Adv Syst Sci Appl 2022; **04**; 233-241
Published online at http://ijassa.ipu.ru.

Estimation of the Posterior Probabilities of Classes by the Approximation of the Anderson Discriminant Function

Valery Zenkov

V.A. Trapeznikov Institute of Control Sciences of RAS, Moscow, Russia
Email: zenkov-v@yandex.ru

Abstract: The approximation of the Anderson discriminant function at a given point in the feature space of two classes using a supervised training sample allows estimating the posterior probabilities of classes as easily as converting Fahrenheit degrees to Celsius degrees. These probabilities make it possible to further solve the classification problem with subjectively specified costs of classification errors and criteria. The training sample is simply converted to a regression analysis sample by replacing the class numbers with differences in error costs. A nonparametric method for approximating the discriminant function at a point is proposed. It does not require the specification of an approximation function at a point more complex than a linear one. Examples are given.

Keywords: machine learning, Anderson discriminant function, approximation, class posterior probability, weighted least squares.

1. INTRODUCTION

The posterior probabilities (PoP) of the classes are exhaustive information for solving the problem of object classification. They allow attributing a classification object to a particular class, taking into account different matrices of the costs of classification errors, set by the user not only for the entire task as a whole, but also for some categories of classified objects separately, for example, by the status of the classified object. We solve the problem in two steps. At the first stage, for the object of classification, we find the PoP of the classes - objective data. At the second stage, we assign the object to one or another class, taking into account the subjective preferences of the user: different the classification error costs, the method of making a decision taking into account some restrictions, etc.

In connection with the importance of the PoP of the classes for solving the classification problem and for interpreting the decisions made when classifying objects, add-ons to some already existing classification methods are used. So, for the support vector machine, the Platt calibrator is used [1], which estimates the PoP of the classes by the distance of a point in the class feature space to the boundary between classes. In this case, a hypothesis is put forward about the form of the dependence of the PoP of the classes on the distance of the point to the boundary between the classes, and then the parameters of the hypothetical dependence are found using the maximum likelihood method using the training set.

The Anderson discriminant function (ADF), according to the sign of which a point in the class feature space belongs to one of two classes, is a regression dependence. It is convenient in that, according to its approximation, obtained by the weighted least squares method from a supervised training set and by the cost of classification errors used to determine it, we calculate PoP classes as simply as we get degrees Celsius from degrees Fahrenheit.

We define ADF from the Bayesian solution of the classification problem [2], obtained by Theodore Wilburne Anderson. ADF is the difference between the two functions of average losses from classification errors in the space of attributes of the two classes at given costs of losses from classification errors.

We approximate the ADF using a supervised training set of the classification problem, transforming it into a regression analysis set by replacing the class numbers (labels) with the corresponding differences in the cost of classification errors. When approximating the ADF at a point, there is no need to specify the form of the approximating dependence in the class feature space. It is enough to construct a linear approximation of the ADF by the closest points to a given point, or by the Nadaray-Watson method [3] to find the ADF value at this point.

The PoP estimates of the classes at a point based on the ADF approximation at this point do not depend on the choice of error costs used to determine the ADF. You can set the error costs equal, for example, 1 and 1, or not equal, for example, 2 and 5. The ADF and approximations at the point will be different, but the PoP estimates at the point will be the same. This is the main property of ADF.

The classification error costs are used to transform the supervised training set into a regression analysis set. Class numbers in the first case are replaced by -1 and +1 in the first case and by -2 and +5 in the second case. Due to the independence of the estimates of the PoP of the classes on the choice of the costs of classification errors, the classification error can always be set to 0.5 and 0.5.

The method for obtaining PoP estimates from the ADF approximation does not depend on the class imbalance in the training sample. Class imbalance in the training set occurs when the number of points of one class is much less than the number of points of another class. Class imbalance makes it difficult to detect objects of a small class if classification methods are used that are not based on PoP estimations and costs of classification errors.

The method for solving the classification problem at a point in the feature space using a supervised training set, when we use the entire training set for each classified point, is nonparametric, and the resulting model is a nonparametric model.

