Control and Cybernetics

vol. **40** (2011) No. 1

Bayes classification of imprecise information of interval type*

by

Piotr Kulczycki and Piotr Andrzej Kowalski

Systems Research Institute, Polish Academy of Sciences
Newelska 6, 01-447 Warszawa, Poland
e-mail: {kulczycki, pakowal}@ibspan.waw.pl

Cracow University of Technology
Department of Automatic Control and Information Technology
Warszawska 24, 01-447 Warszawa, Poland
e-mail: {kulczycki, pkowal}@pk.edu.pl

Abstract: The subject of the investigation presented here is Bayes classification of imprecise multidimensional information of interval type by means of patterns defined through precise data, e.g. deterministic or sharp. For this purpose the statistical kernel estimators methodology was applied, which makes the resulting algorithm independent of the pattern shape. In addition, elements of pattern sets which have insignificant or negative influence on the correctness of classification are eliminated. The concept for realizing the procedure is based on the sensitivity method, used in the domain of artificial neural networks. As a result of this procedure the number of correct classifications and – above all – calculation speed increased significantly. A further growth in quality of classification was achieved with an algorithm for the correction of classifier parameter values. The results of numerical verification, carried out on pseudorandom and benchmark data, as well as a comparative analysis with other methods of similar conditioning, have validated the concept presented here and its positive features.

Keywords: data analysis, classification, imprecise information, interval type information, statistical kernel estimators, reduction in pattern size, classifier parameter correction, sensitivity method for artificial neural networks.

1. Introduction

The current dynamic development in computer technology offers a continuous increase in both capability and speed of contemporary calculation systems, thus

*Submitted: July 2010; Accepted: February 2011.

allowing ever more frequent use of methods which up to now have only been applied to a relatively limited extent. One of the domains of these methods is the analysis of information which is imprecise in various – depending on the conditioning of a problem – forms, for example uncertain (statistical methods, e.g. Gil and Hryniewicz, 2009; Rice, 1994) or fuzzy (fuzzy logic, e.g. Kacprzyk, 1997; Klir and Yuan, 1995).

Recently, many applications showed an increase in the use of interval analysis. The basis for this concept is the assumption that the only available information on an investigated quantity is the fact that it fulfills the condition $\underline{x} \leq x \leq \bar{x}$, and in consequence this quantity can be associated with the interval

$$[\underline{x}, \overline{x}]. \tag{1}$$

Interval analysis is a separate mathematical domain, with its own formal apparatus based on an axiom of the sets theory (Moore, 1966).

A fundamental application of interval analysis was to ensure the required precision of numerical calculations, through monitoring errors arising from rounding numbers (Alefeld and Hercberger, 1986), however as a result of its continuous development, this field is finding ever wider uses in engineering, econometrics and other related areas (Jaulin et al., 2001). Its main advantage is the fact that by definition it models imprecision of a studied quantity, using the simplest possible formula. In many applications interval analysis shows to be absolutely sufficient, yet does not require many calculations (thus enabling its application in highly complex tasks) and is easy to follow and interpret, while also maintaining a formalism stemming from a convenient mathematical tool. Moreover, it can be noted that its concept is related to statistical interval estimation, and analysis of fuzzy numbers with rectangular membership functions.

Dynamic development is also currently taking place in information technologies in the area of data analysis and exploration (Kulczycki et al., 2007; Kumar and Tinku, 2004). This is due not only to an increase in the possibilities of the methodology used here, but above all to an increase in the accessibility of its algorithms, up to now a domain only available to a relatively small group of specialists. Among the fundamental tasks of data analysis and exploration lies that of classification (Hand, 1997). It consists of assigning a tested element to one of several previously selected groups. They are most often given by patterns, which are sets of elements representative for particular classes. This means that in many problems – including those, where data containing imprecision are investigated – elements forming patterns are defined precisely (e.g. deterministic in probability approach, sharp for the case of fuzzy logic, or in relation to notation (1) fulfilling the equality $\underline{x} = \overline{x}$).

This paper offers a complete procedure for classification of imprecise information represented by the interval vector

$$\begin{bmatrix} [\underline{x_1}, \overline{x_1}] \\ [\underline{x_2}, \overline{x_2}] \\ \vdots \\ [\underline{x_n}, \overline{x_n}] \end{bmatrix}, \tag{2}$$

where $\underline{x_k} \leq \overline{x_k}$ for $k = 1, 2, ..., n$, when the patterns of particular classes are given as sets of precise data (e.g. deterministic or sharp) elements, i.e. with $\underline{x_k} = \overline{x_k}$ ($k = 1, 2, ..., n$). The classification concept is based on the Bayes approach, ensuring the minimum of potential losses occurring through classification errors. For such a formulated task the statistical kernel estimators methodology was employed, thereby freeing the above procedure from arbitrary assumptions regarding pattern forms – their identification becomes an integral part of the presented algorithm. A procedure was also developed for reducing the size of pattern sets by removing the elements having negligible or negative influence on correctness of classification. Its concept is founded on the sensitivity method, used in the domain of artificial neural networks, although the intention is to increase the number of accurate classifications and – above all – the calculation speed. Furthermore, a method was designed to ensure additional improvements in classification results, obtained by correcting the values of classifier parameters. The validity and effectiveness of the algorithms used have been examined numerically. A comparison of results obtained was also carried out against other existing methods under analogous conditions.

The preliminary version of this paper was presented in Kulczycki and Kowalski (2008).

2. Preliminaries

2.1. Statistical kernel estimators

Kernel estimators belong to the group of nonparametric statistical methods. They allow for the calculation and clear illustration of characteristics of a random variable distribution, without knowledge of its membership in a given class.

Consider an n-dimensional random variable X with a distribution characterized by the density f. Its kernel estimator $\hat{f} : \mathbb{R}^n \to [0, \infty)$ is calculated on the basis of the random sample $\{x_i\}_{i=1,2,...,m}$ of size m, and defined – in the basic form – by the formula

$$\hat{f}(x) = \frac{1}{mh^n} \sum_{i=1}^{m} K\left(\frac{x - x_i}{h}\right), \tag{3}$$

where the positive coefficient h is a smoothing parameter, while the measurable function $K : \mathbb{R}^n \to [0, \infty)$, symmetrical with respect to zero, having at this point weak global maximum and fulfilling the condition $\int_{\mathbb{R}^n} K(x) \, \mathrm{d}x = 1$, is termed a kernel.

The interpretation of the above definition is illustrated in Fig. 1 for a one-dimensional random variable ($n = 1$) and an 8-element sample ($m = 8$). In the case of a single realization x_i, the function K (transposed along the vector x_i and scaled by the coefficient h) represents the approximation of distribution of the random variable upon obtaining the value x_i. For m independent realizations, this approximation takes the form of a sum of these single approximations. The constant $1/mh^n$ enables the fulfillment of the condition $\int_{\mathbb{R}^n} \hat{f}(x)\,\mathrm{d}x = 1$, required of the density of a probability distribution.

