Bayesian Classification of Interval-Type Information

P. Kulczycki[1,2], P.A. Kowalski[1,2]

[1] Polish Academy of Sciences, Systems Research Institute,
Centre of Information Technology for Data Analysis Methods, Warsaw, Poland.
[2] Cracow University of Technology,
Department of Automatic Control and Information Technology, Cracow, Poland.
e-mail: {Piotr.Kulczycki, Piotr.A.Kowalski}@ibspan.waw.pl

Abstract: The subject of Bayes classification of imprecise multidimensional information of interval type by means of patterns defined through precise data (i.e. deterministic or sharp) is investigated here. To this aim the statistical kernel estimators methodology was applied, which avoids the pattern shape for the resulting algorithm. In addition, elements of pattern sets which have insignificant or negative influence on 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.

Keywords: *data analysis, classification, interval information, nonparametric methods, kernel estimators, reduction in pattern size, classifier parameter correction, neural networks.*

1. Introduction

The current dynamic development in computer technology offers a continuous increase in both capability and speed of contemporary calculational systems, thus allowing ever more frequent use of methods which up to now had only been applied to a relatively limited extent. One of these methods is the analysis of information which is imprecise in various – depending on a problem's conditioning – forms, for example uncertain (statistical methods [4]) or fuzzy (fuzzy logic [5]). Lately many applications have noted 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 dependence $\underline{x} \leq x \leq \overline{x}$, and in consequence this quantity can be associated with the interval

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

Interval analysis is a separate mathematical domain, with its own formal apparatus based on an axiom of the sets theory [11]. Its main advantage is the fact that by definition it models imprecision of a researched 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 identify and interpret, while also maintaining a formalism stemming from a convenient mathematical tool [9].

Among the fundamental tasks of data analysis lies that of classification [1, 3]. 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 is investigated – elements defining patterns are defined precisely (e.g. deterministic in probability approach, sharp for the case of fuzzy logic, or in relation to notation (1) fulfilling equality $\underline{x} = \overline{x}$).

2. Main contents

This work deals with a complete procedure for classification of imprecise information, defined as 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 $\underline{x_k} \le \overline{x_k}$ for $k = 1,2,...,n$, when the patterns of particular classes are given as sets of precise data (i.e. 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 a minimum of potential losses occurring through classification errors. For a such formulated task the statistical kernel estimators methodology [8, 12, 14] 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 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 – calculation speed. Furthermore a method was designed to ensure additional improvements in classification results, obtained by correcting the values of classifier parameters.

Numerical testing wholly confirmed the positive features of the method investigated here. 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 thanks 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 positively to those obtained by applying support vectors machines as well as by the two natural methods.

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 difficult measurements 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 nature is limited in precision.

In particular, the method investigated here can be applied for purposes of the diagnosis process [2, 6, 7, 13]. Namely, let interval (1) or interval vector (2) represent a quantity or n quantities, respectively, whose values attest to the current or – in the case of fault prognosis – predicted technical state of a supervised device. Because of measurement errors and natural fluctuations, the interval form can be justified in many practical tasks. Let also patterns of particular classes represent the types of possible faults. The classification procedure described here allows for precise diagnostic readings to be obtained, with regard to interval character of investigated quantities.

More details concerning the basic version of the above presented method can be found in the paper [10].

References

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

[2] Gertler J.: *Fault Detection and Diagnosis in Engineering Systems*. Dekker, New York (1998).

[3] Hand D.J.: Construction and Assessment of Classification Rules. Wiley, Chichester (1997).

[4] Hryniewicz O.: *Reliability Sampling;* in Encyclopedia of Statistics in Quality and Reliability, Wiley, Chichester (2008).

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

[6] Korbicz J., Koscielny J.M., Kowalczuk Z., Cholewa W. (Editors): *Fault Diagnosis. Models, Artificial Intelligence Applications*. Springer, Berlin (2004).

[7] Kulczycki P.: *Wykrywanie uszkodzen w systemach zautomatyzowanych metodami statystycznymi*. Alfa, Warsaw (1998).

[8] Kulczycki P.: *Estymatory jadrowe w analizie systemowej*. WNT, Warsaw (2005).

[9] Kulczycki P., Hryniewicz O., Kacprzyk J. (Editors): *Techniki informacyjne w badaniach systemowych.* WNT, Warsaw (2007).

[10] Kulczycki P., Kowalski P.A.: *Bayes classification of imprecise information of interval type*. Control and Cybernetics, Vol. 40 (2011), pp. 101-123.
[11] Moore R.E.: *Interval Analysis*. Prentice-Hall, Englewood Cliffs (1966).
[12] Silverman B.W.: *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London (1986).
[13] Tadeusiewicz R., Ogiela M.R.: *Medical Image Understanding Technology*. Springer-Verlag, Berlin (2004).
[14] Wand M.P., Jones M.C.: *Kernel Smoothing*. Chapman and Hall, London (1995).