Bayesian model for rare events recognition with use of logical decision functions class (RESIM 08, Rennes, France)

Vladimir Berikov, Gennady Lbov

Sobolev Institute of mathematics,

Novosibirsk, Russia

{berikov, lbov}@math.nsc.ru

What are the problems we are solving?

- Pattern recognition
- Regression analysis
- Time series analysis
- Cluster analysis

in hard-to-formalize areas of investigation

Hard-to-formalize areas of investigations

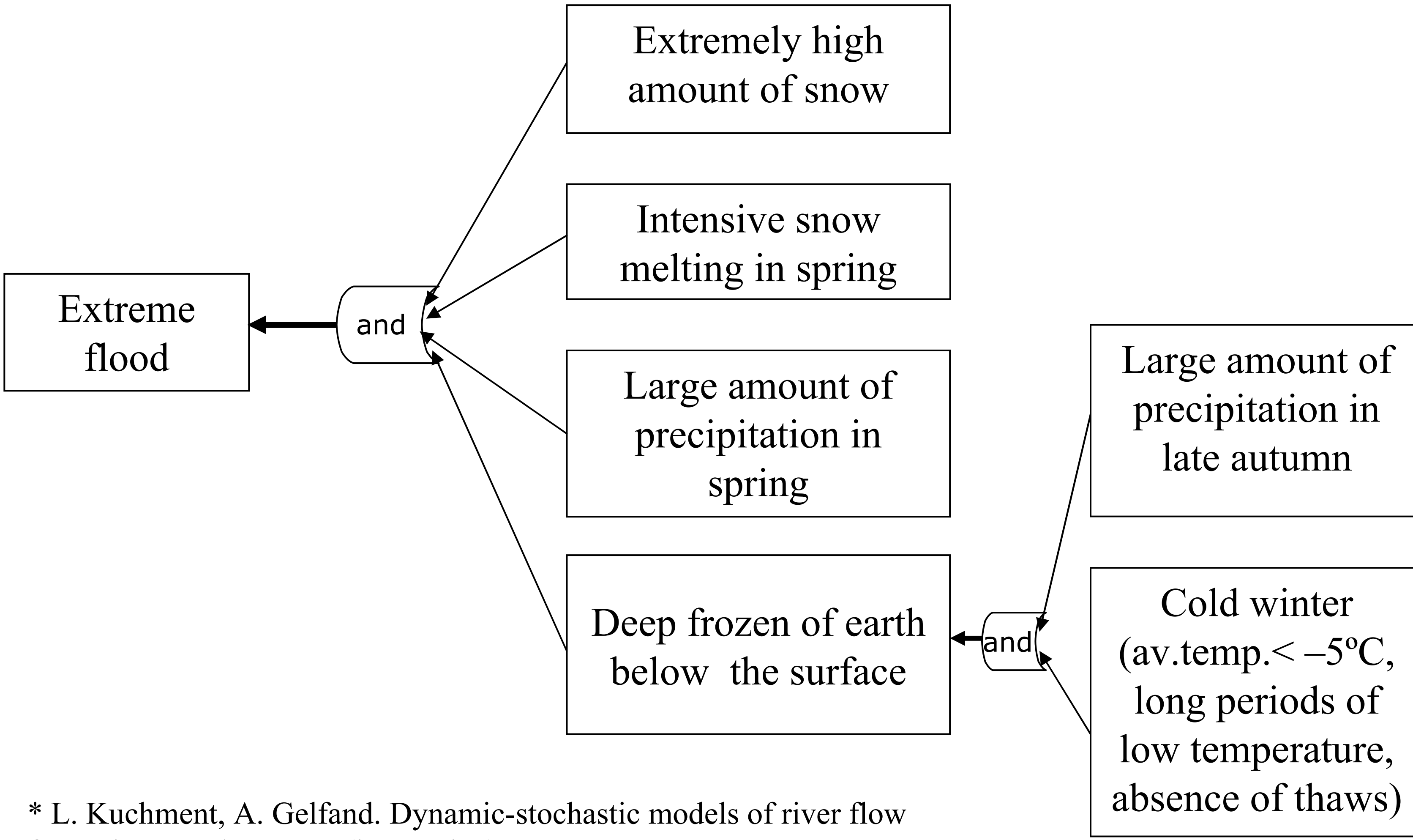
- lack of knowledge about the objects under investigation, that makes it difficult to formulate the mathematical model of the objects;
- large number of heterogeneous (either quantitative or qualitative) features; small sample size;
- heterogeneous types of expert knowledge;
- nonlinear dependencies between features;

- presence of unknown values of features;
- desire to present the results in the form understandable by a specialist in the applied area

Peculiarities of rare events:

- unbalanced data set;
- non-symmetric loss function;
- rare event with large losses = “extreme” event