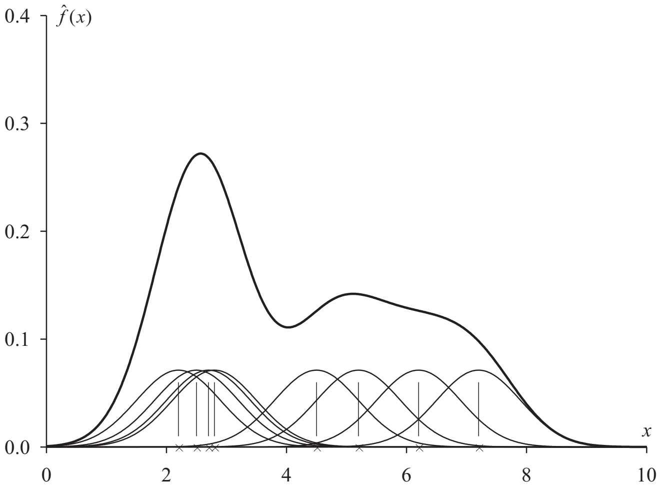

Figure 1. Kernel estimator (3)

The choice of the form of the kernel K and of the value for the smoothing parameter h is most often made based on the criterion of minimization of integrated square error (Kulczycki, 2005; Silverman, 1986; Wand and Jones, 1994).

Thus, the form of the kernel K has practically no influence on the statistical quality of estimation. In this paper the generalized (one-dimensional) Cauchy kernel is applied:

$$K(x) = \frac{2}{\pi\,(x^2+1)^2}, \tag{4}$$

in the multidimensional case defined using the product kernel concept

$$K(x) = K\left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}\right) = \mathcal{K}(x_1)\cdot\mathcal{K}(x_2)\cdot\ldots\cdot\mathcal{K}(x_n), \tag{5}$$

where $\mathcal{K}$ denotes the one-dimensional kernel given by formula (4).

The value of the smoothing parameter h can be calculated in practice with the tried and tested algorithms available in literature. The effective and convenient plug-in method (Kulczycki, 2005 – Section 3.1.5; Wand and Jones, 1994 – Section 3.6.1) is recommended here. In the multidimensional case, regarding application of the product kernel in this paper, the smoothing parameter will be naturally denoted as h_1, h_2, ..., h_n respectively for subsequent coordinates, and can be obtained separately for each of them by the above suggested method.

In practice, one employs additional procedures to increase generally the quality of the kernel estimator, and also fit its features to those of the considered reality. In this paper the modification of the smoothing parameter (Kulczycki, 2005 – Section 3.1.6; Silverman, 1986 – Section 5.3.1) will be applied, significantly improving the properties of the kernel estimator, particularly in areas where it assumes small values. In classification tasks this takes place especially near the boundaries of specific classes, which makes this procedure particularly useful here. Consider, then, the nonnegative modifying coefficients

$$s_i = \left(\frac{\hat{f}_*(x_i)}{\bar{s}} \right)^{-c} \qquad \text{for } i = 1, 2 \ldots, m, \tag{6}$$

where the constant $c \geq 0$ is called modification intensity, $\hat{f}_*$ denotes the kernel estimator in its basic form (3), and $\bar{s}$ – the geometrical mean of the quantities $\hat{f}_*(x_i)$ with $i = 1,\ 2\ \ldots,\ m$. The final definition of estimator (3) with product kernel (5) then takes the form:

$$\hat{f}(x) = \frac{1}{m\, h_1\, h_2 ... \, h_n} \sum_{i=1}^{m} \frac{1}{s_i^n} \mathcal{K}\left(\frac{x_1 - x_{i,1}}{h_1 s_i} \right) \mathcal{K}\left(\frac{x_2 - x_{i,2}}{h_2 s_i} \right) ... \mathcal{K}\left(\frac{x_n - x_{i,n}}{h_n s_i} \right), \tag{7}$$

where the natural notations

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad x_i = \begin{bmatrix} x_{i,1} \\ x_{i,2} \\ \vdots \\ x_{i,n} \end{bmatrix} \qquad \text{for } i = 1,\ 2,\ ...,\ m \tag{8}$$

are used, and together with formula (4) will be employed later on in this paper. The case of $c = 0$, entailing $s_i \equiv 1$, implies the lack of smoothing parameter modification, while with an increase of the parameter c the intensity of this procedure grows. Corollaries resulting from the mean-square criterion primarily point to the value

$$c = 0.5\,. \tag{9}$$

Fig. 2 shows an interpretation of the above procedure. In the areas where elements of the random sample are dense, for the elements x_i it is true that

$\hat{f}_*(x_i) > \bar{s}$, and therefore, as a result of formula (6), also $s_i < 1$. This leads to a narrowing of the kernels assigned to them, which in turn allows for better characterization of specific properties of the distribution. In contrast, in the areas where the elements of the random sample are sparse, one has $\hat{f}_*(x_i) < \bar{s}$ and consequently $s_i > 1$. This causes "flattening" and thus – advantageous to estimation quality – additional smoothing of the kernel estimator in peripheral regions (primarily the so-called "tails") of the distribution.

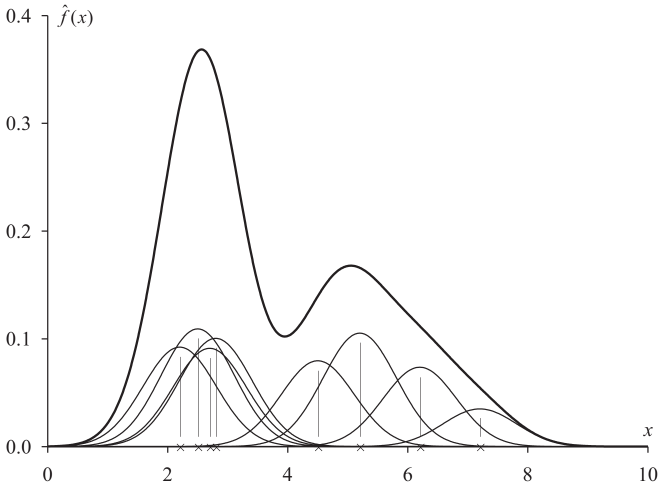

Figure 2. Kernel estimator with smoothing parameter modification

Statistical kernel estimators are dealt with in Kulczycki (2005), Silverman (1986), Wand and Jones (1994). Information on the subject of their applications in standard classification tasks can be found in Devroye et al. (1996), Duda et al. (2001), Ledl (2004), McLachlan (2004); the publication by Kulczycki (2008) can be also recommended.

