A Metaheuristic for Classification of Interval Data in Changing Environments

Piotr Kulczycki and Piotr A. Kowalski

Abstract The Bayes approach is arguably the classification method most used in unspecialized applications, thanks to its robustness, simplicity, and interpretability. The main problem here is establishing proper probability values. This paper deals with adapting the above method for cases where the classified data is of interval type, with changing environments (evolving data stream, concept drift, nonstationarity). The probability values are estimated using nonparametric methods, thanks to which the procedure becomes independent of characteristics of learning subsets representing particular classes. They can also be supplemented with new, current observations, added while performing the algorithm. The investigated process also removes elements with negligible or even negative impact on accuracy of results, which increases the effectiveness of adaptation in conditions of changing reality. It is possible to differentiate the meanings of particular classes. The method allows any number of them. The particular attributes of data elements may be continuous, categorical, or both.

Keywords Data analysis · Classification · Interval data · Changing environment · Adaptation

1 Introduction

One of the main tasks of contemporary data analysis is classification [2, 5]. Suppose that we have a data set, whose particular elements are assigned labels explicitly, indicating membership of particular, previously defined subsets, constituting

P. Kulczycki (✉) · P.A. Kowalski
AGH University of Science and Technology, Faculty of Physics and Applied Computer Science, Kraków, Poland
e-mail: kulczycki@agh.edu.pl; kulczycki@ibspan.waw.pl

P.A. Kowalski
e-mail: pkowal@agh.edu.pl; pakowal@ibspan.waw.pl

P. Kulczycki · P.A. Kowalski
Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

specific classes. Such a label should be forecast for another element submitted for testing which does not already have one. This procedure of mapping a label suggesting membership to a class, to an investigated element is called a classifier.[1] If the concept of the classifier is based on a rough method, giving no strict guarantee of finding the best or even a correct solution, it can be categorized as heuristic [23], while if a few different concepts combine, where some act as servants to others, then it becomes metaheuristic. Finally, when computational intelligence methodology [11] is used, the data set mentioned at the beginning becomes a learning set. Its subsets assigned to particular classes are referred to as patterns.

This publication concerns the classification of data given in interval form [10], including also the multidimensional case. The fundamental benefit of this type of data is its simplicity, transparency, and possibility of using well-developed mathematical apparatus. Besides actual interval analysis, the case investigated here also includes a probabilistic approach with uniform distribution as well as fuzzy logic for a rectangular membership function. On the other hand in this publication, patterns consist of elements which are uniquely determined (including single-point distribution or crisp numbers for probabilistic and fuzzy approaches, respectively). This corresponds to many situations occurring in practice, for instance when patterns are formed from elements precisely measured some time ago (e.g., exchange rates, outside temperature), but the forecast, ambiguous in nature, is classified and presented in interval form [17].

Changeability in time of analyzing data is assumed here. Literature terms this a changing environment [21], occasionally also evolving data stream [3], concept drift [29], nonstationarity [19], or relates it with the adaptation process [4]. Such a problem is most commonly connected to permanent supplementation of a data set with new elements, which are naturally the most up to date and therefore the most valuable. In the methodology presented below, each of the patterns' element receives coefficients proportional to their influence on correct results. Those elements with smallest coefficients are removed, although an exception is made for those with successively growing values, as their character is in accordance with the trend of changes in the environment.

The metaheuristic proposed here will construct Bayes classifier [5], with a deservedly high opinion among researchers. It possesses a range of advantages, both theoretical (ensuring minimum expectation value of losses resulting from classification errors, albeit for incompletely fulfilled assumption of the attributes' independence) and practical (the idea is simple, robust, and being easy to interpret, is easy to modify). This method allows any number of classes and enables to differentiate their meaning from a practical perspective. The probability values existing in the classifier will be established by means of the nonparametric kernel estimators methodology [16]. Patterns can therefore be of any shape, including consisting of separate parts. Particular attributes of processed data may be

---

[1]Sometimes this procedure performs the function of reflecting reality with mathematics and information technology, which explains why it is occasionally called a model.

continuous, categorical, or a combination of both. It is worth noting that, thanks to the correctly chosen measure of similarity, it is possible to treat categorical variables as multivalued, including binary. The fixing and adaptation of estimators' parameters are carried out based on optimization procedures [12] and a sensitivity analysis known from the artificial neural networks technique [30].

The initial sections, Sects. 2–5, shortly present a theoretical basis applied later in the Sect. 6, the main section, to create the classification procedure for use in changing environments. Conclusions with numerical verification, followed by final comments, are the subject of Sect. 7.

The concept worked out here connects research for the interval stationary case with the deterministic nonstationary, which are accessible in the papers [18, 19], respectively. Initial results were described in the publication [20]. The specific aspects of using neural networks in the methodology proposed here are the subject of the articles [14, 15], currently in press.

