Predicted Distribution Density Estimation for Streaming Data

Piotr Kulczycki[1,2], Tomasz Rybotycki[2]

[1] AGH University of Science and Technology, Faculty of Physics and Applied Computer Science
[2] Polish Academy of Sciences, Systems Research Institute

kulczycki@agh.edu.pl; kulczycki@ibspan.waw.pl

Abstract. Recent growth in interest concerning streaming data has been forced by the expansion of systems successively providing current measurements and information, which enables their ongoing, consecutive analysis. The subject of this research is the determination of a density function characterizing potentially changeable distribution of streaming data. Stationary and nonstationary conditions, as well as both appearing alternately, are allowed. Within the distribution-free procedure investigated here, when the data stream becomes nonstationary, the procedure begins to be supported by a forecasting apparatus. Atypical elements are also detected, after which the meaning of those connected with new tendencies strengthens, while diminishing elements weaken. The final result is an effective procedure, ready for use without studies and laborious research.

Keywords: Streaming Data, Distribution Density, Nonparametric Estimation, Distribution-Free Procedure, Prediction, Atypical Element (Outlier).

1 Introduction

Technological progress within the scope of numerical techniques has enabled the comprehensive analysis and exploration of data with different natures. Recently interest in a specific type, characterized by successive and unlimited inflow of sequential elements, named streaming data, has grown. In current practice, data of this type may be nonstationary (evolving in time), therefore, their characteristics are variable, which additionally makes all analysis considerably more difficult. Frequently the character of streaming data undergoes changes from stationary to nonstationary and *vice-versa*, implying further research challenges. Moreover, the nature of permanently incoming, often unverified data causes that they may also contain atypical elements, mostly as a result of errors of different kinds. Their automatic removal may, however, result in the elimination of valuable information about newly forming tendencies. Finally, effective analysis of streaming data fulfilling requirements of contemporary applications needs a range of significant factors, frequently absent in classic problems with an assumed finite dataset size, to be taken into account. This makes the analysis of streaming data extremely valuable from the applicational point of view, but also demanding from a research perspective.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_43

The subject of this paper is the synthesis of a procedure enabling the determination of distribution density of streaming data, both stationary and nonstationary, also with these both cases appearing alternately. The mathematical apparatus is based on the procedures of contemporary data analysis and mathematical statistics, allowing calculation of density without any assumption concerning the specific type of distribution. The particular elements, applied later during the creation of the procedure, will be presented in chapter 2. Successively, the nonparametric method of kernel estimators, procedure for atypical (rare) elements detection, the statistical test of stationarity, and elements of forecasting theory, will be presented in the subsequent four sections of this chapter. The concept of the procedure developed here for the predicted estimation of distribution density of streaming data is the subject of chapter 3. This procedure is modular in nature. In the succeeding four sections, the concepts of fixing the size of a reservoir, the outdatedness of its elements, introduction of forecasting methods, and detection of atypical elements strengthening the importance of those connected with newly arising tendencies and the weakening associated with disappearing trends, have been elaborated. In consequence, a procedure for determining the current distribution density of streaming data will be created, while in the case of discovery of nonstationarity, procedures for adaptation and forecasting are activated in order to effectively match to the changing environment. The calculation complexity of all algorithms used are linear and quadratic; the whole cycle of the calculations is enclosed within few seconds. The memory requirements do not exceed the typical capabilities of contemporary computer systems. The final conclusions and numerical results of the designed method will be briefly described in chapter 4.

The estimation of distribution density of streaming data is a current topic being studied and various methods have been applied, e.g. histogram [19] or wavelets [22], however, concepts based on kernel estimators dominate (see [5] for a rich bibliography) consisting of a proper selection of incoming elements [20], specialized clustering [9, 24], local approach [7], and using calculational intelligence methods, e.g. self-organizing maps [8]. Fundamental information concerning streaming data can be found in the recent books [3, 21].

2 Mathematical Preliminaries

2.1 Nonparametric Estimation, Kernel Estimators

Consider a set consisting of the m elements being the n-dimensional vectors with continuous attributes:

$$x_1, x_2, \ldots, x_m \in \mathbb{R}^n \quad . \tag{1}$$

The kernel estimator $\hat{f}: \mathbb{R}^n \to [0, \infty)$ of the density of a dataset (1) distribution, is defined then as [11, 23]:

$$\hat{f}(x) = \frac{1}{m} \sum_{i=1}^{m} K(x, x_i, h) \quad , \tag{2}$$

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_43

where after separation into coordinates

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad x_i = \begin{bmatrix} x_{i,1} \\ x_{i,2} \\ \vdots \\ x_{i,n} \end{bmatrix} \text{ for } i = 1,2,\dots,m \ , \quad h = \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_n \end{bmatrix} , \tag{3}$$

while the positive constants h_j are the so-called smoothing parameters; the kernel K will be used here in the product form:

$$K(x, x_i, h) = \prod_{j=1}^{n} \frac{1}{h_j} K_j \left(\frac{x_j - x_{i,j}}{h_j} \right) , \tag{4}$$

whereas the one-dimensional kernels $K_j: \mathbb{R} \to [0, \infty)$, for $j = 1,2,\dots,n$, are measurable with unit integral $\int_{\mathbb{R}} K_j(y)\, \mathrm{d}y = 1$, symmetrical, and non-increasing for $[0, \infty)$; (in consequence: non-decreasing for $(-\infty, 0]$). For the needs of further considerations, the definition (2) will be generalized to the weighted form:

$$\hat{f}(x) = \frac{1}{\sum_{i=1}^{m} w_i} \sum_{i=1}^{m} w_i \, K(x, x_i, h) \quad , \tag{5}$$

where the introduced parameters $w_i \geq 0$ are not all equal to 0. Assuming $w_i \equiv 1$, one simply obtains the formula (2). The kernel estimator allows us to calculate the density on the basis of the dataset (1) without any arbitrary assumption concerning the type of its distribution.