2.2. Sensitivity analysis of neural networks

When modeling multidimensional problems using artificial neural networks, particular components of an input vector are most often characterized by differentiated significance of information, and in consequence influence differently the result of the data processing. In order to eliminate redundant – from the point of view of the problem investigated – input vector components, sensitivity analysis of the network with respect to particular learning data is often used. The basic factor for network reduction is sensitivity of the output function with regard to particular input data.

The essence of the sensitivity method (Zurada, 1992) consists in defining – after the network learning phase – the influence of the particular inputs u_i

for $i = 1, 2, ..., m$ on the output value y, which is characterized by the real coefficients

$$S_i = \frac{\partial\, y(u_1, u_2, ..., u_m)}{\partial\, u_i} \quad \text{for } i = 1, 2, ..., m. \tag{10}$$

Next, one aggregates the particular coefficients $S_i^{(p)}$ originating from successive iterations of the previous phase and corresponding to the sensitivity of subsequent learning data, with $p = 1, 2, ..., P$. The result is the final coefficient $\bar{S}_i$ given by the formula

$$\bar{S}_i = \sqrt{\frac{\sum\limits_{p=1}^{P} (S_i^{(p)})^2}{P}} \quad \text{for } i = 1, 2, ..., m. \tag{11}$$

After the sorting operation for the vector $\bar{S}_i$ according to decreasing values, analysis is performed of the relevance of particular components to the result of network operation, and then the least important inputs are eliminated.

In the general case the above algorithm can be used repeatedly to achieve further reduction. However, during empirical testing of the classification method developed here, such action did not bring positive results and so was forsaken.

The application of the above method led to an increase in speed, as well as reduction of errors of learning and generalization, while at the same time reducing the input dimension of the artificial neural network by removing information of little significance or even eliminating data (input vector components) having unfavorable influence on the correctness of the obtained result. Detailed considerations concerning the above procedure are found in Engelbrecht et al. (1995), Zurada (1992).

3. An algorithm for interval classification

3.1. One-dimensional case

This section considers the one-dimensional case, i.e. when $n = 1$. Let therefore be given the quantity having undergone the classification procedure, for the case here considered, represented by the (one-dimensional) interval

$$[\underline{x}, \overline{x}], \tag{12}$$

with $\underline{x} \leq \bar{x}$; if $\underline{x} = \bar{x}$ then the classic case is obtained where the quantity is precise (e.g. deterministic or sharp). Assume also that the real number sets (patterns):

$$x_1^1, x_2^1, \dots, x_{m_1}^1 \tag{13}$$
$$x_1^2, x_2^2, \dots, x_{m_2}^2 \tag{14}$$
$$\vdots$$
$$x_1^J, x_2^J, \dots, x_{m_J}^J \tag{15}$$

represent subsequent J marked classes of the sizes m_1, m_2, ..., m_J, respectively. The upper index, introduced in the above notation, characterizes membership of an element in a given class. As stated before, the task of classification consists of deciding to which of these groups the tested element (12) should be assigned.

Let now $\hat{f}_1$, $\hat{f}_2$, ..., $\hat{f}_J$ denote kernel estimators of a probability distribution density, calculated successively based on sets (13)-(15) treated as random samples – a description of the methodology used for their construction is contained in Section 2.1. In accordance with the classic Bayes approach (Duda et al., 2001), the classified element $\tilde{x} \in \mathbb{R}$ should then be assigned to the class, for which the value

$$m_1\hat{f}_1(\tilde{x}), m_2\hat{f}_2(\tilde{x}),\ \dots\ ,\ m_J\hat{f}_J(\tilde{x}) \tag{16}$$

is the biggest. In the case of information of interval type, represented by element (12), one can infer that this element belongs to the class, for which the expression

$$\frac{m_1}{\overline{x}-\underline{x}}\int_{\underline{x}}^{\overline{x}} \hat{f}_1(x)\,\mathrm{d}x, \frac{m_2}{\overline{x}-\underline{x}}\int_{\underline{x}}^{\overline{x}} \hat{f}_2(x)\,\mathrm{d}x, ..., \frac{m_J}{\overline{x}-\underline{x}}\int_{\underline{x}}^{\overline{x}} \hat{f}_J(x)\,\mathrm{d}x \tag{17}$$

is the biggest.

Considering the limit transitions $\overline{x} \to \tilde{x}^+$ and $\underline{x} \to \tilde{x}^-$ for a fixed $\tilde{x} \in \mathbb{R}$, due to the continuity of the function K used here – given by formula (4) – consequently implying the continuity of the kernel estimator $\hat{f}_j$, one obtains

$$\lim_{\substack{\underline{x} \to \tilde{x}^- \\ \overline{x} \to \tilde{x}^+}} \frac{1}{\overline{x}-\underline{x}}\int_{\underline{x}}^{\overline{x}} \hat{f}_j(x)\,\mathrm{d}x = \hat{f}_j(\tilde{x}) \quad \text{for } j = 1, 2, ..., J. \tag{18}$$

The expressions specified in formula (17) reduce therefore to the classic type (16).

In formula (17), the positive expression $1/(\overline{x}-\underline{x})$ can be omitted as irrelevant in an optimization problem, and so the compared quantities become

$$m_1\int_{\underline{x}}^{\overline{x}} \hat{f}_1(x)\,\mathrm{d}x, m_2\int_{\underline{x}}^{\overline{x}} \hat{f}_2(x)\,\mathrm{d}x, ..., m_J\int_{\underline{x}}^{\overline{x}} \hat{f}_J(x)\,\mathrm{d}x. \tag{19}$$

Moreover, for any $j = 1, 2, ..., J$ one can note

$$\int_{\underline{x}}^{\overline{x}} \hat{f}(x)\,\mathrm{d}x = \hat{F}(\overline{x}) - \hat{F}(\underline{x}), \tag{20}$$

where

$$\hat{F}(x) = \int_{-\infty}^{x} \hat{f}(y)\,\mathrm{d}y. \tag{21}$$

The above value can be analytically calculated, by substituting the equalities defining the kernel estimator (7) (for $n = 1$) and kernel (4) used here, yielding

$$\hat{F}(x) = \sum_{i=1}^{m} \left[\frac{(x^2 - 2xx_i + x_i^2 + h^2 s_i^2)\ \mathrm{arctg}\left(\frac{x-x_i}{s_i h}\right) + hs_i(x - x_i)}{x^2 - 2xx_i + x_i^2 + h^2 s_i^2} + \frac{\pi}{2} \right], \tag{22}$$

where again the positive constant $1/m\pi$ has been omitted. Finally, it should be acknowledged that the considered element belongs to the class, for which the corresponding expression in formula (19) is the biggest, whereby the integral appearing there for any $j = 1, 2, ..., J$ can be effectively calculated using equalities (20) and (22). The above completes the classification algorithm for the one-dimensional case. For the illustration of the interpretation see Fig. 3.