2. ANDERSON DISCRIMINANT FUNCTION

We define ADF as the difference of two functions of average losses of classification errors in a problem with two classes:

$$f_{12}(x,C) = G_1(x,C) - G_2(x,C) = M_{k|x}(C_{1k} - C_{2k}), \tag{2.1}$$

where $G_1(x,C) = C_{12}(1-p(1|x))$, $G_2(x,C) = C_{21}p(1|x)$ there are average losses at the point, if the point is classified as class 1 or, respectively, class 2; C_{ij} is the cost of the error when a point from a class j is mistakenly referred to a class i. $C_{ij} > 0$, $i \# j$, $C_{ii} = 0$. $p(k|x)$ is the PoP of the class k at the point x, $p(1|x) + p(2|x) = 1$, $p(k|x) = P_k p(x|k)/p(x)$, $p(x) = \sum_k P_k p(x|k)$, Pk is the prior probabilities of classes, $p(x|k)$ is the conditional distributions of class features; k is the class numbers 1 or 2; $M_{k|x}(\cdot)$ is the mathematical expectation of at a point.

If $f_{12}(x, C) \le 0$, then the point referred to class 1, otherwise to class 2.

If $f_{12}(x, C) \le 0$, then the point referred to class 1, otherwise to class 2.

3. ADF PROPERTIES

3.1. ADF is by definition a regression Function.

To transform the training set of the supervised classification problem into a regression analysis set, due (2.1), on the set, the class number 1 should be replaced by $-C_{21}$ and the class number 2 by C_{12}.

3.2. The first class PoP and the ADF defined for given C12 and C21 are identically related

$$p(1|x) \equiv \left(C_{12} - f_{12}(x,C)\right)/\left(C_{12}+C_{21}\right). \quad (3.1)$$

Corollary 3.1:

If we set the cost of errors under the condition $C_{12}+C_{21}=1$*, that it does not lead to a loss of generality, then the equal (3.1) is simplified*

$$p(1|x) = p^* - f_{12}(x,p^*) \quad (3.2)$$

and on the boundary between the classes, where $f_{12}(x,p^*) = 0$, $p^* = p(1|x) = C_{12}$ *and* $1-p^* = C_{21}$*. The PoP of the first class on the class boundary is p*.*

Corollary 3.2:

The PoP of the classes at the point x are not addicts on for which $p^* > 0$*, or* C_{12} *and* C_{21}*, the ADF is defined.*

For the formal proof, it suffices to substitute expression (2.1) for ADF in (3.1) or (3.2).

Corollary 3.3:

When choosing the error costs C for classifying a point by the PoP of the classes obtained at the first stage of solving the classification problem, at the second stage there will be indistinguishability of classes in the feature space if

$$\left(min_x f_{12}(x,C) > 0\right) \text{ v } \left(max_x f_{12}(x,C) < 0\right). \quad (3.3)$$

In this case, all points will need to be assigned to the same class.

4. ADF APPROXIMATION METHODS

The ADF can be approximated in at least three ways:

– To approximate the ADF at a point in the class feature space to obtain estimates of the PoP of the classes at this point according to (3.1) or (3.2). This is a nonparametric model for solving a problem.

– To approximate the ADF in the entire area of the training set, having specified the type of the approximating function and choosing the approximation method. For this parametric model, a method is proposed for approximating the ADF in the initially unknown neighborhood of its zero values, to solve more accurately the classification problem for the selected type of approximating dependence and the given costs of classification errors and not pursue the goal of evaluating the PoP of the classes.

– To approximate several ADF for the given PoP values of the first class at the boundary between the classes, for example, 0.1, 0.3,…, 0.9 and for a given type of approximating dependence. This is a variant of a series of ADF approximations. For a given point, the PoP of the classes is found by the method of inter- or extrapolation over neighboring approximations of the ADF, between which the point is located.

The nonparametric model is slower it uses the entire training set in the test stage. Parametric ones work faster, but they require the choice of the type of the approximating function for ADF in the training stage.

4.1. ADF Approximation at Point

This version builds a nonparametric model. Its distinctive feature is the need to use all or part of the supervised training set to estimate the PoP of the classes at each given point.