Example of event tree for extreme flood forecast for rivers in Central Russia *

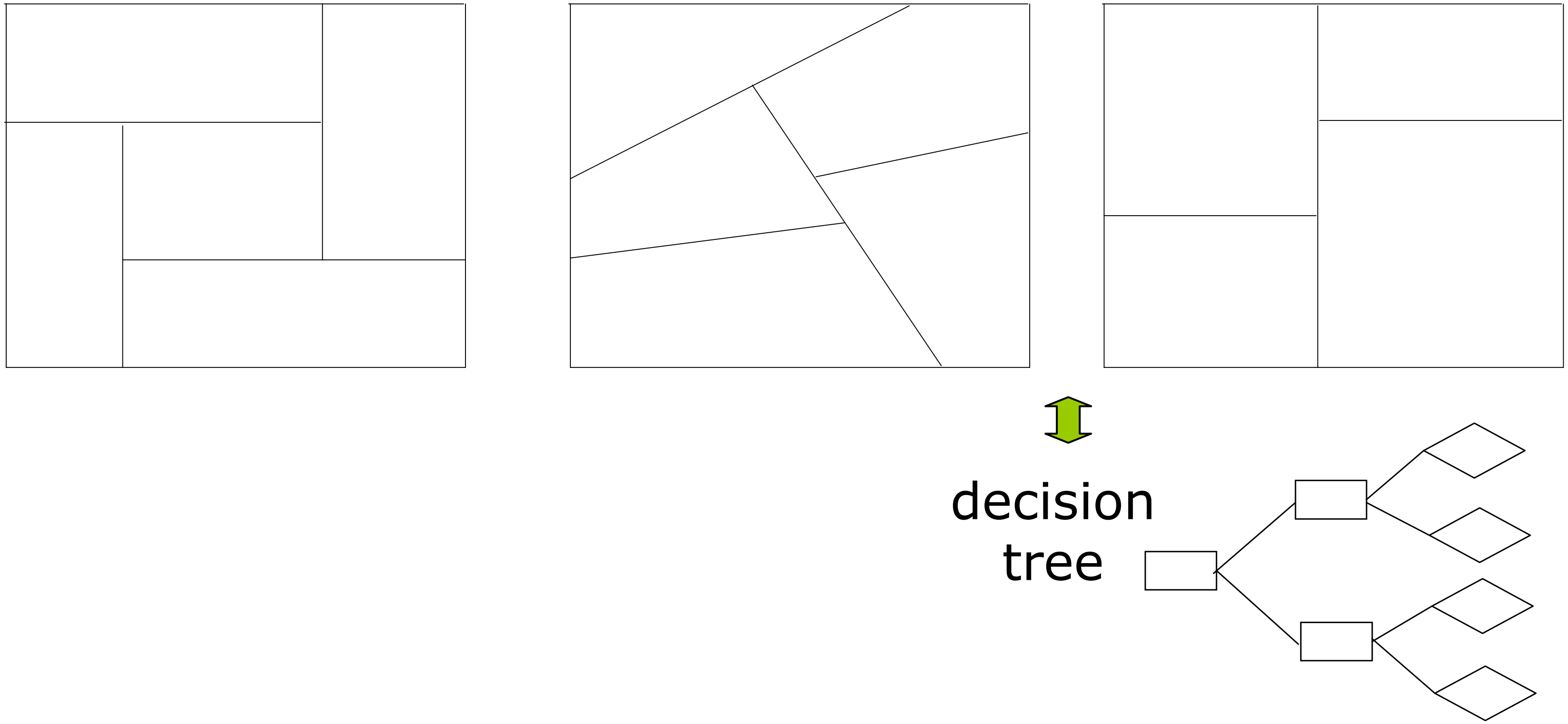


* L. Kuchment, A. Gelfand. Dynamic-stochastic models of river flow formation. Nauka, 1993. (in Russian)

Logical decision functions (LDF)

If P_1 and P_2 and…and P_m Then Y=1

e.g. P_j~ "$X_2>1$", P_j ~"$X_5 \in \{c,d\}$"



Basic problems in constructing LDF

- How to choose optimal complexity of LDF? (validity of quality criterion)
- How to find optimal LDF from given family (validity of algorithm)

Pattern recognition problem

Expert knowledge

Supposed class *of* distributions Λ

Class *of* decision functions *Φ*

Learning sample

X_1	X_2	…	X_n	Y
3	5	…	0	1
6	2	…	1	0
6	2	…	0	1
………………				
1	9	…	1	0

Goal: find $f \in \Phi$ having minimal risk of wrong recognition

Mathematical setting

- general collection Γ of objects
- features $X=(X_1,\ldots,X_j,\ldots,X_n)$, D_X

X_j — quantitative / qualitative / ordered

- Y, $D_Y=\{w^{(1)},\ldots, w^{(i)},\ldots, w^{(K)}\}$, $K \geq 2$ - number of patterns
- learning sample $(a^{(1)},\ldots,a^{(N)})$, N – sample size;
  $x^{(i)}=X(a^{(i)})$, $y^{(i)}=Y(a^{(i)})$
- $\theta=p(x,y)$ – distribution of (X,Y) ("strategy of nature")
- class of distributions Λ