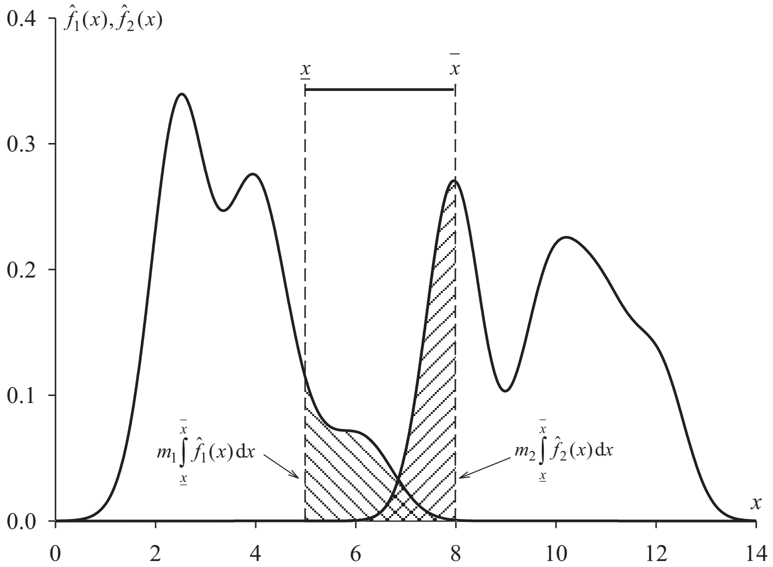

Figure 3. Interpretation for classification procedure according to formula (19)

3.2. The multidimensional case

The concept, presented in the previous subsection, can be naturally generalized for the multidimensional case, i.e. when $n > 1$. Thus, if information of interval type is represented by the interval vector

$$\begin{bmatrix} [\underline{x_1}, \overline{x_1}] \\ [\underline{x_2}, \overline{x_2}] \\ \vdots \\ [\underline{x_n}, \overline{x_n}] \end{bmatrix} \tag{23}$$

and sets (13)-(15) contain the elements of the space $\mathbb{R}^n$, then one can infer that the considered element is assigned to the class with the biggest value for the expression

$$m_1 \int_E \hat{f}_1(x)\,\mathrm{d}x, m_2 \int_E \hat{f}_2(x)\,\mathrm{d}x, ..., m_J \int_E \hat{f}_J(x)\,\mathrm{d}x, \tag{24}$$

where $E = [\underline{x_1}, \overline{x_1}] \times [\underline{x_2}, \overline{x_2}] \times ... \times [\underline{x_n}, \overline{x_n}]$. This is slightly different, though, for the algorithm for calculating the integrals appearing above. However, owing to the properties of the product kernel used here, for any fixed $j = 1, 2, ..., J$ and the kernel K, the following equality is true:

$$\int_E K(x)\,\mathrm{d}x = [\mathcal{F}(\overline{x_1}) - \mathcal{F}(\underline{x_1})][\mathcal{F}(\overline{x_2}) - \mathcal{F}(\underline{x_2})] \ldots [\mathcal{F}(\overline{x_n}) - \mathcal{F}(\underline{x_n})], \tag{25}$$

where $\mathcal{F}$ denotes the primitive of the function $\mathcal{K}$, introduced by definition (5). Taking into account the definition of the kernel estimator with product kernel (7), as well as the analytical form of the primitive function contained in formula (22), the above completes the procedure for classification of interval type information, for the multidimensional case, too.

3.3. Calculational complexity of the algorithm

From the point of view of calculational complexity, it is worth underlining the two-phased nature of the method presented in this paper. The first stage contains the complex procedures for constructing the classifier, which are executed once at the beginning. The most time-consuming is the algorithm for calculating the smoothing parameter using the plug-in method of complexity $O(nm^2)$. The same complexity characterizes the calculations of the smoothing parameter modification procedure.

On the contrary, the procedure for calculating the values of the kernel estimator has the complexity $O(n\,m)$. Considering that the number of such operations is equal to the number of assumed classes J, the calculational complexity of the second phase is linear with respect to all three parameters: n, m and J, where m characterizes here the size of particular patterns. This implies a relatively short calculation time, which, after an earlier execution of the first phase, in most practical problems allows for the application of the investigated algorithm in real time, in an on-line regime.

4. Procedures for increasing classification quality

4.1. Reducing pattern size

In practice, some elements of sets (13)-(15), constituting patterns of particular classes, may have insignificant or even negative – in the sense of classification correctness – influence on the quality of obtained results. Their elimination should therefore imply a reduction in the number of erroneous assignments, as well as decreasing calculation time. To this aim the sensitivity method for learning data, used in artificial neural networks, described in Section 2.2, will be applied.

To meet the requirements of this procedure, the definition of kernel estimator will be generalized below with the introduction of the nonnegative coefficients w_1, w_2, ..., w_m, normed by the condition

$$\sum_{i=1}^{m} w_i = m, \tag{26}$$

and mapped onto the particular elements of the random sample. The basic form of the kernel estimator (3) then takes the form

$$\hat{f}(x) = \frac{1}{mh^n} \sum_{i=1}^{m} w_i K\left(\frac{x - x_i}{h}\right). \tag{27}$$

Formula (7) undergoes analogous generalization. The value of the coefficient w_i may be interpreted as indicating the significance of the i-th element of the pattern for classification correctness. Note that if $w_i \equiv 1$, then definition (27) is reduced to the initial form (3).

In the method designed here, for the purpose of reduction of sets (13)-(15), separate neural networks are built for each investigated class. In order to ensure coherence of the notation below, let now the index $j = 1, 2, \dots, J$ characterizing particular classes, be arbitrarily fixed.

The constructed network has three layers and is unidirectional, with m inputs (corresponding to particular elements of a pattern), a hidden layer whose size is equal to the integer part of the number $\sqrt{m}$, and also one output neuron. This network is submitted to a learning process using a data set comprising of the values of particular kernels for subsequent pattern elements, while the given output constitutes the value of the kernel estimator calculated for the pattern element under consideration. Apart from the above topology, as a result of empirical research, the maximum number of epochs was assumed as 100, the maximum learning error 0.01, the learning speed 0.3, and the momentum coefficient as 0.1. After finishing the learning process, the thus obtained network is subject to sensitivity analysis with respect to the learning data, in accordance with the method presented in Section 2.2. The resulting coefficients $\bar{S}_i$ describing sensitivity, obtained on the basis of formula (11), constitute the fundament for calculating preliminary values

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

after which they are normed to

$$w_i = m \, \frac{\tilde{w}_i}{\sum\limits_{i=1}^{m} \tilde{w}_i} \tag{29}$$

with the aim of guaranteeing fulfillment of condition (26). It is worth noting that the form of formulas (10)-(11) guarantees in practice that not all of the quantities $\tilde{S_i}$ are equal zero, and so the above defined operations are feasible. The shape of formula (28) results from the fact that the network created here is the most sensitive to atypical and redundant elements, which – taking into account the form of kernel estimator (27) – implies a necessity to map the appropriately smaller values $\tilde{w}_i$, and in consequence w_i, to them. The coefficients (29) characterize – according to the idea presented during formulation of generalized form (27) – the significance of particular elements of the pattern, for the correctness of the classification procedure.