2 Kernel Estimators

The nonparametric method of statistical kernel estimators enables the establishment of characteristics—mainly density of distribution—without any prior knowledge concerning its type. Thus, let an n-dimensional continuous random variable be given. Suppose that its distribution has a density, denoted by f. Having the random sample

$$x_1, x_2, \ldots, x_m \tag{1}$$

one can obtain its kernel estimator [16, 26, 28] defined as

$$\hat{f}(x) = \frac{1}{mh^n} \sum_{i=1}^{m} K\left(\frac{x - x_i}{h}\right), \tag{2}$$

whereas the function $K: \mathrm{R}^n \to [0, \infty)$, named a kernel, is measurable, symmetrical with respect to zero, has a weak global maximum at this point, and fulfills the condition $\int_{\mathrm{R}^n} K(x)\, \mathrm{d}x = 1$; the constant $h > 0$ is called a smoothing parameter.

The generalized one-dimensional Cauchy kernel

$$K(x) = \frac{2}{\pi (x^2 + 1)^2}, \tag{3}$$

will be used in the following. This type of kernel lends itself especially well to the classification problem, thanks to the presence of so-called "heavy tails", valuable in areas of potential division into particular classes, actually lying on peripheries of distributions associated with them. For the multidimensional case, the product approach will be used. The kernel is then defined as

$$K(x) = K\left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}\right) = K_1(x_1)\, K_2(x_2)\, \ldots\, K_n(x_n), \tag{4}$$

where K_1, K_2, ... , K_n represent one-dimensional kernels (3). Note that the expression h^n must be substituted in definition (2) by $h_1 \cdot h_2 \cdot\, \ldots\, \cdot h_n$, i.e., the product of smoothing parameters for consecutive coordinates. Observe also that thanks to the continuity of the kernel (3)–(4), the estimator $\hat{f}$ defined by equality (2) is also continuous.

Due to the planned correction in the smoothing parameter h, for calculation of its value the so-called simplified method is enough [16—Sect. 3.1.5; 28—Sect. 3.2.1]. In the one-dimensional case, as well as for particular coordinates in the multidimensional case, the smoothing parameter can be then calculated from a simple formula:

$$h = \left(\frac{W(K)}{U(K)^2} \frac{8\sqrt{\pi}}{3\,m}\right)^{1/5} \widehat{\sigma}, \tag{5}$$

while $W(K) = \int_{\mathrm{R}} K(x)^2\, \mathrm{d}x$, $U(K) = \int_{\mathrm{R}} x^2 K(x)\, \mathrm{d}x$, and $\widehat{\sigma}$ is an (one-dimensional) estimator of standard deviation obtained on the basis of sample (1). For the Cauchy kernel (3) one has $W(K) = 1$ and $U(K) = 5/4\pi$.

Kernel estimators are fully presented in the classic monographs [16, 26, 28], also including among others comments on the choice of kernel type [16—Sect. 3.1.3; 28—Sects. 2.7 and 4.5], algorithms for calculation of the smoothing parameter [16—Sect. 3.1.5; 28—Chap. 3 and Sect. 4.7], and additional concepts for fitting this type of estimator to specific conditions (e.g., boundary of random variable support) and procedures generally increasing its quality. In this latter group, it is worth highlighting the procedure for a smoothing parameter modification [16—Sect. 3.1.6; 26—Sect. 5.3.1], narrowing of particular kernels in dense areas (which enables better characterization of individual features of distribution), and also "flattening" them in sparse regions to additionally smooth the estimator on the peripheries ("tails") of distribution. The potential addition of this aspect to the material presented below is obvious and has been described in detail in the paper [19].

Kernel estimators can also be constructed for different than continuous types of attributes, in particular categorical (nominal and ordered), which through the appropriate selection of similarity measure offers a wide range of generalizations to multivalued variables, including binary. Various compositions of the above types are also possible. The explanations for this topic can be found in the publications [7, 22, 24]. The supplementation of this aspect to the considerations presented in this work is obvious.

3   Bayes Classification

The classification process consists of creating a decision rule, which will map to the tested element an additional label, demonstrating supposed membership to one of the earlier defined classes. These classes are represented by patterns, i.e., sets of elements already possessing such labels. At the beginning consider a continuous random variable. First, the one-dimensional case (relating to the previous section: $n = 1$) will be investigated. Consider therefore the tested quantity, given in the form of the interval

$$[\underline{x}, \bar{x}], \tag{6}$$

while $\underline{x} \leq \bar{x}$. Note that when $\underline{x} = \bar{x}$, it becomes precise (i.e., deterministic or sharp). Let also J classes of the sizes $m_1, m_2, \ldots, m_J$ be represented by patterns composed of real numbers:

$$x_1^1, x_2^1, \ldots, x_{m_1}^1 \tag{7}$$

$$x_1^2, x_2^2, \ldots, x_{m_2}^2 \tag{8}$$

$$\vdots$$

$$x_1^J, x_2^J, \ldots, x_{m_J}^J. \tag{9}$$

(Note that the upper index in the notations (7)–(9) denotes membership to a fixed class). Bayes classification consists of mapping the tested element (6) to the j-class ($j = 1, 2, \ldots, J$) if the largest is the j-th value among