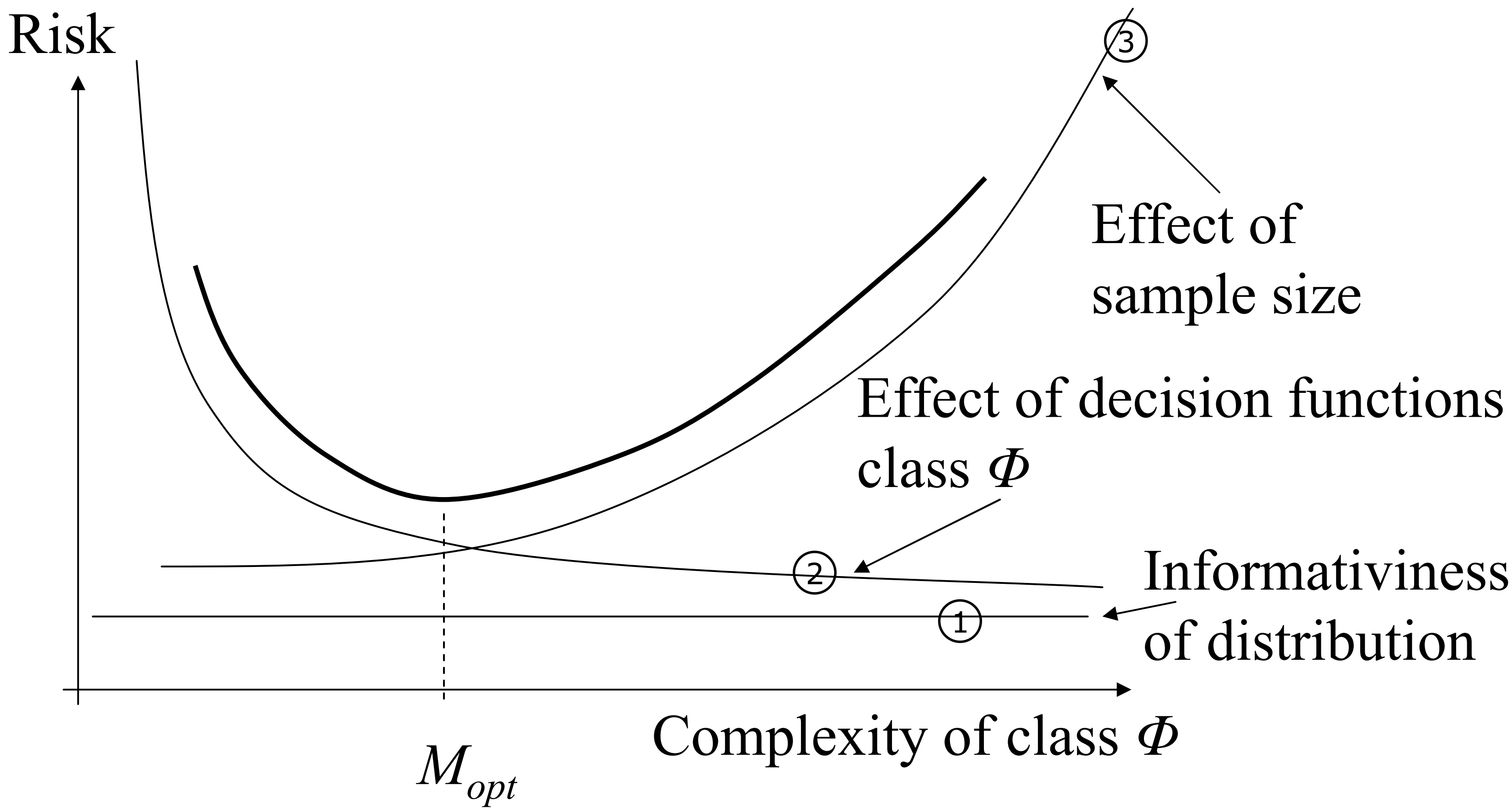
- a priori probabilities of patterns $p^{(1)},\ldots,p^{(K)}$
- decision function $f: D_X \to D_Y$
- class of decision functions Φ
- Loss function $L_{i,j}$ (decision $Y = i$, but actually $Y = j$)
  Y=1 ~ "extreme event"; Y=2 ~ "ordinary event"
  $L_{1,2} \ll L_{2,1}$
- expected losses (risk) $R_f(\theta)=\mathbf{E}\, L_{f(X),Y}$
- optimal Bayes decision function f_B: $R_{f_B}(\theta)=\inf_f R_f(\theta)$
- learning method μ: $f = \mu(s)$

