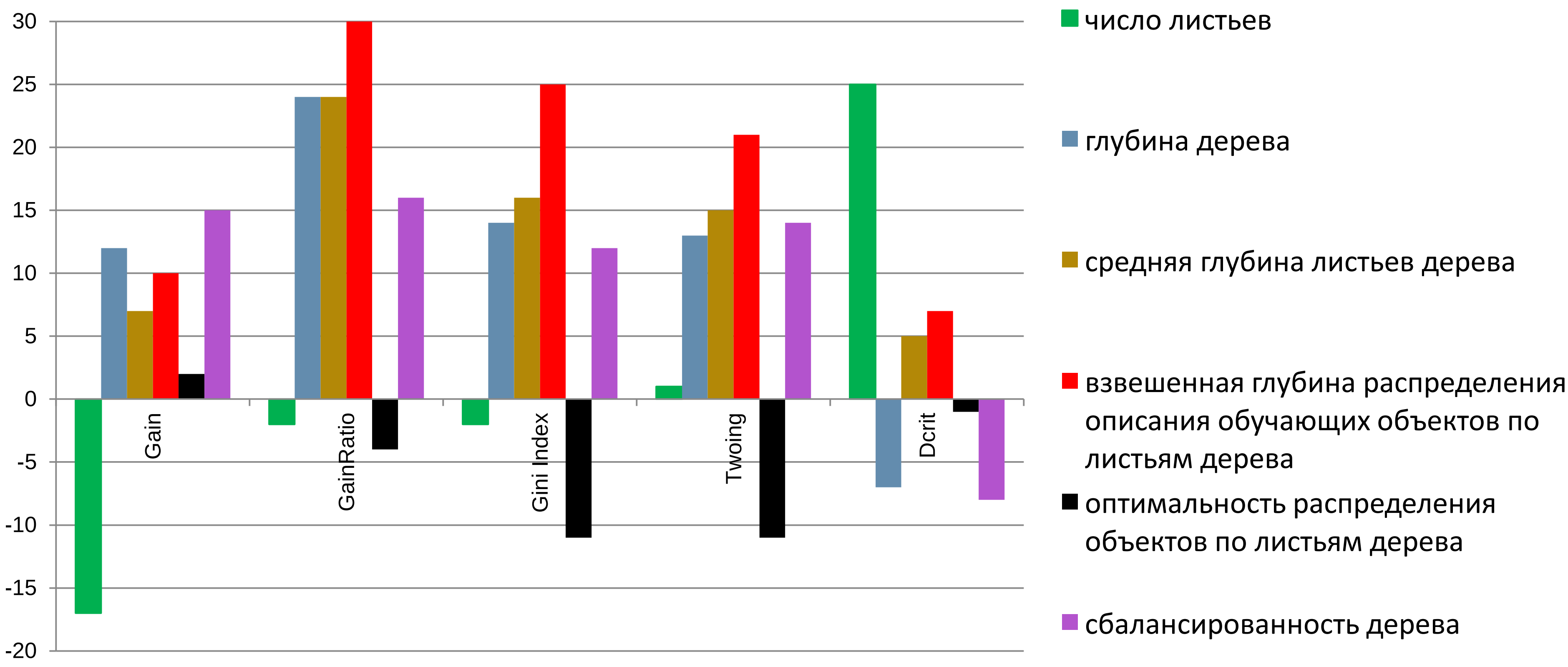


Модель 7 (4 класса по 50 объектов в каждом классе, I тип модели):



Численные эксперименты. Структурные свойства РД в зависимости от критерия ветвления

Пусть μ − число листьев, k_i − глубина i-ого листа, m_i − число обучающих объектов, попадающих в i-ый лист, $\max_{i=1,\dots,\mu} k_i$ − глубина РД, $\sum_{i=1}^{\mu} k_i/\mu$ − средняя глубина РД, $\sum_{i=1}^{\mu} k_i m_i/m$ – взвешенная глубина распределения описаний обучающих объектов по листьям РД, $\left|\sum_{i=1}^{\mu}\left(\frac{k_i}{\mu}-\frac{k_i m_i}{m}\right)\right|$ – «оптимальность» распределения обучающих объектов. Пусть $\Delta(x,y,i)$ − число задач, на которых РД с критерием x не хуже РД с критерием y по i-ой характеристике. По оси ординат – значение величины $\Delta(MDC,y,i)-\Delta(y,MDC,i)$, по оси абсцисс – критерий y.