$$m_1 f_1(\tilde{x}), m_2 f_2(\tilde{x}), \ldots, m_J f_J(\tilde{x}), \tag{10}$$

where $f_1, f_2, \ldots, f_J$ denote probability density with the condition of its membership to the class $1, 2, \ldots, J$, respectively. In the metaheuristic investigated here, these densities will be defined by kernel estimators methodology, described in Sect. 2, where successive patterns (7)–(9) will be used as samples (1). Suppose therefore such estimators of the above densities as $\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_J$. Then expressions (10) take the form

$$m_1 \hat{f}_1(\tilde{x}), m_2 \hat{f}_2(\tilde{x}), \ldots, m_J \hat{f}_J(\tilde{x}). \tag{11}$$

In turn for interval type of data, denoted in the form of element (6), one can conclude that it belongs to the j-class when the biggest is the j-th value from among

$$\frac{m_1}{\bar{x} - \underline{x}} \int_{\underline{x}}^{\bar{x}} \hat{f}_1(x)\, dx, \ \frac{m_2}{\bar{x} - \underline{x}} \int_{\underline{x}}^{\bar{x}} \hat{f}_2(x)\, dx, \ \ldots, \ \frac{m_J}{\bar{x} - \underline{x}} \int_{\underline{x}}^{\bar{x}} \hat{f}_J(x)\, dx. \tag{12}$$

If one uses the continuous kernel K, then formula (12) becomes the generalization of (11). In fact, here the kernel estimator $\hat{f}_j$ is also continuous, therefore for any fixed $\tilde{x} \in [\underline{x}, \overline{x}]$, if the length of interval (6) is reduced to 0 by $\underline{x} \to \tilde{x}$ and $\overline{x} \to \tilde{x}$, then one obtains

$$\lim_{\substack{\underline{x} \to \tilde{x} \\ \overline{x} \to \tilde{x}}} \frac{1}{\overline{x} - \underline{x}} \int_{\underline{x}}^{\overline{x}} \hat{f}_j(x) \, dx = \hat{f}_j(\tilde{x}) \quad \text{for } j = 1, 2, \ldots, J. \tag{13}$$

The expressions (12) transform into (11).

Furthermore, the positive expression $1/(\overline{x} - \underline{x})$ can be removed as having no influence on which factor in formula (12) is the largest. Then it becomes equivalent to

$$m_1 \int_{\underline{x}}^{\overline{x}} \hat{f}_1(x) \, dx, \ m_2 \int_{\underline{x}}^{\overline{x}} \hat{f}_2(x) \, dx, \ \ldots, \ m_J \int_{\underline{x}}^{\overline{x}} \hat{f}_J(x) \, dx. \tag{14}$$

Moreover, for every $j = 1, 2, \ldots, J$ we have

$$\int_{\underline{x}}^{\overline{x}} \hat{f}(x) \, dx = \widehat{F}(\overline{x}) - \widehat{F}(\underline{x}) \tag{15}$$

with

$$\widehat{F}(x) = \int_{-\infty}^{x} \hat{f}(y) \, dy. \tag{16}$$

Substituting to the above dependency the definition for kernel estimator (2) (for $n = 1$) with Cauchy kernel (3) and removing once again the positive constant $1/m\pi$ irrelevant here, one can obtain the following analytical formula:

$$\widehat{F}(x) = \sum_{i=1}^{m} \left[\frac{(x^2 - 2xx_i + x_i^2 + h^2) \arctan\left(\frac{x-x_i}{h}\right) + h(x - x_i)}{x^2 - 2xx_i + x_i^2 + h^2} + \frac{\pi}{2} \right]. \tag{17}$$

In summary: the tested element (6) should be mapped to the j-class ($j = 1, 2, \ldots, J$) if the j-th value is the largest from expressions (14). The integrals appearing there can be calculated using formula (15) with substitution of dependence (17). This completes the classification algorithm in the one-dimensional case.

Now consider the multidimensional case, i.e., $n > 1$, when the interval vector

$$\begin{bmatrix} [\underline{x_1}, \overline{x_1}] \\ [\underline{x_2}, \overline{x_2}] \\ \vdots \\ [\underline{x_n}, \overline{x_n}] \end{bmatrix} \tag{18}$$

is tested, while elements of patterns (7)–(9) belong to the space R^n. Then expressions (14) are

$$m_1 \int_E \hat{f}_1(x)\,dx, \; m_2 \int_E \hat{f}_2(x)\,dx, \; \ldots, \; m_J \int_E \hat{f}_J(x)\,dx, \tag{19}$$

where $E = [\underline{x_1}, \overline{x_1}] \times [\underline{x_2}, \overline{x_2}] \times \cdots \times [\underline{x_n}, \overline{x_n}]$. To calculate the above integrals, observe that for the product kernel (4), the following is true:

$$\int_E K(x)\,dx = [I_1(\overline{x_1}) - I_1(\underline{x_1})][I_2(\overline{x_2}) - I_2(\underline{x_2})] \; \ldots \; [I_n(\overline{x_n}) - I_n(\underline{x_n})], \tag{20}$$

where I_i means the primitive function of the one-dimensional kernel K_i for $i = 1, 2, \ldots, n$. Equalities (15) and (17) provide analytical formulas for obtaining the values of these integrals, which completes the procedure for classification of interval data in the continuous random variable case.