In [4], to estimate the ADF at a point, we select such costs of classification errors for which the ADF approximation at a given point takes zero value. But more preferable, in our opinion,

is the ADF approximation by a plane (at least by a second-order polynomial) in the vicinity of a given point.

The criterion for approximating the ADF at a given set point x_j is the weighted root mean square error. At the training stage, for each set point x_j used as a test one, there is a vector of coefficients λ_j approximating the ADF in the form $(1, x_j)\,\lambda_j$ for given values of the parameters W and S of the weight function, for example, an exponent:

$$Q(\lambda, x_j) = \min_{\lambda_j} \sum_{n=1, n\neq j}^{n=N} [C_{1k_n} - C_{2k_n} - (1, x_n)\lambda_j]^2 \exp(-W \parallel x_j - x_n \parallel^S), \quad (4.1)$$

where the row vector $(1, x_n)$, in addition to the row vector of features x_n of the training set of components, contains a component one. N is the number of rows in the set without row x_j at the training stage, as required by the LOO overfitting method (Leave-One-Out). k_n is the class number 1 or 2 in the set line n. x_n is the feature vector in the set line n. $W \geq 0$ is a weighting factor that sets the rate of decay of the weighting function (in this case, the exponent is from the distance of the training set point to a given point). S is an exponent to which the distance from the sampling point to a given point is raised. $C_{1k_n} - C_{2k_n}$ is the estimate of the ADF in the set row n, obtained by replacing the class number k_n, 1 or 2, in the row by the difference in error costs determined by the given p^*.

Having obtained the vector of coefficients λ_j by minimizing criterion (4.1) for the given parameters of the method W and S, we calculate the estimate of the ADF at the point x_j, and, using (3.1), (3.2), we find the estimate of the PoP of the first class at the point x_j. If the number of classes $K > 2$, then similarly we obtain the PoP of the class at this point, using the method one class against all the others.

The best values of the parameters W and S are selected, for example, according to the minimum criterion

$$R = N^{-1}(C_{12}N_2 + C_{21}N_1) \quad (4.2)$$

where N_1 is the number of points of the first class, erroneously assigned to the second class, N_2 is the number of points of the second class, erroneously assigned to the first class. Often $S = 1$ and the value of the only parameter W is selected. The type of the weighting function is not necessarily exponential.

4.2. ADF Approximation in the Neighborhood of Zero Values

In some cases (a large volume of the training set, the presence of a priori information about the conditional distributions of features of classes, or about the type of surface separating the classes), it is more expedient to build a parametric model. It can be in the form of a discriminant function of a given form with unknown parameters determined from the training set. To classify a point, we can use the Bayes formula to recalculate the reconstructed conditional distributions of class features to the PoP of the classes at the specified point and, taking into account the costs of classification errors, assign the point to a particular class.

We propose in [5] a heuristic method for constructing a parametric model of the classification problem based on a supervised training set, used for the approximation of the discriminant function in the vicinity of its zero values for given costs of classification errors. In this case, we usually do not talk about evaluating the PoP of the classes.

The supervised training set does not contain information about the position of the ADF zero points. For given values of classification errors, there may be no boundary between classes and all points should be assigned to the same class.

To solve the problem, a heuristic method [5] of approaching the boundary between classes is used. First, based on the given costs of classification errors, the training set is transformed into a set of the regression analysis problem by replacing the class numbers (labels) with the corresponding differences of the costs of classification errors (2.1). Then, several iterations of solving the problems of approximation of ADF by the weighted least squares method are

started. The weight function from the distance of the point to zero on the previously obtained ADF approximation is used as the weights of the set points. The measure of the distance of a point to the zero of the previous approximation is the module of the previous approximation of the ADF at this point. Solution criterion at each step with an exponential weighting function:

$$Q(\lambda_i) = \min_{\lambda} \sum_{n=1}^{n=N} \{[C_{1k_n} - C_{2k_n} - \lambda_i' \varphi(x_n)]^2 \exp(-W_i \mid \lambda_{i-1}' \varphi(x_n) \mid^S)\}, \quad (4.3)$$

where i is the iteration number, $i = \{1,2,...,I\}$. At the first step, the problem is solved without a weighting function $(W_0 = 0)$; at subsequent steps, the weighting function gives more weight to points closer to the zero region of the previous ADF approximation. I is a given number of iterations $I < \infty$. W_i is a given weighting factor at a step S is a given exponent, usually $S = 1$. The weighting factor in iterations can be constant or change, for example $W_i = w*i,\ i = \{0,1,...,I\}$. For each iteration, the losses are found from the obtained ADF approximation (4.3). The best value of λ is the vector to which the lower losses correspond (4.2).