- Probability distribution is unknown;
- Learning sample has limited size

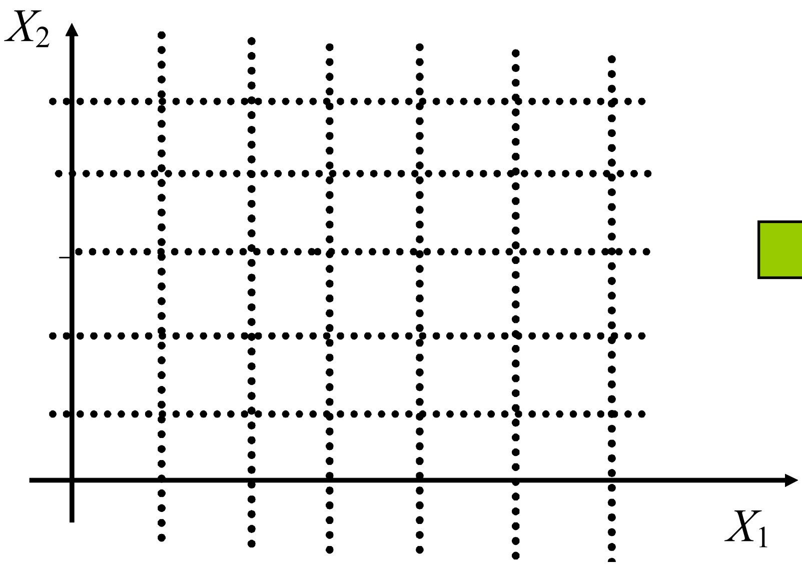
One should reach a compromise between the complexity of class and the accuracy of decisions on learning sample

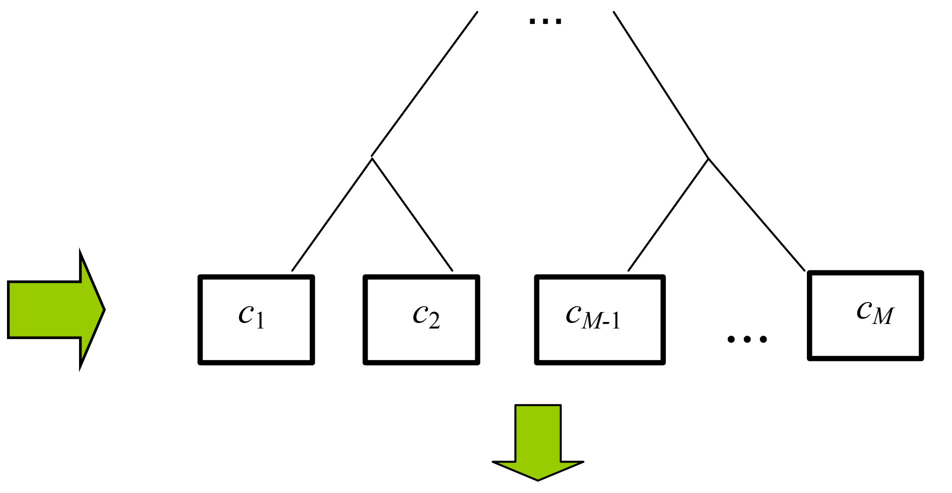
Complexity:

- Number of parameters of discriminant function;
- Number of features;
- VC dimension;
- <u>Maximal number of leaves in decision tree</u>;
- ...


Risk
3
Effect of
sample size
Effect of decision functions
class Φ
2
Informativiness
1
of distribution
Complexity of class Φ
M_{opt}

Recognition on a finite set of events





discrete unordered variable X

partition example

(independent from learning sample)

Bayesian approach: define meta-distribution on class of distributions Λ

X, Y; $D_X=\{1,\ldots,j,\ldots,M\}$ – cells, $D_Y=\{1,\ldots,i,\ldots,K\}$

$$p_j^{(i)} = \mathbf{P}(X = i, Y = j), \quad \theta = (p_1^{(1)},\ldots,p_j^{(i)},\ldots,p_M^{(K)})$$

$p^{(i)} = \sum_j p_j^{(i)}$ - a priori probability of i-th class

Frequency vector $s = (n_1^{(1)},\ldots,n_j^{(i)},\ldots,n_M^{(K)})$, $\sum_{i,j} n_j^{(i)} = N$

$\Lambda = \{\theta\}$ - family of multinomial distributions.

