Classification of Interval Information with Data Drift

Piotr Kulczycki(✉) and Piotr A. Kowalski

Polish Academy of Sciences, Systems Research Institute, Warsaw, Poland
{kulczycki,pakowal}@ibspan.waw.pl

Abstract. The paper deals with the classification task of interval information, when processed data is gradually displaced, i.e. they originate from a nonstationary environment. The procedure worked out is characterized by its many practical properties: ensuring the minimum expected value of misclassifications; allowing influence on the probability of errors in classification to particular classes; reducing patterns by eliminating elements with insignificant or negative influence on the results' accuracy, enabling an unlimited number of patterns and their shapes. The appropriate modifications of the classifier not only lead to an increase in the effectiveness of the procedure, but above all adapt to data drift.

Keywords: Data analysis · Classification · Interval information · Data drift · Artificial neural network · Sensitivity method · Pattern size reduction · Classifier adaptation

1 Introduction

In most of the methods used today for classification [1], one assumes invariability in time of data stream under processing. However, more and more often, in particular for those models in which new – with the most current being the most valuable – elements are continuously added to patterns, this assumption is successfully ignored [5].

The presented paper proposes the procedure for classification of information given in the form of an interval for data which may have drifted – undergoing successive changes. The idea for a solution stems from the sensitivity method used in neural networks, together with nonparametric kernel estimators. Namely, particular elements of patterns receive weights proportional to their significance for correct results. Elements of the smallest weights are eliminated. In order to account for the data drift, those elements whose weights are currently small but increase successively are kept.

This paper is a novel elaboration of research presented in the paper [6] for the interval stationary case, and in the publication [7] for the deterministic nonstationary case.

2 Preliminaries

2.1 Statistical Kernel Estimators

Consider an n-dimensional random variable, with a distribution given by the density f. Its kernel estimator $\hat{f} : \mathbb{R}^n \to [0, \infty)$ is calculated on the basis of the random sample

H. Christiansen et al. (Eds.): CONTEXT 2015, LNAI 9405, pp. 495–500, 2015.
DOI: 10.1007/978-3-319-25591-0_38

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

and defined as

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

where the positive coefficient h is known as a smoothing parameter, while the measurable function $K : \mathbb{R}^n \to [0, \infty)$ symmetrical with respect to zero, having at this point a weak global maximum and fulfilling the condition $\int_{\mathbb{R}^n} K(x)\,\mathrm{d}x = 1$ is termed a kernel. For details see the monographs [4, 8, 10].

In this paper the one-dimensional Cauchy kernel is applied, for the multidimensional case generalized by the product kernel concept [4 – Sect. 3.1.3, 10 – Sects. 2.7 and 4.5]. For calculation of the smoothing parameter, the simplified method assuming normal distribution [4 – Sect. 3.1.5, 10 – Sect. 3.2.1] can be used, thanks to the positive influence of this parameter correction procedure applied in the following. For general improvement of the kernel estimator quality, modification of the smoothing parameter [4 – Sect. 3.1.6, 8 – Sect. 5.3.1] will be used, with the intensity $c \geq 0$. As its initial standard value $c = 0.5$ can be assumed.

2.2 Bayes Classification of Interval Information

Consider J sets

$$\{x_1'\ x_2', \ldots, x_{m_1}'\},\ \{x_1''\ x_2'', \ldots, x_{m_2}''\},\ \ldots,\ \{x_1''\cdots'\ x_2''\cdots', \ldots, x_{m_J}''\cdots'\} \tag{3}$$

representing assumed classes. The sizes $m_1, m_2, \ldots, m_J$ should be proportional to the "contribution" of particular classes in the population. Because of practical aspects, one can assume that the elements from sets (3) belong to the space $\mathbb{R}^n$. Representative elements, consisting of patterns, are characterized by considerable precision and are either deterministic in nature, or of interval type with length of this interval so small that it can be identified with its midpoint without any influence on the quality of the result.

Let now $\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_J$ denote kernel estimators of densities, calculated successively based on sets (3) treated as samples (1), according to the methodology from Sect. 2.1.