4.3. Estimation of PoP of the class at a point from a series of ADF Approximations

In this parametric method [6], based on a supervised training set, we construct several approximations of the ADF specified for multiple PoP of the first class at the boundaries between classes. This completes the training process and the training set is no longer used.

The constructed series of approximations can be used to estimate the PoP of the first-class at a given point in the feature space by two methods: the interpolation method and the smoothing method.

In the interpolation method for estimating the PoP of the first class at a given point by the signs of the ADF approximations in the series, at this point, there is a pair of neighboring approximations, between which the point is located. If a pair is found, then the interpolation method yields an estimate for PoP of the first class. A measure of the distance to neighboring boundaries and known by the construction of PoP of the first class on these boundaries is the module of the values of neighboring ADF approximations at a given point. If a given point is outside the limits of the ADF approximations, then the estimate of the PoP of the class is found by extrapolation from the first or last ADF approximation in the series, checking the result for non-negativity and not exceeding one.

In the smoothing method, the weighted p_j* values used to obtain the j-approximations of the ADF are used to estimate the PoP of the classes at a given point. As a measure of the distance of a point to the boundary j between classes, the modulus of the value of the j-approximation of the ADF at a given point $\left| F_j(x, p_j*) \right|$ is used. As a weighting function, for example, the exponent is used

$$w_j(x) = \exp(-W \mid \mathrm{F}_j(x, p_j^*) \mid), \quad j = 1 \div J, \quad (4.4)$$

where W is the parameter setting the smoothing properties of the method.

The weighted estimated PoP at a point using a series of ADF approximation and (4.4) will be

$$p(1 \mid x) = \sum_j p_j^* w_j(x) \ / \ \sum_j w_j(x). \quad (4.5)$$

The best value of W can be selected according to (4.2), performing classifications according to the PoP of the classes (4.5).

5. EXAMPLES

5.1. Approximation of the ADF at a Point

We present the case of solving the problem by minimizing (4.3) and (4.2) in Fig. 5.1 with normal conditional distributions of two-dimensional features of three classes. One of them, class A, is located between classes B and C. Class A is designated by the number 1, classes B and C will be combined into one class and designated by the number 2. Class middles: $m_a = (0,\ 0)$, $m_b = (-3,\ 0)$, $m_c = (3,\ 0)$. The covariance matrices of the classes are the same. The prior probabilities of the classes $P_a = 0.5$, $P_b = P_c = 0.25$. The set generated by this parameter contains 120 points. The figure shows the theoretical densities of conditional distributions of class features, theoretical ADF and PoP of the first class, as well as estimates of PoP at given points, obtained from $S = 1$ and $W = 3$ (4.2).