The above material can be easily transposed from continuous to categorical variables. Here, an interval element should be understood to be the set sum of several categories. In this situation, testing an element of such type, one should add the kernel estimators values for all categories belonging to the created sum (or their combinations if there are a number of categorical attributes), and then apply criterion (11). The procedure is similar for a combination of continuous and categorical attributes: for fixed categories belonging to the set one should—using the above-presented methodology—calculate kernel estimator values for continuous attributes, add them, and finally apply criterion (11).

Finally, generalize expressions existing in (11) and (19), introducing the coefficients $z_1, z_2, \ldots, z_J > 0$ in the following manner:

$$z_1 m_1 \int_{\underline{x}}^{\overline{x}} \hat{f}_1(x)\,dx, \; z_2 m_2 \int_{\underline{x}}^{\overline{x}} \hat{f}_2(x)\,dx, \; \ldots, \; z_J m_J \int_{\underline{x}}^{\overline{x}} \hat{f}_J(x)\,dx \tag{21}$$

$$z_1 m_1 \int_E \hat{f}_1(x)\,dx, \; z_2 m_2 \int_E \hat{f}_2(x)\,dx, \; \ldots, \; z_J m_J \int_E \hat{f}_J(x)\,dx, \tag{22}$$

respectively. Taking as standard values $z_1 = z_2 = \ldots = z_J = 1$, formula (21) brings us to (14), and (22) to (19). By appropriately changing the value z_i, one can appropriately influence the probability of assigning elements from the i-th class to other wrong classes, although potentially at the cost of increasing the total number of misclassifications. This concept can be applied in such situations where particular classes are associated with phenomena of different significance to the investigated task, or diverse conditioning. In the case of changing environments, moving patterns represent a much more difficult scenario. They may contain elements which are no longer current, or have already appeared, but will only become typical in the future. The adaptation procedure for such patterns is significantly less efficient than for unchanging patterns, where instead of the necessity for updating they can be successively improved by removing less effective elements. In the presented problem, the coefficient z_i values should be, respectively, proportional to the speed of changes of the i-th classes. The value, 1.25 can be proposed as initial; generally for the most applicational tasks $z_1, z_2, \ldots, z_J \in [1, 1.5]$.

Bayes classification is highly regarded among practitioners. It is uncomplicated, easily interpretable, and often provides results better than many more refined procedures. Together with kernel estimators, with a very small value of the smoothing parameter, it is reminiscent of the nearest neighbor algorithm, whereas when it is large, it is similar to average (mean) linkage. Thanks to the proper choice of the smoothing parameter, it seems possible to obtain better results than in the case of those two effective methods. Within the proposed metaheuristic, this aspect is reflected in the optimal correction of the above parameter, presented in the next section.

More details concerning Bayes classification is included in the publications [1, 5]; see also [9, 13]. A somewhat broader presentation of the material of the above section can be found in the paper [18].

4   Correction for Smoothing Parameters

With the aim of improving quality of results as well as creating the possibility of keeping up with environment changes, the metaheuristic investigated here applies a correction procedure to the smoothing parameters values, using optimizing algorithms, suiting the value (5) to the classification problem.

Thus, suppose n correcting coefficients $b_1, b_2, \ldots, b_n > 0$, which will be used to multiply the particular smoothing parameters $h_1, h_2, \ldots, h_n$ calculated using formula (5), respectively. Note that the case $b_1 = b_2 = \ldots = b_n = 1$ means a lack of correction. Assume the natural performance index

$$J(b_1, b_2, \ldots, b_n) = \#\{\text{incorrect classifications}\}, \tag{23}$$

where # denotes here the number of elements, and the task of minimization of its value. First, on the grid created for the values $b_j = 0.25, 0.5, \ldots, 1.75$ for every

coordinate $j = 1, 2, \ldots, n$, one should calculate the values of the above index, and then choose the best five. Next, treating these points as initial, static optimization methods in the space R^n ought to be used. The value of index (23) can be calculated by the classic leave-one-out method. Due to these values being integers, a modified Hook–Jeeves procedure [12], with initial step taken as 0.2, was applied. Other conceptions are described in the survey paper [27]. After finishing the above five "runs" of the Hook–Jeeves procedure, one should select one of these values of the correcting coefficients $b_1, b_2, \ldots, b_n$ for which functional (23) value for the end point is the smallest.

