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 *

- Extremely high amount of snow
- Intensive snow melting in spring
- Large amount of precipitation in spring
- Deep frozen of earth below the surface
- Large amount of precipitation in late autumn
- Cold winter (av.temp.< –5°C, long periods of low temperature, absence of thaws)

Logical decision functions (LDF)

If $P_1$ and $P_2$ and...and $P_m$ Then $Y=1$
e.g. $P_j \sim "X_2>1"$, $P_j \sim "X_5 \in \{c,d\}"$
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

Learning sample

<table>
<thead>
<tr>
<th>X₁ X₂ ... Xₙ</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 5</td>
<td>0 1</td>
</tr>
<tr>
<td>6 2</td>
<td>1 0</td>
</tr>
<tr>
<td>6 2</td>
<td>0 1</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>1 9</td>
<td>1 0</td>
</tr>
</tbody>
</table>

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

Expert knowledge

Supposed class of distributions \( \Lambda \)

Class of decision functions \( \Phi \)
Mathematical setting

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

  - $X_j$ quantitative
  - $X_j$ qualitative

- $Y$, $D_Y=\{w^{(1)},\ldots,w^{(i)},\ldots,w^{(K)}\}$, $K \geq 2$ - 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 $\Lambda$
• a priori probabilities of patterns $p^{(1)}, \ldots, p^{(K)}$
• decision function $f : D_X \to D_Y$
• class of decision functions $\Phi$
• Loss function $L_{i,j}$ (decision $Y = i$, but actually $Y = j$)
  $Y=1 \sim "extreme \ event"; \ Y=2 \sim "ordinary \ event"$
  $L_{1,2} \ll L_{2,1}$
• expected losses (risk) $R_f(\theta) = E L_{f(X),Y}$
• optimal Bayes decision function $f_B : R_{f_B}(\theta) = \inf_f R_f(\theta)$
• learning method $\mu : f = \mu(s)$
• 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;
- Maximal number of leaves in decision tree;
- ...
Complexity of class $\Phi$

Risk

Effect of decision functions class $\Phi$

Effect of sample size

Informativeness of distribution

$M_{opt}$

Complexity of class $\Phi$
Recognition on a finite set of events

partition example

(discrete unordered variable $X$)

(independent from learning sample)
Bayesian approach: define meta-distribution on class of distributions $\Lambda$

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

$p^{(i)}_j = P(X = i, Y = j), \quad \theta = (p^{(1)}_1, \ldots, p^{(i)}_j, \ldots, p^{(K)}_M)$

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

Frequency vector $s = (n^{(1)}_1, \ldots, n^{(i)}_j, \ldots, n^{(K)}_M), \sum n^{(i)}_j = N$

$\Lambda = \{\theta\}$ - family of multinomial distributions.
Random vector $\Theta$ is defined on $\Lambda$

\[
p(\theta) = \frac{1}{Z_{i,j}} \prod_j (p^{(i)}_j)^{d^{(i)}_j-1},
\]

where $d^{(i)}_j > 0$ (Dirichlet distribution).
when \( d_j^{(i)} = d = 1 \) – uniform a priori distribution

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

*empirical error minimization* method \( \mu^\ast \)

Suppose that both sample \( S \) and strategy of nature \( \Theta \) 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 $K=2$, $d_j^{(i)} = d_j$, where $d > 0$;

**Proposition 1.** $E P_{f_B}(\Theta) = I_{0.5}(d+1,d)$, where $I_x(p,q)$ – beta distribution function.

For example, if $E P_{f_B} = 0.15$, then $d=0.38$
Expected error probability $\mathbb{E}P_{\mu^*(S)(\Theta)}$ as a function of complexity (Theorem 1)

$K=2, d=1,$
$p^{(1)}=0.05, p^{(2)}=0.95,$
$L_{1,2}=1, L_{2,1}=20,$
$L_{1,1}=L_{2,2}=0$
**Theorem 2.** A posteriori mathematical expectation of risk function \(E_{\Theta|s}R_f(\Theta)\) 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 \(K=2\), \(P_{f,s} \approx \bar{n} + d \frac{M}{N}\), where \(\bar{n}\) is error frequency;

For \(d=1\), \(P_{f,s} \approx \bar{n} + (K - 1) \frac{M}{N}\) (LDF quality criteria)
Interval estimates of risk

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

\[
P(R_{\mu(S)}(\Theta) \leq \varepsilon) \geq \eta,
\]

\[
\varepsilon = \varepsilon(\mu, N, M, \eta) \text{ (Theorem 3)}
\]
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


Thank you for your attention