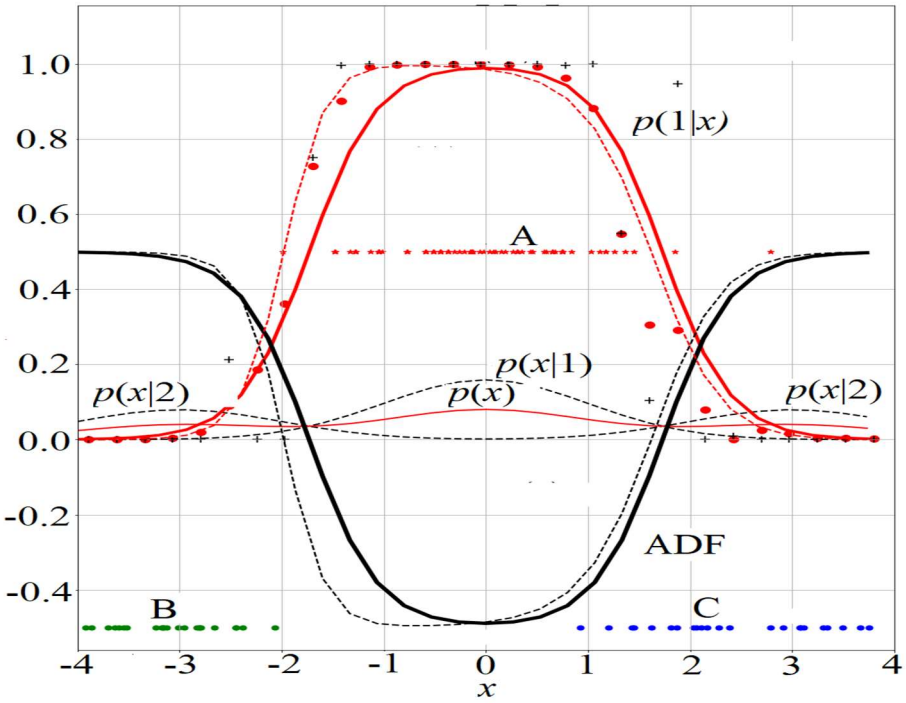

Fig. 5.1

The root-mean-square errors (RMS) in the estimation of PoP of the first-class about with concern to the theoretical ones were 0.09 and 0.18, respectively, for the linear and polynomial approximation of the second-order of the ADF. The linear approximation of ADF turned out to be more accurate in this set.

The estimates of PoP of the first class at the given points are located near the curve of PoP of the first class, obtained from the estimates of the parameters of the normal laws of conditional distributions of features (the dashed curve next to the solid theoretical curve of PoP of the first class).

Since there are two features, the flat figure shows a section by a plane passing along the axis of the values of the first feature and the ordinate.

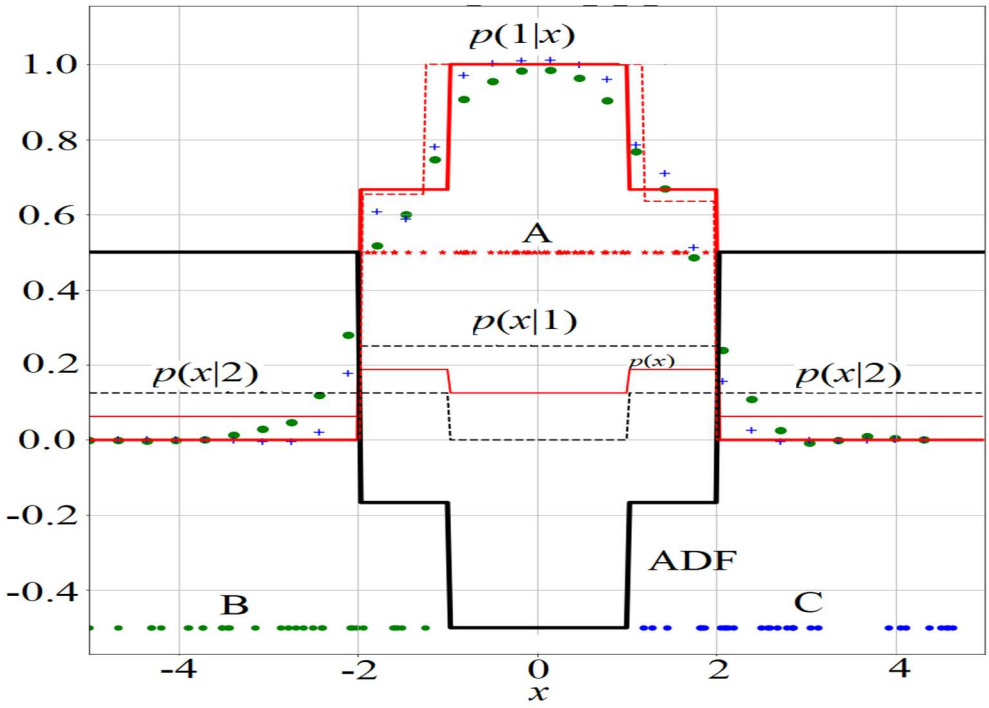

Fig. 5.2

Fig. 5.2 shows an example with a one-dimensional feature distributed according to a uniform law of probability distribution. Classes A, B and C have averages $m_a = 0$, $m_b = -3$, $m_c = 3$. The limits of maximum deviations from the average in both directions are the same and equal to 2. The classes are intersected in the feature space. The prior probabilities of the classes are the same as in the previous case. There are 120 points in the training set, $S = 1$ and $W = 3$. Visually, the estimates of PoP of the first class by a second-order polynomial in being more accurate.