Empirical research confirmed the natural assumption that the pattern set should be relieved of those elements for which $w_i < 1$. (Note that, thanks to normalization made with formula (29), the mean value of coefficients w_i equals 1.) Decreasing of such an assumed threshold value resulted in a significant drop in the degree of pattern size reduction, while in the vicinity of the value 1 the influence on classification quality was practically unnoticeable, however, considerable decrease implied a sizable rise in the number of errors. On the other hand, an increase in this value caused a sharp fall in classification quality, due to a loss of valuable and non-redundant information included in the pattern.

Another procedure – besides the above presented algorithm for reduction of pattern size based on sensitivity analysis – from the group of methods dedicated to kernel estimators is the weighted Parzen windows algorithm contained in Babich and Camps (1996). The concept worked out in this paper was compared with the above algorithm as well as with other methods of eliminating elements of pattern sets, for example by the natural percentage reduction or the algorithm of k-nearest neighbors described in Mitra et al. (2002). In all cases, the results obtained on the basis of the sensitivity method were significantly better for the task of classification of interval information. An additional noteworthy positive aspect of this concept was the lack of necessity of introducing arbitrary parameters of fundamental relevance, which often requires laborious preliminary research.

4.2. Correcting the smoothing parameter and modification intensity values

The literature on the subject often presents the opinion that the classic universal methods of calculating the smoothing parameter value – most often based on a quadratic criterion – are not proper for the classification task. For example, in Ghosh et al. (2006) experimental research conducted on two classes was presented showing the significant difference between the value of this parameter when calculated by minimizing integrated square error, and when obtained by minimizing the number of misclassifications. However, the latter method is difficult in practical use for the multidimensional case, due to an extraordinarily long calculation time – a problem which becomes more important with the bigger number of classes. Available literature does not suggest a definitive solution for such a task.

This work proposes introducing $n + 1$ multiplicative correcting coefficients for the values of the parameter defining the intensity of modification procedure c and smoothing parameters for particular coordinates h_1, h_2, …, h_n, with respect to the optimal ones, calculated using the integrated square error criterion. Denote them as $b_0 \geq 0$, $b_1, b_2, \ldots, b_n > 0$, respectively. Note that $b_0 = b_1 = \ldots = b_n = 1$ means, in practice, no correction. Next, through a comprehensive search using a grid with a relatively large discretization value, one finds the most advantageous points with respect to minimal incorrect classifications. The final phase is a static optimization procedure in the $(n + 1)$-dimensional space, where the initial conditions are the points chosen above, while the performance index is given as

$$J(b_0, b_1, \ldots, b_n) = \#\{\textit{incorrect classifications}\}, \tag{30}$$

where # denotes the power (size) of a set. The value of the above functional for a fixed argument is calculated with the help of the classic leave-one-out method. This value is an integer – to find the minimum a modified Hook-Jeeves algorithm (Kelley, 1999) was applied.

Following experimental research it was assumed that the grid used for primary search has intersections at the points 0.25, 0.5, …, 1.75 for every coordinate. For such intersections the value of functional (30) is calculated, after which the obtained results are sorted, and the five best become subsequent initial conditions for the Hook-Jeeves method, where the value of the initial step is taken as 0.2. After finishing every one of the above five "runs" of this method, the value of functional (30) for the end point is calculated, and finally the one with the smallest value among them is shown.

The above algorithm was comprehensively tested and compared with a related exemplary method available in literature (Marzio and Taylor, 2005), and achieved better results in both correctness of classification and speed. This mainly results from fixing the value calculated on the basis of the minimal integrated square error criterion as a starting point and searching for a solution in its neighborhood, by creating a separate algorithm with features specific to the problem under research. It is worth noting that the procedure of smoothing parameter modification, introduced here into the classification task, greatly improves the quality of the results, and was adapted in the here proposed algorithm by merely increasing the dimension of the search-space by one. Application of the leave-one-out method also seems to be advantageous, since – as opposed to procedures based on additional validation samples – it does not reduce pattern sizes.

5. Numerical verification

Verification of correctness of the method presented in this paper for classifying interval information was conducted with numerical simulation.

First a typical one-dimensional case will be presented in detail, where the samples representing two patterns are obtained from generators with normal distributions $\mathrm{N}(0, 1)$ and $\mathrm{N}(2, 1)$. Note that the theoretical division point is placed at a distance of only one standard deviation from both expected values. Classified elements were obtained through generation by one of the aforementioned generators with normal distribution of the first pseudorandom number, as well as the second – taken from a generator with uniform distribution – defining the location of the first as within an interval of arbitrarily assumed length. This represents information of interval type when there are no circumstances for the considered imprecision, although its size is known. Such an interpretation seems to be the most appropriate for the majority of practical interval analysis applications. One hundred sets of 1000 elements obtained from each pattern were subjected to classification. The results are shown in Table 1, which also contains in the gray column – for comparison – results for testing precise elements.

Note that in particular columns of Table 1, as size increases, mean classification error and its standard deviation decrease. Another tendency – natural for tasks with interval data – is the increase in mean classification error as interval length increases. This is present in all rows. Obviously, the greater the imprecisation data, the worse the quality of analysis. However, it should also be emphasised that standard deviation shrinks as the interval length grows – this is the result of an ever more effective averaging, „stabilizing" the obtained results. The above is not valid for the last columns, where the interval length equals 5, which also seems obvious, given that the patterns have distributions $\mathrm{N}(0, 1)$ and $\mathrm{N}(2, 1)$ – such a large inaccuracy of the classified element is inadequate for treatment in the thus defined classification problem.