Random vector Θ is defined on Λ

$$p(\theta) = \frac{1}{Z}\prod_{i,j}(p_j^{(i)})^{d_j^{(i)}-1},$$

where $d_j^{(i)} > 0$ (Dirichlet distribution).

when $d_j^{(i)} \equiv d = 1$ – uniform a priori distribution

μ - deterministic learning method: $f = \mu(s)$

empirical error minimization method μ^*

Suppose that both sample S and strategy of nature Θ are random

Risk of wrong recognition – random function $R_{\mu(S)}(\Theta)$

Suppose that a priori probability of rare event ***is known*** $(p^{(1)};\ p^{(2)} = 1 - p^{(1)})$

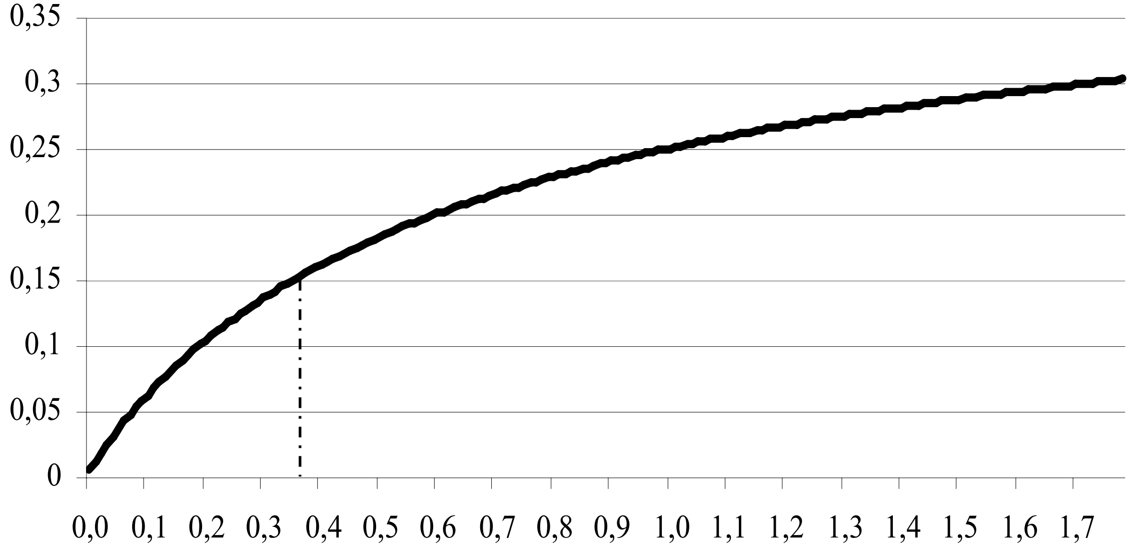
Directions in model investigation:

- How to set model parameters $d_j^{(i)}$?
- How to define optimal complexity of the class of LDF?
- How to substantiate quality criterion?
- How to get more reliable estimates of risk?
- How to extend model on regression analysis, time series analysis, cluster analysis?

Setting a priori distribution with respect to expected probability of error for Bayes decision function f_B

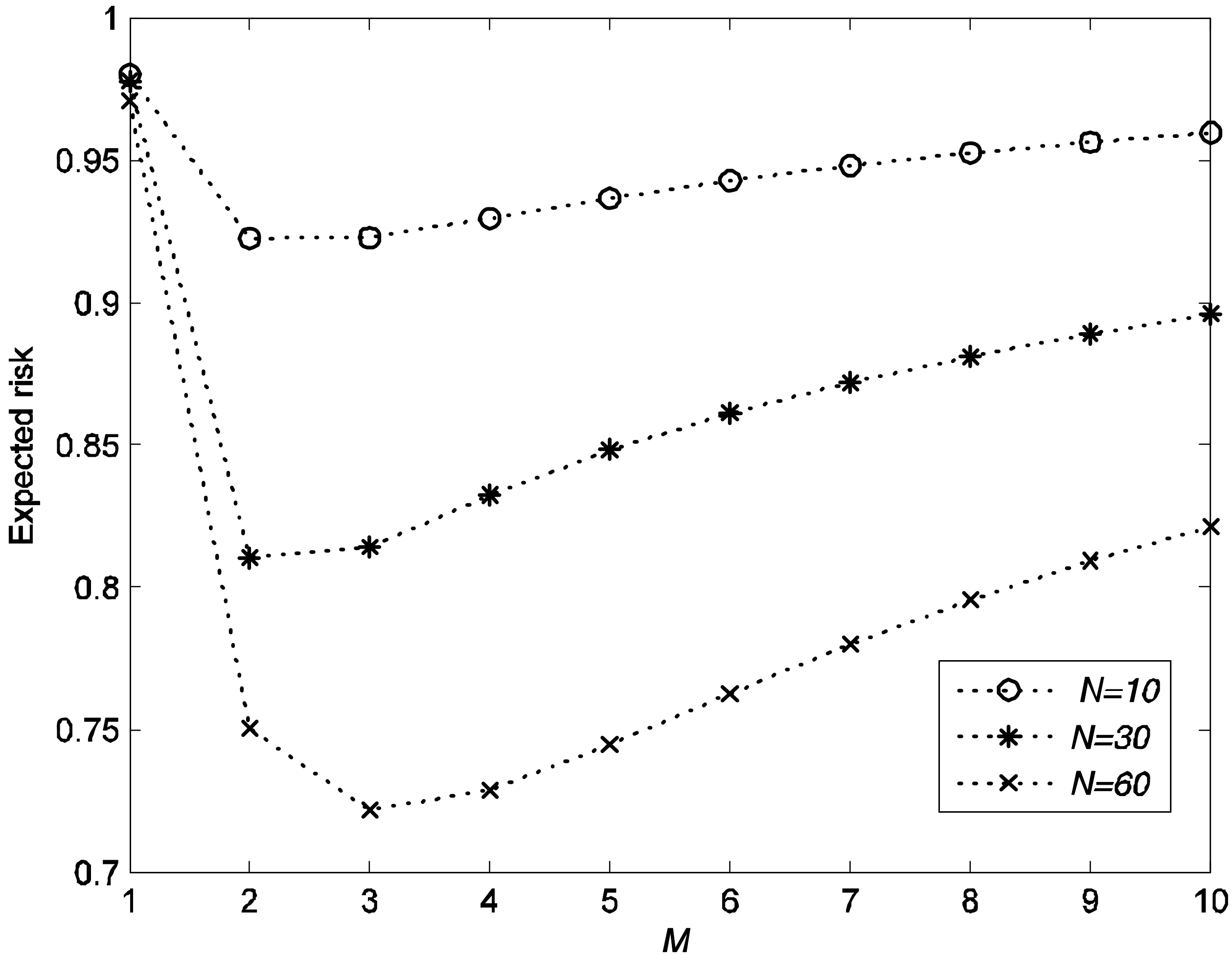
Let $K=2$, $d_j^{(i)} \equiv d$, where $d > 0$;