Explanations for Fig. 5.1 and Fig. 5.2. The sampling points of the three classes A, B, and C are located at levels p^* and $1 - p^*$. The conditional distributions of features of classes 1 and 2 (one against combining the other two) are the dashed lines $p(x|1)$ and $p(x|2)$ in the figures. The solid thin line is the distribution density of x, $p(x)$. The inscription ADF denotes the position of the theoretical ADF (solid thick line with convexity downward). $p(1|x)$ is the theoretical PoP of the first class (solid thick line with a convex upward). Close to theoretical dashed lines is ADF and PoP constructed from the training set, for the known conditional distributions of class attributes and the given costs of classification errors, $p^* = 0.5$.

The ADF constructed for the training set in Fig. 5.2 is not shown. It is a mirror image of the PoP constructed for the training set (dashed line). Round points near PoP of the first-class – estimates of PoP by approximating ADF at given points; crosses — estimates of PoP using a second-order polynomial to approximate the ADF.

5.2. Approximation of the ADF in the Vicinity of Zero Values

When choosing the type of ADF approximating dependence, one can use the features, according to their correlation coefficients with the desired value. So in the example [5] with the dimension of the feature space of 217, the training set consisted of only 252 lines. Three were selected with a correlation coefficient estimates of the desired value not lower than 0.64 and with a cross-correlation, not higher than 0.66. The classification error according to the ADF approximation in the vicinity of the zeros according to the linear approximation was 6.8 %, according to the second-order polynomial – 4.8%. For comparison, the classification

method based on the hypothesis of conditional normal distributions of the selected features of the classes, according to the polynomial and according to the linear model, was 10.3% and 9.9%, respectively.

We made [7] a comparison on 15 examples of the quality of solving classification problems by the ADF approximation method in the vicinity of zeros and the support vector machine method (SVM) for linear ADF approximation and approximation by a second-order polynomial at equal costs of classification errors. In 14 examples, the first method gave a lower classification error than the second. A few examples may not be a sufficient basis for judging the advantages of one method over another.

5.3. Estimation of the PoP of the First Class at a Point from a Series of the ADF approximations

As a model example, Fig. 5.3, [6] the classification problem with two classes with normal conditional distributions of class attributes in a two-dimensional space with different covariance matrices was considered.

The average values of the features of the first class $m_1 = (0,\ 0)$. The covariance matrix is $S_1 = ((1,0);(0,1))$. For the second class: $m_2 = (0,\ 2)$, $S_2 = ((4,1);(1,1))$. The prior probabilities of the classes are $P_1 = 0,6$, $P_2 = 0,4$.