Численные эксперименты. Качество РД с помощью LOO

$$\Theta = \sum_{i=1}^{l} q_i \Big/ l,\ q_i$$ – процент правильно распознанных объектов класса K_i

№ Задачи ($\lvert K_1\rvert$; …$\lvert K_l\rvert$; n)	MDC	Gain	Gain Ratio	Gini	Twoing	Dcrit	№ Задачи ($\lvert K_1\rvert$; …$\lvert K_l\rvert$; n)	MDC	Gain	Gain Ratio	Gini	Twoing	Dcrit
№ 1 (48; 12; 69)*	61,5	64,6	56,3	64,6	64,6	46,9	№ 19 (51; 218; 21)*	52,2	61,2	52,8	55,8	56,1	57,2
№ 2 (23; 173; 9)*	61	60,7	57,3	62,8	62,8	54,8	№ 20 (60; 15; 39; 5)*	70,5	64,3	69,3	68,2	68,2	68,2
№ 3 (23; 173; 17)*	58,6	54,7	54	44,8	45,4	44	№ 21 (47; 30; 7)*	83,6	89,1	89,7	89,1	88	83,6
№ 4 (152; 190; 15)*	79,1	78,4	81,4	78,7	78,7	74,7	№ 22 (40; 40; 18)*	68,8	72,5	72,5	68,8	68,8	55
№ 5 (16; 17; 12)*	51,3	51,5	57,5	48,5	48,5	75,9	№ 23 (11; 47; 15)	83,2	89,1	78,9	90,1	90,1	87,7
№ 6 (48; 23; 8)*	77,5	76,4	63,7	65,7	65,7	59,3	№ 24 (39; 22; 18)	66,3	69,3	80	72,6	72,6	73,1
№ 7 (89; 42; 9)*	76,3	78,6	75,1	78,1	78,1	59,5	№ 25 (52; 25; 8)*	55,7	48,9	44,5	54,8	56,8	64,5
№ 8 (76; 33; 24; 7)	86,9	86,9	90,6	88,3	88,3	89,5	№ 26 (59; 71; 48; 13)	95,6	93,4	95,6	88,4	88,1	91,2
№ 9 (86;31;22; 20; 8; 13)	37,3	34,6	41,5	27,7	31,5	36,4	№ 27 (458; 241; 9)*	94,6	93,1	94,4	94,5	94,5	94
№ 10 (120; 150; 13)	77,4	76,2	74,4	75	75	77,4	№ 28 (307; 383; 15)*	80	81,3	81,4	77,9	77,9	83,3
№ 11 (32; 123; 19)*	75,2	61,3	75,5	78,2	78,2	63,7	№ 29 (500; 268; 8)	69,5	72,4	66,2	70,6	70,6	68,3
№ 12 (218; 126; 9)*	95,1	96,1	94,2	94,8	94,8	95,4	№ 30 (70; 76; 17; 13; 9; 29; 9)	65	69,3	56,6	59,4	66,9	58,8
№ 13 (38; 107; 35)	69,8	75,4	77,4	77,3	77,3	68	№ 31 (126; 225; 34)	85	85,7	93,6	86,3	86,1	80,1
№ 14 (35; 72; 35)	59,7	58,2	65,4	65,3	65,3	57,7	№ 32 (300; 330; 309; 315; 310; 269; 302; 304; 276; 285; 7)	100	100	100	100	100	100
№ 15 (38; 35; 35)	60,2	60,3	82,3	58,8	58,8	68,6	№ 33 (20; 20; 20; 88; 44; 20; 20; 92; 20; 20; 20; 44; 20; 91; 91; 15; 14; 16; 8; 35)*	92,1	91,4	92,8	91,6	91,3	81,6
№ 16 (38; 72; 35)	76,4	84	76,4	76,5	76,5	69,1	№ 34 (626; 332; 9)	86,2	86,4	83	85	85	86,7
№ 17 (30; 102; 24)*	63,6	55,8	59,1	55	55	62	№ 35 (2; 81; 61; 4; 18)	60,9	46,8	58	51,4	57	52,9
№ 18 (51; 218; 24)*	47,8	58	54,6	54	54	56,2	№ 36 (112; 61; 72; 49; 52; 20; 34)	93,6	93,5	90,6	90,5	91,6	89,3
Среднее значение по всем задачам	72,7	72,8	73,2	71,9	72,5	70,4	**Число наилучших результатов по всем задачам**	10	10	13	5	5	6