However, the above-presented correction of the smoothing parameters procedure is not necessary, it increases classification accuracy, enhances adaptation, and furthermore enables the use of a simplified method for calculating smoothing parameters values (5), based on the square criterion, which is not always beneficial to the classification task [8]. Its influence could have particular significance in abrupt or atypical changes of environment. When applying the modification procedure for the smoothing parameter (see the penultimate paragraph of Sect. 2), the above action undergoes moderate generalization in accordance with the concept described in the paper [19].

5 Pattern Size Reduction

In practical tasks, several elements of patterns (7)–(9) might be unimportant, and in some cases may even have negative influence for classification quality. Their proper selection and removal can improve the correctness of results, and also—thanks to a reduction in pattern sizes—significantly accelerate calculations. To this end, we shall generalize the definition of kernel estimator (2) to the following form:

$$\hat{f}(x) = \frac{1}{mh^n} \sum_{i=1}^{m} w_i K\left(\frac{x - x_i}{h}\right),$$
(24)

where the coefficients $w_1, w_2, \ldots, w_m \geq 0$ introduced above are normed such that

$$\sum_{i=1}^{m} w_i = m.$$
(25)

In the special case $w_i \equiv 1$, formula (24) reduces to its initial definition (2). The parameters w_i are intended to characterize the influence of the respective i-th elements of the patterns on the accuracy of results. In order to calculate their values, the sensitivity analysis, familiar from the theory of artificial neural networks [6, 30], will be applied. Its aim is to define—after the learning phase—the influence of the particular inputs u_i of a neural network on its output value y, described in the natural way by the quantity

$$S_i = \frac{\partial y(x_1, x_2, \ldots, x_m)}{\partial x_i} \text{ for } i = 1, 2, \ldots, m, \tag{26}$$

and then to aggregate information in the form of the coefficients

$$\overline{S}_i = \sqrt{\frac{\sum_{p=1}^{P} (S_i^{(p)})^2}{P}} \text{ for } i = 1, 2, \ldots, m, \tag{27}$$

where $S_i^{(p)}$ with $p = 1, 2, \ldots, P$ denotes the value (26) for particular iterations. A detailed description of the sensitivity method, together with the appropriate formulas, is presented in the publications [6, 30]. The configuration of neural networks and specific aspects associated with this topic are presented in the separate papers [14, 15]. To every class characterized by patterns (7)–(9) an individual network is assigned. For the sake of simplified notation, the index $j = 1, 2, \ldots, J$ of particular classes will be fixed hereinafter.

In order to define the values of the parameters introduced in definition (24), first calculate auxiliary quantities

$$\tilde{w}_i = \left(1 - \frac{\overline{S}_i}{\sum_{j=1}^{m} \overline{S}_j} \right), \tag{28}$$

finally normed—in consideration of condition (25)—to

$$w_i = m \frac{\tilde{w}_i}{\sum_{i=1}^{m} \tilde{w}_i}. \tag{29}$$

The concept of the above formulas stems from the fact that neural networks are most sensitive to redundant and atypical elements which, from a classification point of view, are mainly of negative significance, therefore they receive the values $\tilde{w}_i$ and in consequence w_i should be proportionately small. Note also that due to the shape of formulas (26)–(27), in practice not all coefficients $\overline{S}_i$ are equal to zero, which guarantees the nominator in dependence (28) is not equal to zero.

Finally, those elements of patterns (7)–(9) for which $w_i < 1$ are removed. The limit value 1 results from the fact that, thanks to the form of normalization (29), the arithmetic mean of parameters equals 1. Empirical research carried out confirmed this theoretically conditioned point of view [14, 15].

6 Classification Metaheuristic

This crucial section collates the material presented in this paper. Procedures presented earlier in Sects. 2–5, will be joined in the classifying metaheuristic designed for the changing environment case. An illustration is provided in Fig. 1. Blocks drawn with a continuous line denote operations performed on all elements of patterns, with a dashed line—on particular classes, while a dotted line symbolizes operations for each element of those patterns.

To start, one should fix the so-called reference sizes of patterns (7)–(9), denoted hereinafter as m_1^*, m_2^*, $\ldots, m_J^*$. They are the sizes of patterns defined during the reduction procedure presented in Sect. 5. Of course, initial patterns must be of a size no smaller than the reference ones. These values may be changed, with the natural boundary that their increase cannot be smaller than the amount of new elements. To begin one can propose $m_1^* = m_2^* = \cdots = m_J^* = 25 \cdot 2^n$. Greater values may cause an increase in calculation time, while smaller a drop in accuracy of results.

Initial patterns (7)–(9) constitute preliminary data submitted for investigated procedure. First, the values of the smoothing parameters h_1, h_2, $\ldots$, h_n are calculated according to the material of Sect. 2. This action is denoted in Fig. 1 as block A. The subsequent block B symbolizes computation for the coefficients b_1, b_2, $\ldots$, b_n values, realizing a correction of the smoothing parameters, worked out in Sect. 4.

