DOI https://doi.org/10.30525/978-9934-26-043-8-5

ESTIMATOR OF MULTIDIMENSIONAL BAYESIAN THRESHOLD IN TWO-CLASS CLASSIFICATION

Kubaychuk O. O.
Candidate of Sciences in Physics and Mathematics, Associate Professor,
Professor at the Special Department № 1
Institute of Specialized Communication and Information Security
of the National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute"
Kyiv, Ukraine

Object classification by its numerical characteristic is an important theoretical problem and has practical significance, for example, the definition of a person as "not healthy", if the temperature of its body exceeds 37°C. To solve this problem we consider the threshold-based rule. According to this rule, an object is classified to the first class if its characteristic does not exceed a threshold $t=37°C$; otherwise, an object is classified to belong to the second class. The empirical Bayesian classification (EBC) and minimization of the empirical risk (MER) are widely used methods to estimate the best threshold. The case when the learning sample is obtained from a mixture with varying concentrations is considered in [1, pp. 40-47] and the asymptotic of both methods of estimating is investigated.

The technique of a nonparametric analysis of mixtures where concentrations changes from observation to observation develops, actively. The problem of distributions estimating in case at known concentrations is considered in [2, p. 618; 3, pp. 123-128], estimates of concentrations in two-component mixtures in [4, p. 71]. The correction algorithms for weighted empirical distribution functions are proposed in [5, p. 315], [6, p. 518].

However, it is often necessary to classify an object in case of more than one threshold, for example, the definition of a person as "not healthy", if the temperature of its body exceeds 37°C or lower then 36°C. Another example: the person is sick, if the level of its haemoglobin exceeds 84 units or lower than 72 units. In particular, this problem is discussed in [7, pp. 47-50], [8, pp. 78-85]. The case of two thresholds and three prescribed classes was studied in [9, p. 6262].

In the current paper, is assumed that the object may belong to one of two prescribed classes. An unknown class number containing 0 is denoted by

$ind(O)$. A classification rule (briefly, classifier) is a function $g: R^n \to \{1,2\}$ that assigns a value to $ind(O)$ by using characteristic ξ. In general, classification rule is defined as a general measurable function, but we restrict the consideration to the so-called threshold-based classification rules of the forms

$$g^1_{2m-1,\mathbf{t}}(\xi)=\begin{cases}1, & \xi\in\left\{\bigcup\limits_{k=1}^{m-1,m>1}[t_{2k},t_{2k+1}]\right\}\bigcup\{(-\infty,t_1]\},\\ 2, & \xi\in\left\{\bigcup\limits_{k=1}^{m-1,m>1}(t_{2k-1},t_{2k})\right\}\bigcup\{(t_{2m-1},+\infty)\},\end{cases}$$

$$g^2_{2m-1,\mathbf{t}}(\xi)=\begin{cases}1, & \xi\in\left\{\bigcup\limits_{k=1}^{m-1,m>1}(t_{2k-1},t_{2k})\right\}\bigcup\{(t_{2m-1},+\infty)\},\\ 2, & \xi\in\left\{\bigcup\limits_{k=1}^{m-1,m>1}[t_{2k},t_{2k+1}]\right\}\bigcup\{(-\infty,t_1]\},\end{cases}$$

if $n=2m-1, m\in N$ and

$$g^1_{2m,\mathbf{t}}(\xi)=\begin{cases}1, & \xi\in\left\{\bigcup\limits_{k=1}^{m-1,m>1}(t_{2k},t_{2k+1})\right\}\bigcup\{(-\infty,t_1)\}\bigcup\{(t_{2m},+\infty)\},\\ 2, & \xi\in\left\{\bigcup\limits_{k=1}^{m}[t_{2k-1},t_{2k}]\right\},\end{cases}$$

$$g^2_{2m,\mathbf{t}}(\xi)=\begin{cases}1, & \xi\in\left\{\bigcup\limits_{k=1}^{m}[t_{2k-1},t_{2k}]\right\},\\ 2, & \xi\in\left\{\bigcup\limits_{k=1}^{m-1,m>1}(t_{2k},t_{2k+1})\right\}\bigcup\{(-\infty,t_1)\}\bigcup\{(t_{2m},+\infty)\},\end{cases}$$

If $n=2m, m\in N$, where $\mathbf{t}=(t_1,t_2,\ldots t_n)$ is the multidimensional threshold [10, pp. 342-351], [11, p. 61].

When determining the best threshold, one faces the problem of estimating the threshold by using a learning sample, whose members are classified correctly. We consider the Bayesian empirical classification method, in assumption, that a learning sample is obtained from a mixture with varying concentration.