First consider the one-dimensional case $(n = 1)$. In accordance with the classic Bayes approach [1], ensuring a minimum of expected value of losses, the tested element $[\underline{x}, \overline{x}]$, with $\underline{x} < \overline{x}$, should be ranked to the class for which the value

$$m_1 \int_{\underline{x}}^{\overline{x}} \hat{f}_1(x)\,\mathrm{d}x,\ m_2 \int_{\underline{x}}^{\overline{x}} \hat{f}_2(x)\,\mathrm{d}x,\ \ldots,\ m_J \int_{\underline{x}}^{\overline{x}} \hat{f}_J(x)\,\mathrm{d}x \tag{4}$$

is the greatest. The above can be generalized by introducing to expressions (4) the positive coefficients $z_1, z_2, \ldots, z_J$:

$$z_1 m_1 \int_{\underline{x}}^{\overline{x}} \hat{f}_1(x)\,\mathrm{d}x,\ z_2 m_2 \int_{\underline{x}}^{\overline{x}} \hat{f}_2(x)\,\mathrm{d}x,\ \ldots, z_J m_J \int_{\underline{x}}^{\overline{x}} \hat{f}_J(x)\,\mathrm{d}x. \tag{5}$$

Taking as standard values $z_1 = z_2 = \ldots = z_J = 1$, formula (5) brings us to (4). By appropriately increasing the value z_i, a decrease can be achieved in the probability of erroneously assigning elements of the i-th class to other wrong classes. Thanks to this, it is possible to favor classes which are in some way noticeable or more heavily conditioned. For the classification, these are in a natural way classes defined by non-stationary patterns, it is worth increasing coefficients relating to more varying patterns. In such case, the initial value 1.25 can be proposed for further research.

In the multidimensional case, i.e. when $n > 1$, the tested element is

$$\begin{bmatrix} [\underline{x_1}, \overline{x_1}] \\ [\underline{x_2}, \overline{x_2}] \\ \vdots \\ [\underline{x_n}, \overline{x_n}] \end{bmatrix} \tag{6}$$

with $\underline{x_k} < \overline{x_k}$ for $k = 1, 2, \ldots, n$, and criterion (5) takes the following form:

$$z_1 m_1 \int_{\underline{x}}^{\overline{x}} \hat{f}_1(x)\,\mathrm{d}x,\ z_2 m_2 \int_{\underline{x}}^{\overline{x}} \hat{f}_2(x)\,\mathrm{d}x,\ \ldots,\ z_J m_J \int_{\underline{x}}^{\overline{x}} \hat{f}_J(x)\,\mathrm{d}x, \tag{7}$$

where $E = [\underline{x_1}, \overline{x_1}] \times [\underline{x_2}, \overline{x_2}] \times \ldots \times [\underline{x_n}, \overline{x_n}]$.

For the Cauchy kernel proposed here, generalized in the multidimensional case by the product kernel concept (see Sect. 2.1), the analytical form of quantities occurring in formulas (4), (5) and (7) are possible to obtain; see the paper [6].

2.3 Sensitivity Analysis for Learning Data – Reducing Pattern Size

When modeling by artificial neural networks, particular components of an input vector most often are characterized by diverse significance of information. Using a sensitivity analysis [12], one obtains the parameters $\bar{S}_i$ describing proportionally the influence of the particular inputs ($i = 1, 2, \ldots, m$) on the output value, and then the least significant inputs can be eliminated.

To apply the above procedure, the definition of the kernel estimator will be generalized with the introduction of the nonnegative coefficients $w_1, w_2, \ldots, w_m$, normed so that $\sum_{i=1}^{m} w_i = m$, and mapped to particular elements of random sample (1). The basic form of kernel estimator (2) then takes the form

$$\hat{f}(x) = \frac{1}{mh^n} \sum_{i=1}^{m} w_i K\left(\frac{x - x_i}{h}\right). \tag{8}$$

The coefficient w_i value may be interpreted as indicating the significance (weight) of the i-th element of the pattern to classification correctness.