The next step, described in Sect. 5 (block C in Fig. 1), consists of the calculation of the parameters w_i values, carried out separately for particular classes. After that, these parameters are sorted within each class (block D in Fig. 1). Any sorting procedure [25] can be used here. Following this, shown in Fig. 1 as block E, the m_1^*, m_2^*, $\ldots, m_J^*$ elements corresponding to the largest values w_i are the basis of the principal phase of the investigated procedure—Bayes classification (block F in Fig. 1), which will be discussed in the subsequent paragraph. On the other hand, elements corresponding to smaller values w_i are sent to block U, during which the derivative w_i' is calculated individually for each of them. Newton's interpolation polynomial for the last three observations can be proposed here; its description, together with formulas as well as similar methods are presented in the survey paper [27]. (If for some element, three previous values w_i are not available, then they can be filled with zeroes, artificially increasing a derivative, while at the same time securing such elements against premature removal.) Later the values w_i' are sorted separately for specific classes (block V in Fig. 1), after which—within block W— elements of each pattern in the number

$$q m_1^*, \ q m_2^*, \ \ldots, q m_J^*, \tag{30}$$

respectively, with the largest positive derivative values, return to block A at the beginning. The leftover elements are finally removed, as is shown in block Z.
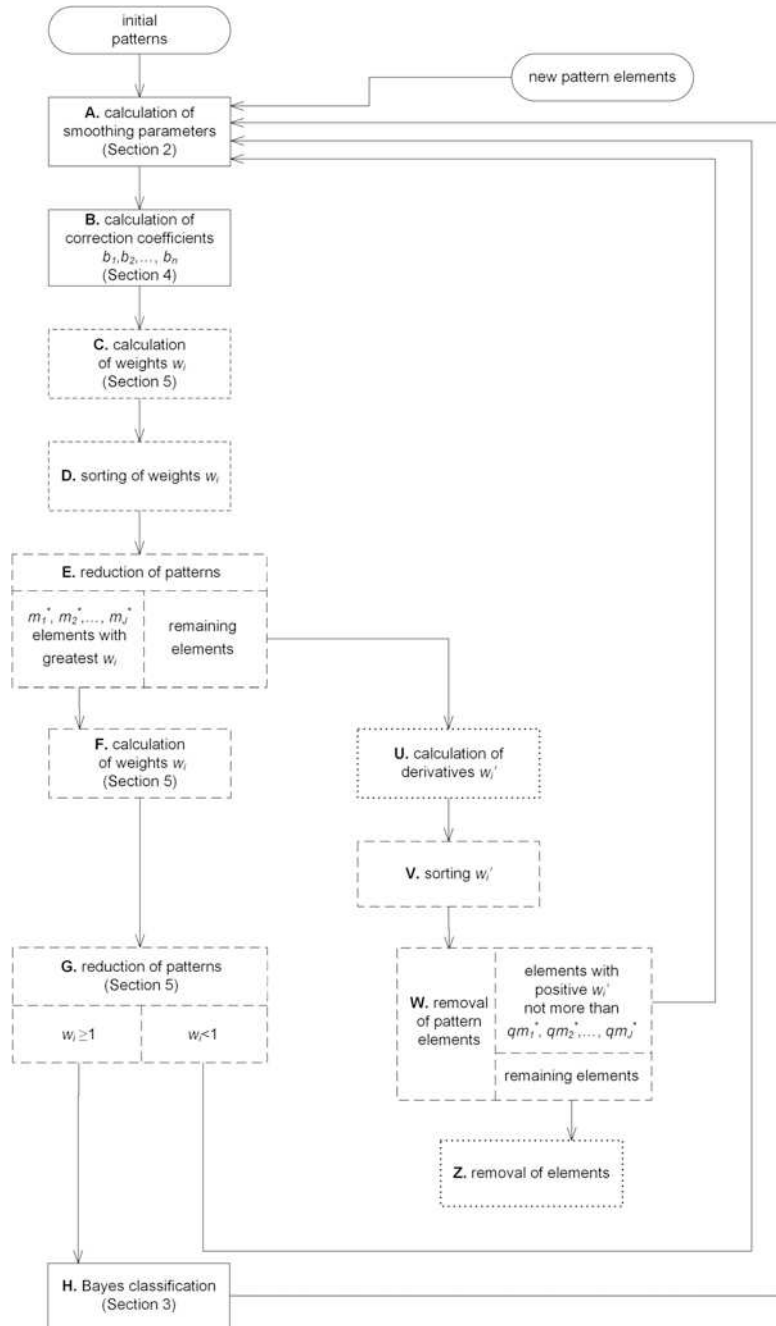
Fig. 1 Classification metaheuristic

The positive parameter q introduced above in formula (30) implies the part played in further tests of elements with small, but successively increasing significance, therefore preceding trends of environment changes, as it were. The initial value $q = 0.2$ is proposed; generally $q \in [0.1, 0.25]$ depending on intensity and uniformity of changes. Bigger values may improve the adaptation process but lengthen calculation time, while smaller ones bring contrary effects.