Исследование качества РД с использованием теории отступов

Опр. Отступом (margin) объекта S называется величина $\mathrm{margin}(S) = \bar{\omega}^q(S) - \max\limits_{j \neq k}\{\bar{\omega}^j(S)\}$, где q – номер правильного класса для S, $\bar{\omega}^i(S)$, $i \in \{1,...,l\}$, – оценка классификатора за принадлежность объекта S классу K_i, $\sum_{j=1}^{l} \bar{\omega}^j(S) = 1$.

Пусть D – распределение на $M \times Y$, $Y = \{1,-1\}$ – метки классов, T – множество обучающих объектов из D. Пусть R_j – множество предикатов для признака x_j, э.к. B_i – конъюнкция предикатов. Обозначим через $U = \{R_1,...,R_n\}$.

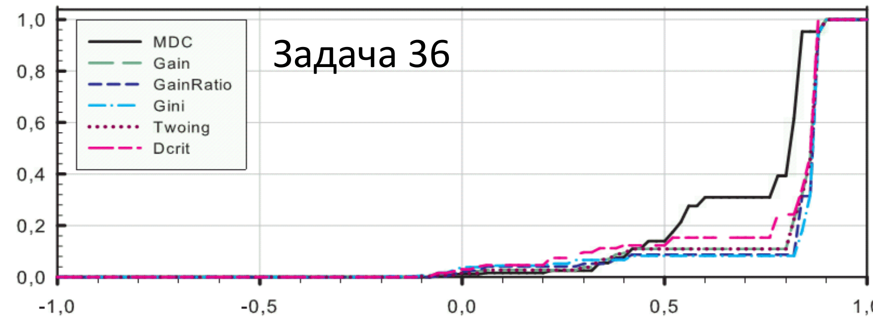
Пусть $P_T(\mathrm{margin} \le \theta)$ – вероятность того, что отступ для случайно выбранного объекта из T не превысит θ, $P_D[\mathrm{error}]$ – вероятность ошибки РД с μ листьями на объекте $S \in M$, $v = \sum_{i=1}^{\mu} \alpha_i d_i VCD(U)$, $\sum_{i=1}^{\mu} \alpha_i = 1$, $\alpha_i \ge 0$, $d = \max\limits_{i=1,...,\mu} d_i$, где d_i – глубина i-ого листа.

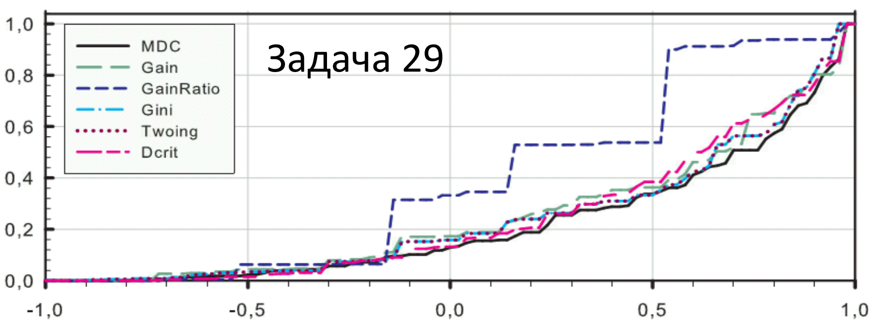
Теорема 1. Для любого $\delta > 0$ и для любого $\theta > 0$ с вероятностью не меньше $1-\delta$ справедливо