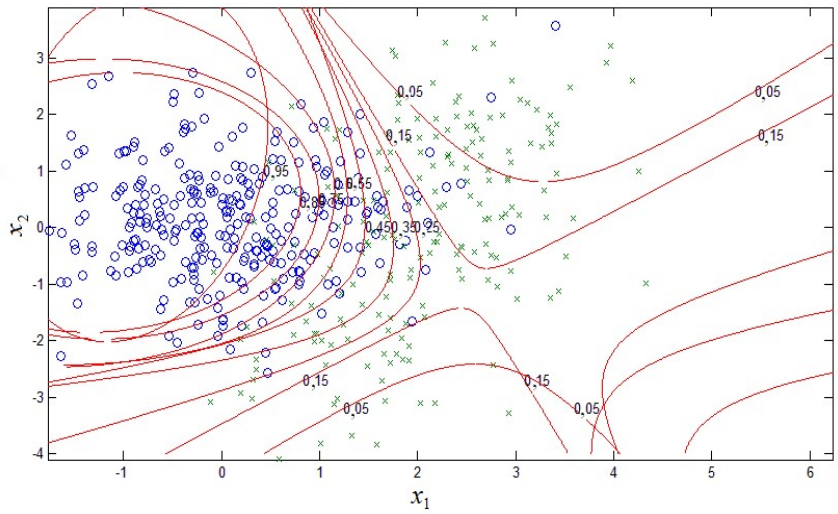

Fig. 5.3

Based on the initial data and the generated set of 500 points, the PoP of the first class was calculated in the following ways:

1. Exact - according to the known distributions according to the Bayes formula;
2. Sample - based on sample estimates of distribution parameters using Bayes' formula for calculating PoP of the first class;
3. Inter-extrapolation for the series of the ADF approximated by a second-order polynomial;
4. By the method of smoothing according to the series of ADF approximations.

A series of the ADF approximations is built for p^* values from 0.95 to 0.05 with a step of 0.1. Circles are points of the first class, crosses are of the second. The root-mean-square error estimate of the PoP of the first class for the training set by the interpolation method for a series of ADF approximations was 0.055. The error in its estimation by the smoothing method was 0.06. The error in estimating PoP of the first class based on sample estimates of feature distributions and using Bayes' formulas was 0.02, i.e. three times less. But in justification of

other methods, it should be remembered that they do not require knowledge of the laws of distribution of the features of classes.

6. CONCLUSION

Approximation of the Anderson discriminant function based on a supervised training set allows you to simultaneously find estimates of the posterior probabilities of classes of classified objects. The posterior probabilities of the classes are comprehensive information for solving the problem of classification according to various criteria.

The supervised training set of the classification problem is transformed into a training set of the regression analysis problem by replacing the class numbers in the training set with the corresponding differences with the costs of classification errors, set arbitrarily at the training stage to determine the approximations of the Anderson discriminant function. At the second stage of solving the problem, according to the estimates of the objective posterior probabilities of classes, the actual classification of the object is performed according to the subjectively specified classification criteria.

Methods for solving the problem related to the use of nonparametric and parametric approaches are proposed. With the nonparametric approach, to approximate the Anderson discriminant function at a point in the class feature space representing the classified object, a training set is used for each classified point, which may be unacceptable for larger set sizes and with certain performance requirements. For a non-parametric approach a rather linear approximating function.

In parametric approaches, weighted least squares are used to obtain a more accurate approximation of a function in the vicinity of its zero values. In another parametric approach, a series of approximations of the Anderson discriminant functions are constructed from a given set of a posteriori probabilities on the class boundaries. Then, for the classified point, the posterior probabilities of the classes are found using the interpolation method or the smoothing method.

Methods are demonstrated by examples.

REFERENCES

1. Platt, J. (2000). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, *Advances in large margin classifiers,* **10**(3), 61–74.

2. Anderson, T.W. (2003) *An Introduction to Multivariate Statistical Analysis*, *3 rd. ed*. New York, NY: John Wiley & Sons.

3. Hardle, V. (1993) *Applied nonparametric regression*. Moscow, Russia: Mir.

4. Zenkov, V.V. (2018). Estimating the probability of a class at a point by the approximation of one discriminant function, *Autom Remote Control,* **79**(9), 1582–1592, https://doi.org/10.1134/S0005117918090047.

5. Zenkov, V.V. (2017). Using Weighted Least Squares to approximate the discriminant function with a cylindrical surface in classification problems, *Autom. Remote Control,* **78**(9), 1662–1673, https://doi.org/10.1134/S0005117917090107.

6. Zenkov, V.V. (2019). Estimation of the posterior probability of a class from a series of Discriminant Anderson functions, *Autom Remote Control*, **80**(3), 447-458. https://doi.org/10.1134/S0005117919030056.

7. Zenkov, V.V. (2020). Applying an approximation of the Anderson Discriminant Function and Support Vector Machines for solving some classification tasks, *Autom Remote Control*, **81(**1), 118–129, https://doi.org/10.1134/S0005117920010105.