Table 1. Results of numerical verification for patterns N(0, 1) and N(2, 1) with notation: *mean classification error ± its standard deviation*

length	**0.00**	**0.10**	**0.25**	**0.50**	**1.00**	**2.00**	**5.0**
m							
10	0.1713 ± 0.0257	0.1720 ± 0.0215	0.1720 ± 0.0215	0.1723 ± 0.0215	0.1729 ± 0.0214	0.1761 ± 0.0208	0.1944 ± 0.0198
20	0.1655 ± 0.0160	0.1669 ± 0.0175	0.1669 ± 0.0176	0.1672 ± 0.0174	0.1680 ± 0.0171	0.1713 ± 0.0161	0.1888 ± 0.0131
50	0.1602 ± 0.0126	0.1605 ± 0.0124	0.1606 ± 0.0122	0.1609 ± 0.0122	0.1617 ± 0.0116	0.1652 ± 0.0108	0.1848 ± 0.0100
100	0.1596 ± 0.0122	0.1601 ± 0.0112	0.1602 ± 0.0112	0.1604 ± 0.0113	0.1615 ± 0.0110	0.1650 ± 0.0098	0.1827 ± 0.0079
200	0.1596 ± 0.0123	0.1602 ± 0.0112	0.1604 ± 0.0112	0.1609 ± 0.0111	0.1618 ± 0.0107	0.1650 ± 0.0091	0.1840 ± 0.0080
500	0.1591 ± 0.0125	0.1595 ± 0.0114	0.1596 ± 0.0111	0.1602 ± 0.0110	0.1613 ± 0.0101	0.1647 ± 0.0088	0.1844 ± 0.0076
1000	0.1579 ± 0.0140	0.1584 ± 0.0131	0.1588 ± 0.0131	0.1591 ± 0.0127	0.1603 ± 0.0118	0.1637 ± 0.0098	0.1833 ± 0.0086

After applying the procedure for reducing pattern sets, presented in Section 4.1, the number of wrong classifications was lowered by approximately 15%, while the size of patterns was reduced by approximately 40%. The conjunction of these results is particularly worth attention: while appropriately reducing pattern sizes, which does imply a significant increase in calculation speed, the classification quality is also importantly improved.

Figs. 4 and 5 show exemplary coefficients assigned to particular elements of both pattern sets as a result of the use of sensitivity analysis of neural networks. One can see that coefficients below the assumed threshold value 1 are mapped primarily to atypical elements, but also to some elements in dense areas of patterns, where they were treated as redundant, and their role was taken over by elements considered to be more representative, with coefficients greater than 1. Interpretation of these results can be a valuable aid in investigations of unusual aspects of the problem considered, e.g. for nonstationary patterns (13)-(15).

A further subject of research was the task of correcting the smoothing parameter and intensity of its modification, presented in Section 4.2. This procedure was carried out after the reduction of patterns. The obtained results caused a further decrease in the number of classification errors to approximately 14%.

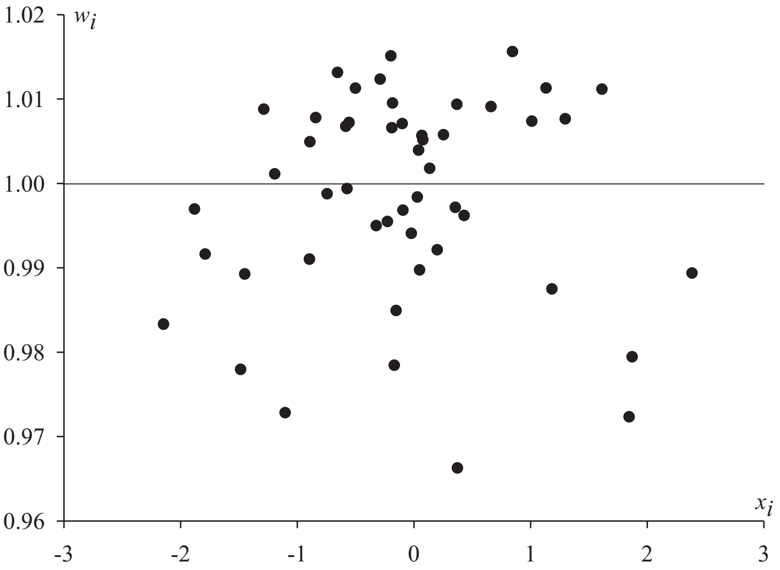

Figure 4. Coefficients w_i assigned to particular elements x_i of the pattern N(0,1)

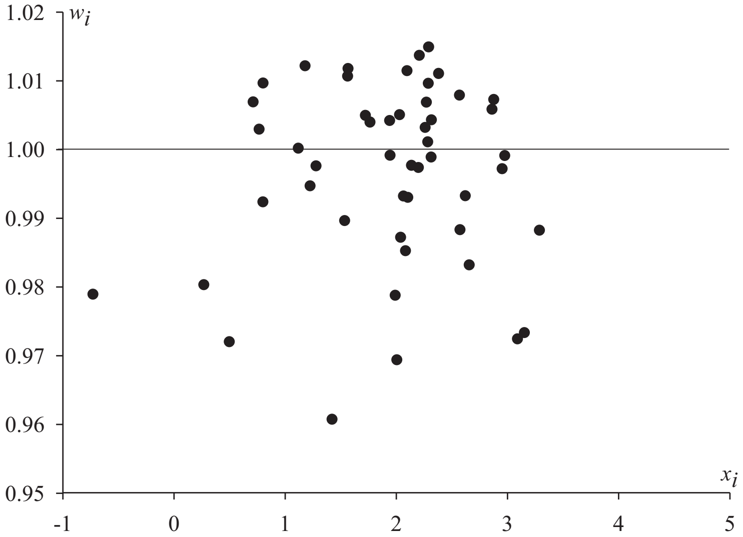

Figure 5. Coefficients w_i assigned to particular elements x_i of the pattern N(2,1)

The above tests were also carried out for the case where interval length was defined randomly, as it was, for example, in Souza and Carvalho (2004). The uniform distribution over the interval of the form $[0, 2d]$, while d denotes the respective lengths of intervals shown in Table 1, was assumed. The obtained results were comparable in relation to those previously presented for fixed lengths. A similar situation occurred with the joining of both concepts, when both the interval length and its “catching point” were random.

Further experiments concerned the symmetry of errors in the cases of evenly and unevenly matched patterns. In the first of them both mean classification error and its standard deviation are not asymmetric, even with very small pattern sizes, which – although generally intuitively correct – is worth mentioning. In the case of unevenly matched patterns, the number of elements incorrectly classified from a class with larger pattern to a class with a smaller pattern is lower and this tendency becomes more visible as the ratio of pattern sizes moves away from 1 – this mainly results from a better quality of larger pattern.

The subsequent study dealt with interval classification of multidimensional information. Generally, increasing the dimension by 1 required – in order to maintain a given accuracy – approximately a fourfold increase in pattern size. This is in accordance with the theoretical properties of kernel estimators, arising from the ever present “multidimensionality curse”; see e.g. (Kulczycki, 2005 – Section 3.1.9; Silverman, 1986 – Section 4.5.2). It is, however, worth underlining that once the above requirement was fulfilled, the properties of the algorithm did not become an issue.

The subject of subsequent investigations concerned multimodal patterns, including cases of incoherent subsets, separated by fragments of other patterns. This required an increase in sample sizes of 10-50% in practice. The result is obvious from an intuitive point of view – every mode and subset of particular patterns was defined by reducing the number of elements, and so the patterns were naturally not as accurate as in the unimodal case. However, apart from an insignificant fall in classification quality, the algorithm itself did not undergo any change, either in the sense of its structure or calculation time, which is characteristic for procedures based on the kernel estimators methodology.