Let us return to Bayes classification, the essence of the procedure presented here. As mentioned at the top of the previous paragraph, this stage sees the arrival of those patterns' elements which have the greatest influence on accurate results. First the parameters' w_i values are once more calculated, in accordance with Sect. 5 (block F in Fig. 1). Then within block G those elements for which $w_i < 1$ are excluded from further processing and sent at the beginning to block A, while those with $w_i \geq 1$ are prescribed to block H, where they form the basis for Bayes classification, described in Sect. 3 (block H in Fig. 1). Testing can be performed on many interval data of type (6) or (18). Next all patterns' elements join block A at the beginning.

The presented procedure can be repeated as soon as new elements are provided to block A. In addition, there are also applied the previously used $m_1^*, m_2^*, \ldots, m_J^*$ elements with the largest values w_i as the most valuable for accuracy of results, as well as approximately $qm_1^*, qm_2^*, \ldots, qm_J^*$ ones having the greatest positive derivative w_i', as not having yet big influence but successively increasing their significance as the environment changes.

The expanded description of the procedure presented above can be found in the paper [19].

7 Verification and Final Comments

The correctness of the method described in this paper underwent comprehensive numerical verification. In particular, it was shown that the classification developed here offers correct results also in cases of nonseparated classes with composite multisegment and multimodal patterns. The character of changing environment may increase successively, abruptly, or also periodically, although the best results are found in the first case. The standard values proposed in this text for the parameters used were obtained as deductions from simulations carried out.

The results differed little in nature from those obtained in the basic case where an element which is uniquely defined, e.g., deterministic or crisp, undergoes testing. It proves proper averaging introduced by formulas (14) and (19).

As an example, presented in Fig. 2, let us consider the illustratory two-dimensional case with two classes, one of which is invariable, with the other also unchanging at the beginning, after the 18th step it starts to change its place, and then—after describing a full orbit around the first class—stops in the 54th step at its initial location. The remaining parameters are accepted in the form proposed above

Fig. 2 Number of
misclassifications at particular
steps of the representative run
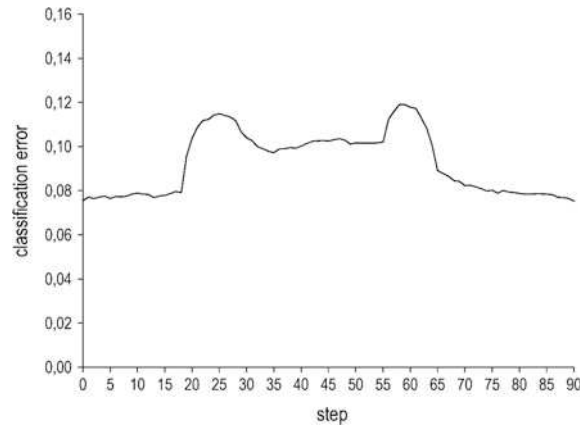
in this text. One can see in Fig. 2 that the number of misclassifications increases
sharply at times when the environment changes its character, i.e., in steps 18 and
54. The prediction function is then ineffective by nature. In the periods of non-
stationarity, i.e., before the 18th and after the 54th step, the rate of errors stabilizes
at a value of 0.08, whereas in the period of constant changes between the 18th and
54th steps, at the higher 0.105. This is still lower than the maximums values 0.12,
which would be maintained without the influence of the adaptation function
designed here.

Further research was undertaken on the influence of size of imprecision of
classified data—represented by the length of intervals—on accuracy of results. In
this aspect also the effects showed themselves to be fully satisfactory. If the interval
length was less than the generally understood distance between centers of specific
patterns (a condition usually fulfilled in practice), then its growth did not cause an
increase in the mean value of incorrect classifications, but in fact the results
underwent some stabilization—the variance of misclassifications decreased. Again
averaging, introduced by formulas (14) and (19), proves to have a positive
influence.

A broader description of particular aspects of the above simulations can be found
in the papers [14, 15, 18, 19].

The metaheuristic proposed in this paper was compared with other classification
methods based on computational intelligence, e.g., Support Vector Machine, as
well as natural, e.g., counting components of patterns which are included in the
tested element. Unfortunately, no method has been found to allow exactly the same
conditionings: uniquely defined patterns elements, interval form of tested element,
changing environment, any number of classes and patterns shapes, categorical
attributes. For this reason, it was possible only to compare with simplifications
fitting suitable methodologies, and so offer the results presented below purely in a
qualitative aspect. The advantage of the metaheuristic proposed in this paper mainly
lies in the smaller number of misclassifications for stabilized variability of envi-
ronment, which in Fig. 2 appears as a significant decrease in errors between 30 and

55 steps. Better results are also achieved here in areas between particular patterns, which are always troublesome for classification, as well as for long intervals representing specific attributes of tested elements. Thanks to the calculational complexity of particular procedures of the metaheuristic under investigation, the proposed method is especially destined for those cases where slow learning is permitted, but the classification process itself must be fast. This is achieved in great part by obtaining an analytical form of formulas (15)–(17). The computational complexity of the classification phase alone amounts to $O(nJ\,m)$, and therefore is linear with respect to dimensionality of space, number of classes, and size of their patterns.