Proposition 1. $\mathbf{E}P_{f_B}(\Theta) = I_{0,5}(d+1, d)$, where $I_x(p,q)$ – beta distribution function.



For example, if $\mathbf{E}Pf_B = 0.15$, then d=0.38

Expected error probability $\mathbf{EP}_{\mu^*(S)}(\Theta)$ as a function of complexity (Theorem 1)



K=2, d=1,
$p^{(1)}$=0.05, $p^{(2)}$=0.95,
$L_{1,2}$=1, $L_{2,1}$=20,
$L_{1,1}$=$\mathrm{L}_{2,2}$=0

A posteriori estimates of risk (decision function f is fixed; learning sample is given)

Theorem 2. A posteriori mathematical expectation of risk function $\mathbf{E}_{\Theta|s} R_f(\Theta)$ equals

$$R_{f,s} = \frac{1}{N+D} \sum_{j,q} L_{f(j),q} (n_j^{(q)} + d_j^{(q)}), \text{ where } D = \sum_{i,j} d_j^{(i)}.$$

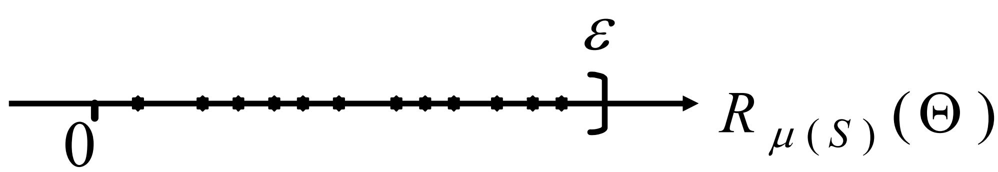
NB. *A posteriori mathematical expectation of risk is optimum Bayes estimate of risk under quadratic loss function* (see Lehmann, E. L., Casella G. Theory of point estimation. Springer Verlag, 1998.)

For K=2, $P_{f,s} \approx \tilde{n} + d\frac{M}{N}$, where $\tilde{n}$ is error frequency;

For d=1, $P_{f,s} \approx \tilde{n} + (K-1)\frac{M}{N}$ (LDF quality criteria)