For the purpose of calculation the weights w_i values, separate neural networks are built for each investigated class. 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. After this, the obtained network undergoes sensitivity analysis on learning data. The resulting coefficients $\bar{S}_i$ describing sensitivity, constitute the fundament for calculating the values

$$\tilde{w}_i = \left(1 - \frac{\bar{S}_i}{\sum_{j=1}^{m} \bar{S}_j}\right) \text{ normed to } w_i = m \frac{\tilde{w}_i}{\sum_{i=1}^{m} \tilde{w}_i}. \tag{9}$$

The shape of the formula defining the parameters $\tilde{w}_i$ results from the fact that the network created here is the most sensitive to atypical and redundant elements, which implies a necessity to map the appropriately smaller values $\tilde{w}_i$, and in consequence w_i, to them. Coefficients w_i represent the significance of particular elements of the pattern to accuracy of the classification. Because – thanks to normalization – the mean value of the coefficients w_i equals 1, the pattern set should be relieved of those elements for which $w_i < 1$.

3 Classification Procedure

This section presents the method for classification of interval information with data drift.

First one should fix the reference sizes of patterns (3), hereinafter denoted by $m_1^*, m_2^*, \ldots, m_J^*$. The patterns of these sizes will be the subject of a basic reduction procedure, described in Sect. 2.3. The sizes of patterns available at the beginning of the algorithm must not be smaller than the above referential values. These values can however be modified during the procedure's operation, with the natural condition that their potential growth does not increase the number of elements newly provided for the patterns. For preliminary research, $m_1^* = m_2^* = \ldots = m_J^* = 25 \cdot 2^n$ can be proposed.

The elements of initial patterns (3) are provided as introductory data. Based on these – according to Sect. 2.1 – the value of the parameter h is calculated (for the parameter c initially assumed to be equal 0.5). Next, corrections in the parameters c and $h_1, h_2, \ldots, h_n$ values are made by introducing $n+1$ multiplicative correcting coefficients. Denote them as $b_0 \geq 0, b_1, b_2, \ldots, b_n > 0$, respectively. Their values can be calculated by a static optimization procedure in the $(n+1)$-dimensional space, where

the initial conditions at the beginning are the points of a grid, or the previous values in the following steps, while the performance index is given as the number of misclassifications. To find the minimum a modified Hook-Jeeves algorithm [2] was applied.

The next procedure is the calculation of the parameters w_i values mapped to particular patterns' elements, separately for each class, as in Sect. 2.3. Following this, within each class, the values of the parameter w_i are sorted, and then the appropriate $m_1^*, m_2^*, \ldots, m_J^*$ elements of the largest values w_i are designated to the classification phase itself. The remaining ones undergo further treatment, which will be presented later, after Bayes classification has been dealt with.

The reduced patterns separately go through a procedure newly calculating the values of parameters w_i, shown in Sect. 2.3. Next, these patterns' elements for which $w_i \geq 1$ are submitted to further stages of the classification procedure, while those with $w_i < 1$ are sent to the beginning of the algorithm for further processing in the next steps of the algorithm, after adding new elements of patterns. The final, and also the principal part of the procedure worked out here is Bayes classification, presented in Sect. 2.2. Obviously many tested elements of interval type can be subjected to classification separately. After the procedure has been finished, elements of patterns which have undergone classification are sent to the beginning of the algorithm, to further avail of the next steps, following the addition of new elements of patterns.

Now – in reference to the end of the paragraph before the last – it remains to consider those elements whose values w_i were not counted among the $m_1^*, m_2^*, \ldots, m_J^*$ largest for particular patterns. Thus, for each of them the derivative $w_i^{'}$ is calculated. A method based on Newton's interpolation polynomial [9] is suggested here. If the element is "too new" and does not possess enough earlier values w_i, then the gaps should be filled with zeros, which prevents premature removal. Next for each separate class, the elements $w_i^{'}$ are sorted. The respective

$$qm_1^*, qm_2^*, \ldots, qm_J^* \tag{10}$$

elements of each pattern with the largest derivative values, on the additional requirement that the value is positive, go back to the beginning of the algorithm for further calculations carried out after the addition of new elements. If the number of elements with positive derivative is less than $qm_1^*, qm_2^*, \ldots, qm_J^*$, then the number of elements going back may be smaller (including even zero). The remaining elements are finally eliminated from the procedure. In the above notation q is a positive constant influencing the proportion of patterns' elements with little, but successively increasing meaning. The standard value of the parameter q can be proposed as $q = 0.1$.