References

1. Aggarwal, C.C.: Data classification: algorithms and applications. Chapman & Hall/CRC, London (2014)
2. Aggarwal, C.C.: Data mining. The textbook. Springer, Cham (2015)
3. Aggarwal, C.C., Han, J., Wang, J., Yu, P.S.: A framework for on-demand classification of evolving data streams. IEEE Trans. Knowl. Data Eng. **18**, 577–589 (2006)
4. Bouchachia, A.: Adaptation in classification systems. In: Hassanien, A.E., Abraham, A., Herrera, F. (eds.) Foundations of Computational Intelligence, vol. 2, pp. 237–258. Springer, Berlin (2009)
5. Duda, R.O., Hart, P.E., Storck, D.G.: Pattern classification. Wiley, New York (2001)
6. Engelbrecht, A.P., Cloete, I., Zurada, J.: Determining the significance of input parameters using sensitivity analysis. In: Mira, J., Sandoval F. (eds.) From Natural to Artificial Neural Computation. Lecture Notes in Computer Science, pp. 382–388. Springer, Berlin (1995)
7. Gaosheng, J., Rui, L., Zhongwen, L.: Nonparametric estimation of multivariate CDF with categorical and continuous data. Adv. Econom. **25**, 291–318 (2009)
8. Ghosh, A.K., Chaudhuri, P., Sengupta, D.: Classification using kernel density estimation: multiscale analysis and visualization. Technometrics **48**, 120–132 (2006)
9. Hryniewicz, O., Kaczmarek, K., Nowak, P.: Bayes statistical decisions with random fuzzy data—an application for the Weibull distribution. Maint. Reliab. **17**, 610–616 (2015)
10. Jaulin, L., Kieffer, M., Didrit, O., Walter, E.: Applied interval analysis. Springer, Berlin (2001)
11. Kacprzyk, J., Pedrycz, W. (eds.): Springer handbook of computational intelligence. Springer, Dordrecht (2015)
12. Kelley, C.T.: Iterative methods for optimization. SIAM, Philadelphia (1999)
13. Kobos, M., Mandziuk, J.: Multiple-resolution classification with combination of density estimators. Connect. Sci. **23**, 219–237 (2011)
14. Kowalski, P.A., Kulczycki, P.: A complete algorithm for the reduction of pattern data in the classification of interval information. Int. J. Comput. Methods. **13**(1650018) (2016)
15. Kowalski, P.A., Kulczycki, P.: Interval probabilistic neural network. Neural Comput. Appl. (2017, in press)
16. Kulczycki, P.: Estymatory jadrowe w analizie systemowej. WNT, Warsaw (2005)
17. Kulczycki, P., Hryniewicz, O., Kacprzyk, J. (eds.): Techniki informacyjne w badaniach systemowych. WNT, Warsaw (2007)
18. Kulczycki, P., Kowalski, P.A.: Bayes classification of imprecise information of interval type. Control Cybern. **40**, 101–123 (2011)
19. Kulczycki, P., Kowalski, P.A.: Bayes classification for nonstationary patterns. Int. J. Comput. Methods **12**(1550008) (19 pages) (2015a)

20. Kulczycki, P., Kowalski, P.A.: Classification of interval information with data drift. In: Christiansen, H., Stojanovic, I., Papadopoulos, G.A. (eds.) Modeling and Using Context. Lecture Notes in Computer Science, pp. 495–500. Springer, Berlin (2015b)
21. Kuncheva, L.I.: Classifier ensembles for changing environments. In: Roli, F., Kittler, J., Windeatt, T. (eds.) Multiple Classifier Systems. Lecture Notes in Computer Science, pp. 1–15. Springer, Berlin (2004)
22. Li, Q., Racine, J.S.: Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data. J. Bus. Econ. Stat. **26**, 423–434 (2008)
23. Michalewicz, Z., Fogel, D.B.: How to Solve It: Modern Heuristics. Springer, New York (2004)
24. Ouyang, D., Li, Q., Racine, J.: Cross-validation and the estimation of probability distributions with categorical data. J. Nonparametric Stat. **18**, 69–100 (2006)
25. Sedgewick, R., Wayne, K.: Algorithms. Addison-Wesley, Upper Saddle River (2011)
26. Silverman, B.W.: Density Estimation for Statistics and Data Analysis. Chapman and Hall, London (1986)
27. Venter, G.: Review of optimization techniques. Encyclopedia of Aerospace Engineering, pp. 5229–5238. Wiley, New York (2010)
28. Wand, M.P., Jones, M.C.: Kernel Smoothing. Chapman and Hall, London (1995)
29. Zlobaite, I.: *Learning under Concept Drift: an Overview*, Technical report, Faculty of Mathematics and Informatics, Vilnius University (2009)
30. Zurada, J.: Introduction to Artificial Neural Network Systems. West Publishing, St. Paul (1992)