Interval estimates of risk

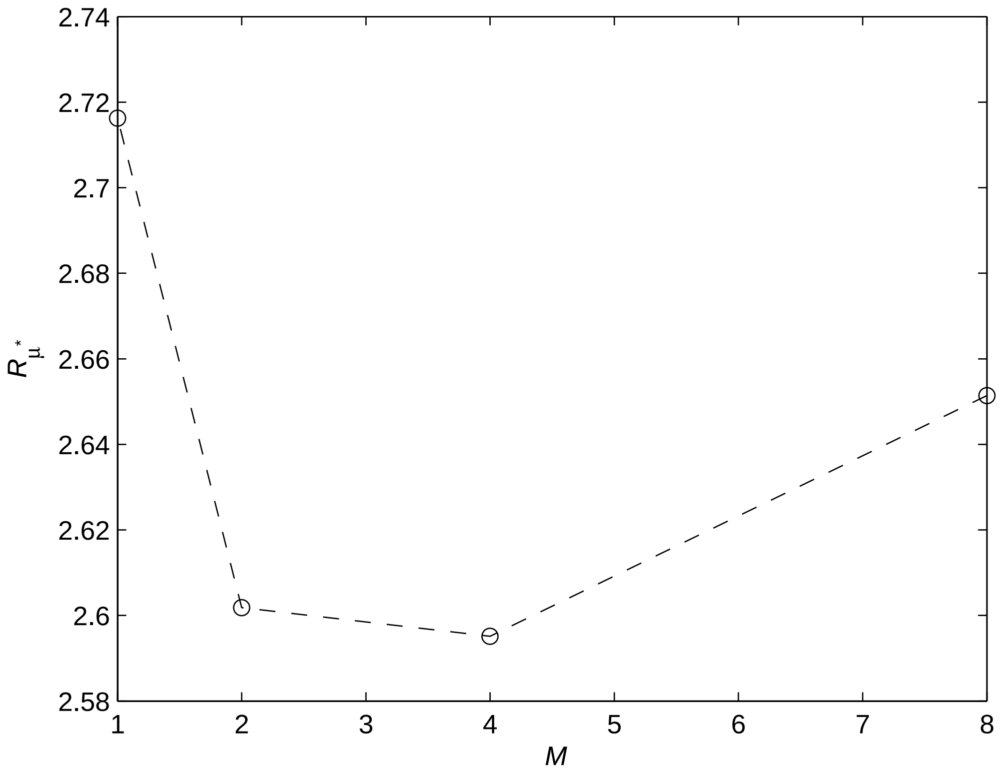
- Upper risk bound over strategies of nature and over samples of size N



$$\mathrm{P}(R_{\mu(S)}(\Theta) \le \varepsilon) \ge \eta,$$
$$\varepsilon = \varepsilon(\mu, N, M, \eta) \text{ (Theorem 3)}$$

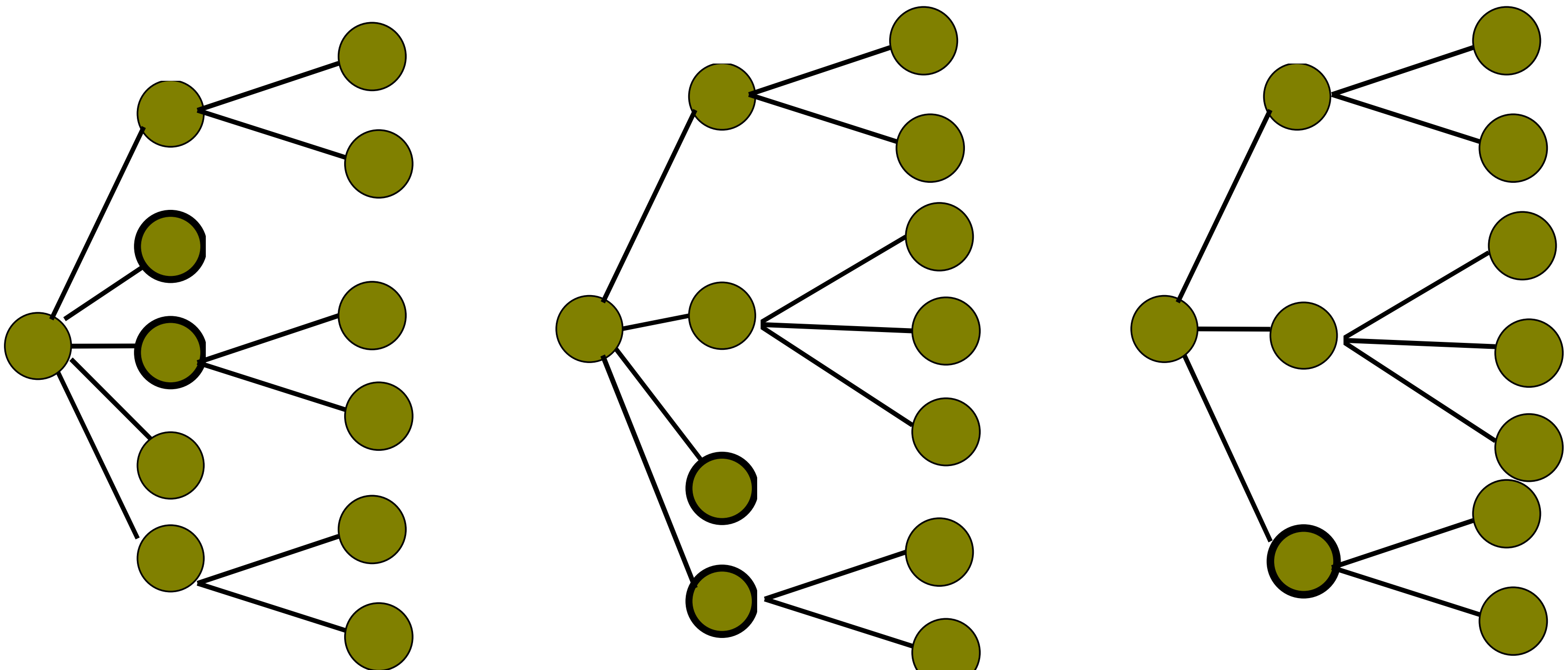
Ordered regression problem
(intermediate between pattern recognition and regression analysis)

Y – ordered discrete variable; loss function $L_{i,q}=(i-q)^2$, where $i,q=1,2,\ldots,K$ (Theorem 4).