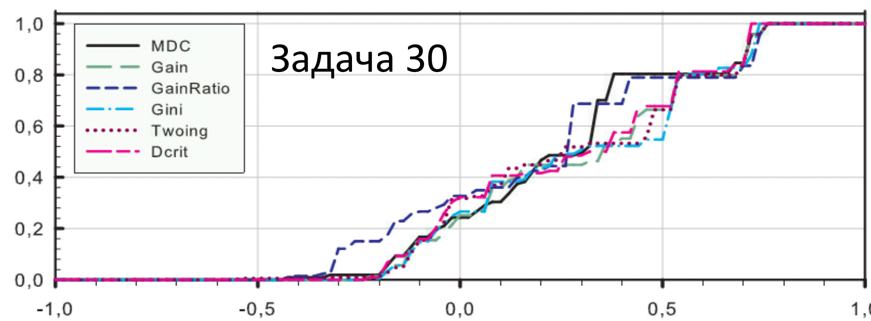
$$P_D[\mathrm{error}] \le 2P_T(\mathrm{margin} \le \theta) + c/m\left(1/\theta^2 (v \ln m + \ln d) \ln (m\theta^2/v) + \ln(1/\delta)\right),$$

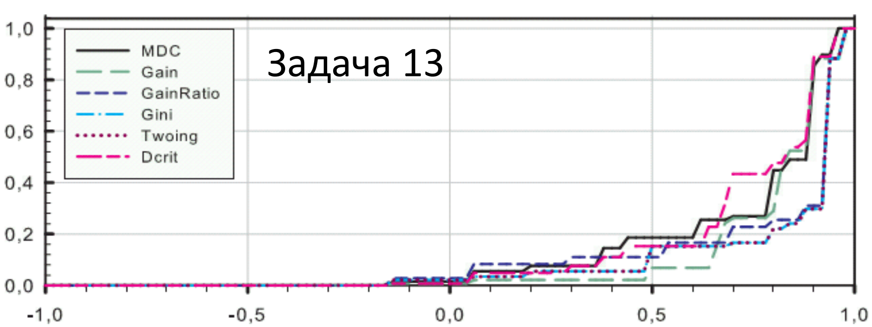
где $c > 0$ – константа.

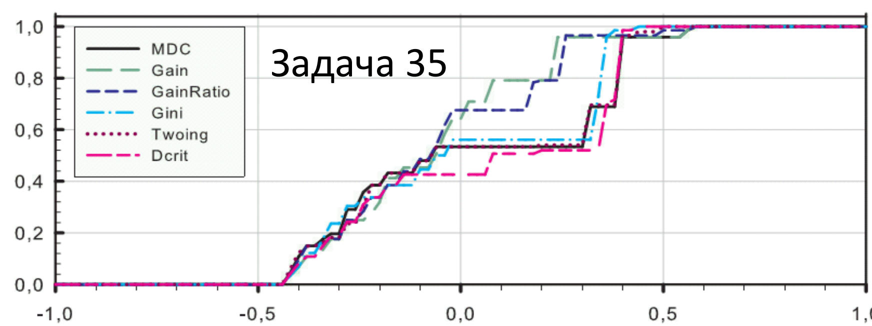
Численные эксперименты. Качество РД с использованием анализа отступов обучающих объектов











По оси ординат – $P_T(\text{margin} \le \theta)$, по оси абсцисс – значение величины θ.

Теорема 1. $P_D[\text{error}] \le 2P_T(\text{margin} \le \theta) + c/m(1/\theta^2(v \ln m + \ln d)\ln(m\theta^2/v) + \ln(1/\delta))$

Основные результаты

1. Разработан новый критерий выбора признака для построения внутренней вершины РД – критерий максимизации доли объектов различных классов (MDC).
2. На модельных данных проведено исследование особенностей разделения обучающих объектов при синтезе решающего дерева с помощью различных критериев ветвления (Gini Index, Twoing, Gain, GainRatio, критерий равномерного разбиения и критерий MDC).
3. На реальных задачах исследованы структурные свойства и качество решающего дерева в зависимости от применяемого критерия ветвления.
4. Показано, что применение нового критерия MDC позволяет получить сопоставимое по качеству и более оптимальное по структуре РД по сравнению с РД, построенного при использовании таких критериев, как: Gini Index, Twoing, Gain, GainRatio и критерий равномерного разбиения (Dcrit).