Similar results were obtained for greater numbers of classes. These results do not have any specific differentiating features and generally confirm the accuracy of the method. This is particularly worth stressing, however, due to the unsavory habit commonly found in subject literature of presenting methods of classification actually dedicated only to the case of two classes and silent of the fact that their application with respect to greater numbers is practically impossible.

The method was also positively verified for the case with bounded support, realized by symmetrical reflection of kernels (Kulczycki, 2005 – Section 3.1.8; Silverman, 1986 – Section 2.10), as well as when patterns of particular classes were obtained through clustering – then results were even slightly more advantageous, especially regarding effectiveness of the pattern reduction procedure.

Elements wrongly mapped during clustering were a minority in such obtained pattern sets, and were successfully eliminated with the reduction algorithm.

All verification research carried out showed that increasing pattern size resulted in a decrease in both mean value of classification errors as well as their standard deviation, which in practice allows for successive improvement in quality of classification when collecting new data. Furthermore, as the length of interval increased, so, slightly, did the classification error, to a certain extent. From the application point of view the above features are worth underlining, as they indicate that it is possible to enhance classification quality by increasing the available information in terms of greater patterns as well as more accurate interval elements being investigated. Due to the time-consuming nature of the calculations, however, practical tasks require a compromise to be established between the number and accuracy of information available, and the quality of results achieved.

Another subject of study was the comparison of the procedure presented here with other methods of corresponding conditioning, both natural in character as well as available in literature, based on other concepts.

For the former, an algorithm was used where elements of every pattern included in the classified interval element were counted. In this instance the results clearly proved worse than those coming from the method presented in this paper, especially for short intervals.

The procedure described here was also compared with another natural concept consisting in replacing the classified interval element with its middle value, and then using the Bayes classification for precise data. The results of the two methods were similar for smaller lengths, however, as the length increased the procedure proposed in this paper became ever more advantageous, in the sense of both smaller mean values of classification error and their standard deviation, especially for atypical multimodal and/or multicomponent, partly overlapping pattern sets.

The latter group of compared concepts contained a method based on the support vector machine, according to the algorithm presented in Zhao et al. (2005). When this procedure is used three types of decisions regarding assignment of the tested interval element are obtained: to the first class, to the second, or none at all. Comparing these results with those acquired through the procedures in this paper, it must be stated that the former were worse by 5% to even 50%. If, however, it is taken that not making a decision is a proper action, the above ceases to be unconditionally true, although then the conditions of the problem are different. Worth noting is the fact that the method presented here could also be administered to them, most easily by not making a decision in the classification when none of the expressions in formulas (16) or (24) is significantly greater than the others. It is to be underlined that the method presented in Zhao et al. (2005) does not lend itself to generalization to the multidimensional case or to more than two classes.

Finally, verification was performed using benchmark data. Due to the specific conditioning of the method presented here, this type of data was not found in public repositories or on websites. For this reason the interval data used in the tests below was formed from precise data taken from repositories in the same way as described at the beginning of this section – a pseudorandom number obtained from a generator with uniform distribution becomes the location of the above mentioned precise data within an interval of arbitrarily assumed length.

First, the benchmark data was investigated for *Toy 2D*, located at the website *http://www.cse.ust.hk/˜twinsen/assgn2.pdf*. The two-dimensional random samples – learning and testing – represent the phases of the moon. The former contains 2 152 elements belonging to the first class (connected with the first quarter moon) and 2 444 to the second (the third quarter moon), while the latter was formed on the basis of a two-dimensional regular grid and includes 26 130 elements of the first class and 34 371 of the second.

The results of numerical verification are presented in Table 2. In the gray column, for comparison – as with Table 1 – the results are shown for precise data. It can be seen that the loss of information resulting from the introduction of imprecision of interval type did not cause a significant increase in error of classification (carried out here also using the method investigated in this paper) in the first four columns corresponding to the interval lengths 0.1, 0.25, 0.5 and 1.0. However, for the lengths 2.0 and 5.0, where they are the multiples of the range of sample data, such imprecise interval information obviously considerably lowers the quality of classification.

Table 2. Results of numerical verification for *Toy 2D* data

length of interval	**0.00 × 0.00**	**0.10 × 0.10**	**0.25 × 0.25**	**0.50 × 0.50**	**1.00 × 1.00**	**2.00 × 2.00**	**5.00 × 5.00**
mean classification error	0.0681	0.0737	0.0748	0.0766	0.0828	0.1097	0.2226

Further research was conducted on the real data set *Iris Plants Database*, taken from the well-known repository of the *Center for Machine Learning and Intelligent Systems at the University of California, Irvine*, at *http://archive.ics.uci.edu/ml/datasets/Iris*. This data set contains the lengths and widths of petals and sepals of three species of iris *setosa canadensis*, *versicolor* and *virginica*; the first two classes are linearly separable. The set of data is composed of three classes of equal size represented by altogether 150 elements, although here the learning and testing samples have not been defined. Because of this, in the research described below, data was divided randomly into two subsets: learning and test. The results presented in Table 3 are the mean of 1000 divisions performed randomly. The intervals were generated as before.

The results obtained clearly show many advantages of the classification method presented in this paper. The first was that its sensitivity to the "multi-dimensionality curse" turned out to be lower in practice than theory suggested, as classification with a 4-dimensional vector was carried out satisfactorily on patterns of about 40 elements. Another confirmation of the effectiveness of this method comes from the comparison with results described in Kotsiantis and Pintelas (2005) for precise data. In that article the classification error was never below 4.5%. A similar result was obtained for the method presented in this paper for precise data (see gray column in Table 3). Despite a decrease in the accuracy of information classified due to their change into imprecise data, the results did not worsen for the interval up to 1.0 cm, which is particularly worth underlining in conclusion.

Details of numerical verifications can be found in Kowalski (2009).

Table 3. Results of numerical verification for *Iris* data

length of interval	**0.00 × 0.00**	**0.10 × 0.10**	**0.25 × 0.25**	**0.50 × 0.50**	**1.00 × 1.00**	**2.00 × 2.00**	**5.00 × 5.00**
mean classification error	0.041	0.045	0.047	0.048	0.049	0.066	0.156

6. Summary

This paper presents the complete Bayes algorithm – thereby ensuring minimum potential losses – for the classification of multidimensional imprecise information of interval type, where patterns of particular classes are given on the basis of sets of precisely defined elements, with no limits to the number of classes. In addition, two optional procedures are provided, which improve and enhance the quality of classification: reduction in pattern size and correction of the classifier parameter values.

Considering the calculational complexity, it is worth underlining that the investigated method has two phases. Time-consuming algorithms for constructing the classifier are executed only once in the initial stage. The aforementioned optional procedures may be run irregularly, depending on the computer system free computing power. The classification itself of imprecise information is carried out in a relatively short time, which may be of great practical significance in many applicational tasks. This is achieved mainly due to the analytical form of the formulas used.