The above procedure is repeated following the addition of new elements. Besides these elements – as has been mentioned earlier – for particular patterns respectively $m_1^*, m_2^*, \ldots, m_J^*$ elements of the greatest values w_i are taken, as well as up to $qm_1^*, qm_2^*, \ldots, qm_J^*$ elements of the greatest derivative $w_i^{'}$, so successively increasing its significance, most often due to the data drift.

4 Empirical Verification and Comparison

The correct functioning of the concept under investigation have been comprehensively verified numerically, and also compared with results obtained using procedures based on the support vector machine concept. Research was carried out for data sets in various configurations and with different properties, particularly with nonseparated classes, complex patterns, multimodal and consisting of detached subsets located alternately.

Comparative analysis was submitted to detailed investigations. Due to lack of an available algorithm dedicated to interval and drifting data, comparisons were made using two concepts based on the support vector machine method with proper modifications. The first one, intended for deterministic nonstationary data, presented in the article [3], was used with respect to midpoints of classified intervals. The second, from the publication [11], for stationary interval data was applied by removing the oldest elements from patterns and replacing them with the newest. And so, in relation to the first algorithm, the number of misclassifications was up to 20 % lower, while with the second even 50 % (treating no decision – a possibility there – as an unsatisfactory result). The advantage of the procedure presented in this paper was particularly visible in the case of steady drift – taking into account the fact that its idea is based on derivatives of a predictive nature, this observation is completely understandable.

The broader description of particular aspects and the analytical form of formulas can be found in the papers [Kulczycki and Kowalski, 2011, 2015].

References

1. Duda, R.O., Hart, P.E., Storck, D.G.: Pattern Classification. Wiley, New York (2001)
2. Kelley, C.T.: Iterative Methods for Optimization. SIAM, Philadelphia (1999)
3. Krasotkina, O.V., Mottl, V.V., Turkov, P.A.: Bayesian approach to the pattern recognition problem in nonstationary environment. In: Kuznetsov, S.O., Mandal, D.P., Kundu, M.K., Pal, S.K. (eds.) PReMI 2011. LNCS, vol. 6744, pp. 24–29. Springer, Heidelberg (2011)
4. Kulczycki, P.: Estymatory jadrowe w analizie systemowej. WNT, Warsaw (2005)
5. Kulczycki, P., Hryniewicz, O., Kacprzyk, J. (eds.): Techniki informacyjne w badaniach systemowych. WNT, Warsaw (2007)
6. Kulczycki, P., Kowalski, P.A.: Bayes classification of imprecise information of interval type. Control Cybern. **40**(1), 101–123 (2011)
7. Kulczycki, P., Kowalski, P.A.: Bayes classification for nonstationary patterns. Int. J. Comput. Methods **12**(2), 19 (2015). Article ID 1550008
8. Silverman, B.W.: Density Estimation for Statistics and Data Analysis. Chapman and Hall, London (1986)
9. Venter, G.: Review of optimization techniques. In: Blockley, R., Shyy, W. (eds.) Encyclopedia of Aerospace Engineering, pp. 5229–5238. Wiley, New York (2010)
10. Wand, M.P., Jones, M.C.: Kernel Smoothing. Chapman and Hall, London (1995)
11. Zhao,Y., He, Q., Chen, Q.: An interval set classification based on support vector machines. In: 2nd International Conference on Networking and Services, Silicon Valley, USA, 25–30 September 2005, pp. 81–86 (2005)
12. Zurada, J.: Introduction to Artificial Neural Neural Systems. West Publishing, St. Paul (1992)