The distribution functions H_s are assumed to be unknown. One can estimate these functions from the data $\Xi_N = \{\xi_{j:N}\}_{j=1}^{N}$ being a sample from a mixture with varying concentration [12, pp. 226-231], where $\xi_{j:N}$ are independent if N is fixed and

$$P\{\xi_{j:N} < x\} = w_{j:N} H_1(x) + (1 - w_{j:N}) H_2(x),$$

Here $w_{j:N}$ is a known concentration in the mixture of objects of the first class at the moment when an observation j is made.

To estimate the distribution functions H_s, we use weighted empirical distribution functions [13, p. 48], [14, p. 98], [15, pp. 83-92]

$$\widehat{H}_s^N(x) = \frac{1}{N} \sum_{j=1}^{N} a_{j:N}^s 1\{\xi_j < x\}$$

where $1\{A\}$ is the indicator an event A and $a_{j:N}^s$ are known weight coefficients:

$$a_{j:N}^1 = \frac{1}{\Delta_N} \left((1 - S_N^1) w_{j:N} + (S_N^2 - S_N^1) \right),$$

$$a_{j:N}^2 = \frac{1}{\Delta_N} \left(S_N^2 - S_N^1 w_{j:N} \right),$$

$$S_N^k = \frac{1}{N} \sum_{j=1}^{N} (w_{j:N})^k, \quad k = 1, 2, \qquad \Delta_N = S_N^2 - (S_N^1)^2.$$

In the assumptions, described above, we found the conditions of convergence in probability of the estimator for the Bayesian threshold constructed by the method of empirical-Bayesian classification for a sample from a mixture with variable concentrations.

References:

1. Ivan'ko Y., Maĭboroda R. The asymptotic behavior of threshold-based classification rules constructed from a sample from a mixture with varying concentrations. *Theory of Probability and Mathematical Statistics*. 2007. *74*. P. 37-47.

2. Maiboroda R. E. Estimates for distributions of components of mixtures with varying concentrations. *Ukrainian Mathematical Journal*. 1996. *48*(4). P. 618-622.

3. Maiboroda R. E. An asymptotically effective probability estimator constructed from observations of a mixture. *Theory of Probability and Mathematical Statistics*. 1999. 59. P. 121-128.

4. Maıboroda R. E. Projective estimates for changing concentrations of mixtures. *Theory of Probability and Mathematical Statistics*. 1993. 46. P. 71–75.

5. Kubaychuk O. Fast correction algorithms for weighted empirical distribution functions. *Advances and Applications in Statistics*. 2019. 54(2). P. 315-326. http://dx.doi.org/10.17654/AS054020315

6. Kubaychuk O. Fast correction algorithm lr_up(p1,p2) for weighted empirical distribution function // Theoretical and scientific bases of development of scientific thought. Abstracts of V International Scientific and Practical Conference. Rome, Italy 2020. P. 518-523. URL: https://isg-konf.com. Available at : DOI: 10.46299/ISG.2021.I.V

7. Кубайчук О. Асимптотика оцінки для баєсового порогу. *Вісник Київського національного університету імені Тараса Шевченка. Математика. Механіка*. 2008. 19-20. С. 47-50.

8. Kubaychuk O. O. The estimator asymptotic behavior of the empirical risk minimization method for bayesian border. *Research Bulletin of NTUU "Kyiv Polytechnic Institute"*. 2010. 4. P. 78-85.

9. Kubaychuk O. The asymptotic behaviour of threshold-based classification rules in case of three prescribed classes. *Journal of Advances in Mathematics*. 2016. *12*(5). P. 6262-6269.

10. Kubaychuk O. EBC-Estimator of Multidimensional Bayesian Threshold in Case of Two Classes. *Journal of Statistical Theory and Applications.* 2020. *19*(3). P. 342-351.

11. Kubaychuk O. Classification of objects based on threshold rules // Theoretical foundations of modern science and practice. Abstracts of XI International Scientific and Practical Conference. Melbourne, Australia 2020. P.61-63. URL: http://isg-konf.com.

12. Kubaychuk O. Estimation of moments by observations from mixtures with varying concentrations. *Theory of Stochastic Processes.* 2002. Vol. 8, No. 3-4, P. 226-231.

13. Kubaychuk O. O. Estimation of moments from mixtures with the use of improved weighted empirical distribution function. *Visnyk Kyiv Univ. Matematyka. Mekhanika*. 2003. *9*(10). P. 48-52.

14. Maiboroda R., Kubaichuk O. Asymptotic normality of improved weighted empirical distribution functions. *Theory of Probability and Mathematical Statistics*. 2004. *69*. P. 95-102.

15. Maiboroda R., Kubaichuk O. Improved estimators for moments constructed from observations of a mixture. *Theory of Probability and Mathematical Statistics*. 2005. *70.* P. 83-92.