Numerical testing wholly confirmed the positive features of the method worked out. It was carried out with the use of pseudorandom and benchmark data. In particular, the results show that the classifying algorithm can be used successfully for inseparable classes of complex multimodal patterns as well as for those consisting of incoherent subsets at alternate locations. This is due to the application of the statistical kernel estimators methodology, which makes the above procedure independent of the shapes of patterns – their identification is an integral part of the presented algorithm. As shown by numerical verification, the algorithm has beneficial features in the multidimensional case, too. The results also compared advantageously to those obtained by applying support vector machines as well as by the two natural methods.

The reduction in the size of patterns proved to be especially effective in the case of interval classification, as – apart from the obvious profits gained from eliminating pattern elements with negative influence on classification – this type of information is averaging in character and so removing redundant elements has particularly insignificant influence on the quality of the procedure.

The task of classifying interval information based on precise data can be interpreted illustratively with the example where the patterns present actual, precisely measured quantities, while intervals being classified represent uncertainties and imprecision in plans, estimations or measurements difficult to make. In particular, pattern sets may consists of very accurate measurements, in which errors are practically ignored, while the classified interval constitutes a measurement taken from another, much less accurate apparatus or carried out in much worse conditions. Another example of the application of this kind of classification is the possibility of treating precise data as actual information from the past, e.g. temperature or currency exchange rates, while the classified element represents a prognosis, which, by its very nature, is limited in precision.

References

Alefeld, G. and Hercberger, J. (1986) *Introduction to Interval Computations.* Academic Press, New York.

Babich, G.A. and Camps, O.I. (1996) Weighted Parzen Windows for Pattern Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **18**, 567-570.

Devroye, L., Gyorfi, L. and Lugosi, G. (1996) *A Probabilistic Theory of Pattern Recognition.* Springer, New York.

Duda, R.O., Hart, P.E. and Stork, D.G. (2001) *Pattern Classification.* Wiley, New York.

Engelbrecht, A.P., Cloete, I. and Zurada, J. (1995) Determining the Significance of Input Parameters Using Sensitivity Analysis. *International Workshop on Artificial Neural Networks*, Torremolinos (Spain), 7-9 June 1995, LNCS, **930**, 382-388.

Gil, M.A. and Hryniewicz, O. (2009) Statistics with Imprecise Data. In: R.A. Meyers, ed., *Encyclopedia of Complexity and Systems Science 2009*, Springer, Heidelberg, 8679-8690.

Ghost, A.K, Chaudhuri, P. and Sengupta, D. (2006) Classification Using Kernel Density Estimation: Multiscale Analysis and Visualization. *Technometrics*, **48**, 120-132.

Hand, D.J. (1997) *Construction and Assessment of Classification Rules.* Wiley, Chichester.

Jaulin, L., Kieffer, M., Didrit, O. and Walter, E. (2001) *Applied Interval Analysis.* Springer, Berlin.

Kacprzyk, J. (1997) *Multistage Fuzzy Control: A Model-Based Approach to Fuzzy Control and Decision Making.* Wiley, Chichester.

Kelley, C.T. (1999) *Iterative Methods for Optimization.* SIAM, Philadelphia.

Klir, G.J. and Yuan, B. (1995) *Fuzzy Sets and Fuzzy Logic: Theory and Applications.* Prentice Hall, Upper Saddle River.

Kotsiantis, S.B. and Pintelas, P.E. (2005) Logitboost of Simple Bayesian Classifier. *Informatica*, **29**, 53-59.

Kowalski P.A. (2009) *Klasyfikacja bayesowska informacji niedokładnej typu przedziałowego* (*Bayes classification of imprecise information of interval type; in Polish*). Ph.D.-thesis, Systems Research Institute, Polish Academy of Sciences, Warsaw.

Kulczycki, P. (2005) *Estymatory jądrowe w analizie systemowej* (*Kernel estimators in systems analysis; in Polish*). WNT, Warsaw.

Kulczycki, P. (2008) Kernel Estimators in Industrial Applications. In: B. Prasad, ed. *Soft Computing Applications in Industry*, Springer-Verlag, Berlin, 69-91.

Kulczycki, P., Hryniewicz, O. and Kacprzyk, J., eds. (2007) *Techniki informacyjne w badaniach systemowych* (*Information technologies in systems research; in Polish*). WNT, Warsaw.

Kulczycki, P., Kowalski, P.A. (2008) Klasyfikacja informacji niedokładnej typu przedziałowego ze zredukowanymi próbami wzorcowymi. (Classification of imprecise information of interval type with reduced samples; in Polish). In: O. Hryniewicz, A. Straszak and J. Studziński, eds. *Badania operacyjne i systemowe: środowisko naturalne, przestrzeń, optymalizacja*, IBS PAN, Warsaw, ser. *Badania Systemowe*, **63**, 305-314.

Kumar, R.A. and Tinku, A. (2004) *Information Technology: Principles and Applications.* Prentice Hall of India, New Delhi.

Ledl, T. (2004) Kernel Density Estimation*,* Theory and Application in Discriminant Analysis. *Austrian Journal of Statistics*, **33**, 267-279.

Marzio, Di M. and Taylor, C. (2005) On boosting kernel density methods for multivariate data: density estimation and classification. *Statistical Methods and Applications*, **14**, 163-178.

McLachlan, G.J. (2004) *Discriminant Analysis and Statistical Pattern Recognition.* Wiley, Hoboken.

Mitra, P., Murthy, C.A. and Pal, S.K. (2002) Density-Based Multiscale Data Condensation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**, 734-747.

Moore, R.E. (1966) *Interval Analysis.* Prentice-Hall, Englewood Cliffs.

Rice, J.A. (1994) *Mathematical Statistics and Data Analysis.* Duxbury, Pacific Grove.

Ripley, B.D. (1996) *Pattern Recognition and Neural Networks.* Cambridge University Press, Cambridge.

Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Souza, De R.M.C.R. and Carvalho, De F.A.T. (2004) Dynamic clustering of interval data based on adaptive Chebyshev distances. *Electronics Letters*, **40**, 658-660.

Tou, J.T. and Gonzales, R.C. (1974) *Pattern Recognition Analysis.* Addison-Wesley, Reading.

Wand, M.P. and Jones, M.C. (1995) *Kernel Smoothing.* Chapman and Hall, London.

Zhao,Y., He, Q. and Chen, Q. (2005) An Interval Set Classification Based on Support Vector Machines. *Joint International Conference on Autonomic and Autonomous Systems*, *2nd International Conference on Networking and Services*, Silicon Valley (USA), 25-30 September 2005, 81-86.

Zurada, J. (1992) *Introduction to Artificial Neural Systems.* West Publishing, St. Paul.