$N=10$, $d_0=0.1$, $K=6$

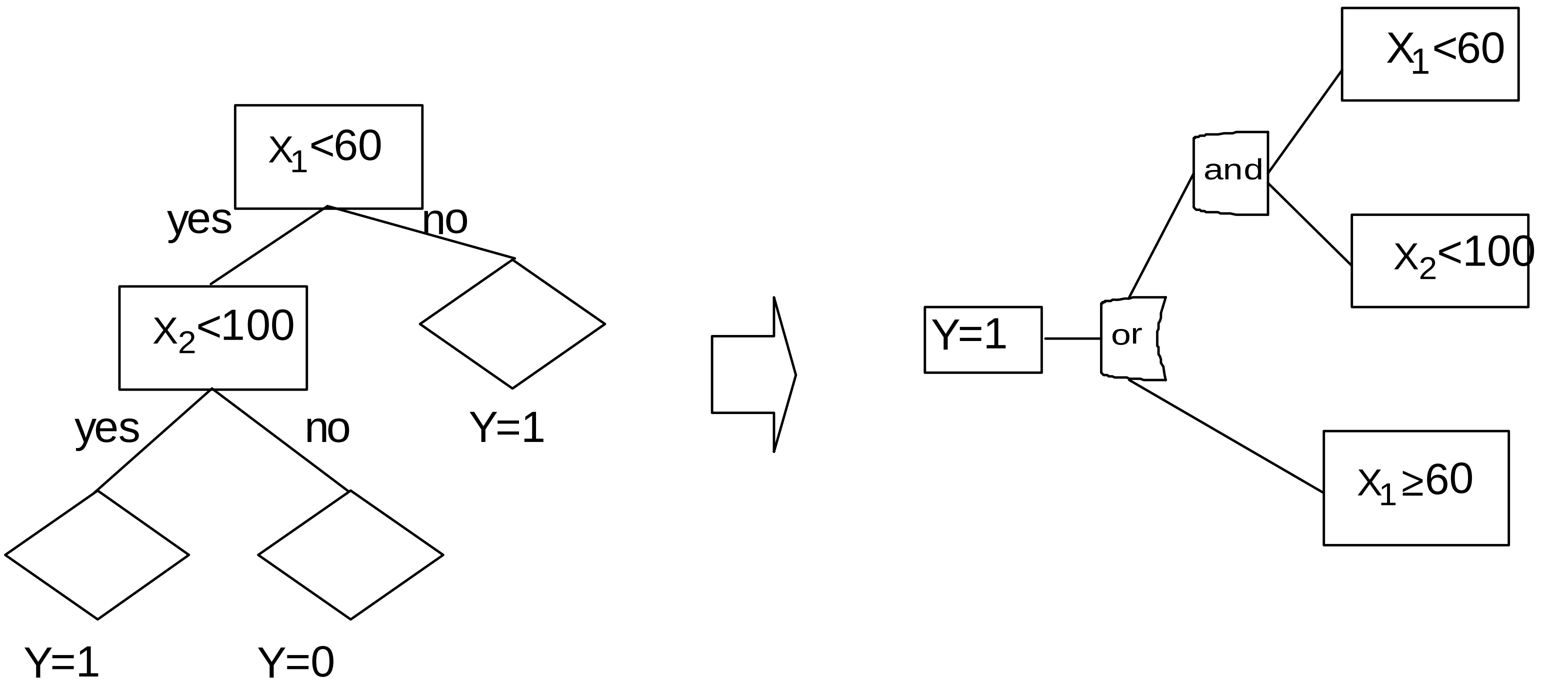
Recursive algorithm for decision tree construction



optimal number of branches;

optimal level of recursive embedding

Decision trees and event trees for rare events analysis



References

- Lbov,G.S. Construction of recognition decision rules in the class of logical functions // International Journal of Imaging Systems and Technology. V.4, Issue 1. 1991. P. 62-64.
- Berikov V.B., Lbov G. S. Bayes Estimates for Recognition Quality on Finite Sets of Events // Doklady Mathematical Sciences, Vol. 71, No. 3, 2005, P. 327–330.
- Berikov V.B., Lbov G. S. Choice of Optimal Complexity of the Class of Logical Decision Functions in Pattern Recognition Problems // Doklady Mathematical Sciences, 2007. V.76, N 3/1. P. 969-971
- Lbov, G.S., Berikov V.B. Stability of decision functions in problems of pattern recognition and analysis of heterogeneous information. Novosibirsk, Sobolev Institute of mathematics. 2005. (in Russian)

Thank you for your attention