Generally, the selection of the kernel K_j form is practically meaningless and the user should, above all, take into account the properties of the desired estimator or/and computational aspects, beneficial for the application being worked out. In the following, the normal (Gauss) kernel

$$K_j(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2} \right) \tag{6}$$

will be applied, as generally used.

The fixing of the smoothing parameter h_j has significant meaning for quality of estimation. Fortunately, many suitable procedures for calculating its optimal value have been worked out. In particular, for simple unimodal distributions and in the preliminary phase of investigation, the normal concept is suggested. Then

$$h_j = \left(\frac{8\sqrt{\pi}}{3} \frac{W(K)}{U(K)^2} \frac{1}{m} \right)^{1/5} \hat{\sigma}_j \ , \tag{7}$$

where $W(K) = \int_{-\infty}^{\infty} K(y)^2 \, \mathrm{d}y$ and $U(K) = \int_{-\infty}^{\infty} y^2 K(y) \, \mathrm{d}y$. For the normal kernel (6) one has $W(K) = 1/2\sqrt{\pi}$ and $U(K) = 1$. The standard deviation estimator $\hat{\sigma}_j$, occurring above, can be calculated for the dataset (1) from the classic formula, potentially extended for the weighted form (5) as follows:

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_43

$$\hat{\sigma}_j^{\,2} = \frac{1}{m-1}\sum_{i=1}^{m}(w_i x_{i,j})^2 - \frac{1}{m(m-1)}\left(\sum_{i=1}^{m} w_i x_{i,j}\right)^2 \quad . \tag{8}$$

In other situations we propose testing the plug-in method [11, Section 3.1.5; 23, Section 3.6.1], where its degree should be equal to the number of separated factors (modes), but in practice not greater than 3; the value 2 can be treated as a standard. A generalization of this method to the weighted form should be made similarly to the formula (8).

In practice various modifications, generalizations and fitting properties of the estimator to specific realities can be applied, e.g. other algorithms for fixing the smoothing parameter, its adaptation, or boundary of the function $\hat{f}$ support. The procedure worked out in this paper has no limits in this range besides requirements regarding time and memory as well as excessive complexity of interpretation, which should be individually considered. The classic textbooks on kernel estimators constitute the monographs [11, 23]. The effective determination of distribution density enables comprehensive data analysis [12, 13] and various valuable applications [14, 15].

2.2 Detection of Atypical Elements (Outliers)

The determination of distribution density enables effective detection of atypical elements [2], which are understood here in the sense of rare occurrence. Unlike distance methods, one can then find atypical observations not only on the peripheries of the population, but in the case of multimodal distributions with wide-spreading segments, also those lying in between these segments, even if they are close to the 'center' of the set.

Consider the dataset (1) containing elements representative of the considered population. Based on the material from section 2.2, the kernel estimator (5) can be calculated. Then, consider also the set of its values for elements of the dataset (1), therefore

$$\hat{f}_{-1}(x_1), \hat{f}_{-2}(x_2), \dots, \hat{f}_{-m}(x_m) \quad , \tag{9}$$

where $\hat{f}_{-i}$ means the kernel estimator $\hat{f}$ calculated excluding the i-th element of the dataset. Next, define the number $r \in (0,1)$ determining the sensitivity of the procedure for identifying atypical elements. This number will simply determine the assumed proportion of atypical elements in relation to the total population; therefore, the ratio of the number of atypical elements to the sum of atypical and typical elements. In practice

$$r = 0.01, 0.05, 0.1 \tag{10}$$

is the most often used. Next, for the set (23) one can calculate the positional estimator for the quantile of the degree r given by the formula

$$\hat{q}_r = \begin{cases} s_1 & \text{for} \quad mr < 0.5 \\ (0.5 + i - mr)s_i + (0.5 - i + mr)s_{i+1} & \text{for} \quad mr \geq 0.5 \end{cases} , \tag{11}$$

where $i = \underline{int}(mr + 0.5)$, while $\underline{int}$ denotes an integral part of a number, and s_i is the i-th value in size of the set (9) after being sorted; thus

$$\{s_1, s_2, \dots, s_m\} = \left\{\hat{f}_{-1}(x_1), \hat{f}_{-2}(x_2), \dots, \hat{f}_{-m}(x_m)\right\} \tag{12}$$

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_43

with $s_1 \leq s_2 \leq \cdots \leq s_m$. Generally, there are no special recommendations concerning the choice of the sorting algorithm used for specifying set (12). However, let us interpret the definition (11), taking into account the values (10). So, it is enough to sort only the $i+1$ smallest values in the set (9), therefore, about 1-10% of its size. One can apply a simple algorithm that subsequently finds the $i+1$ smallest elements of the set (9).

Finally, if for a given tested element $\tilde{x} \in \mathbb{R}^n$, the condition $\hat{f}(\tilde{x}) \leq \hat{q}_r$ is fulfilled, then this element should be considered atypical; for the opposite $\hat{f}(\tilde{x}) > \hat{q}_r$ it is typical.

The details of the above method can be found in the paper [16]. A review of various methods of outlier detection is given in the monograph [2].

2.3 Testing of Stationarity, KPSS Test

Let the real time series $\{X_t\}_{t=1,2,\ldots}$ be given. The stationarity of the stochastic process, from which this series originate, will be verified using the KPSS test [18]. The hypothesis being tested here is the stationarity, with respect to the alternative hypothesis that the process is nonstationary. Generally, the KPSS test is applied in two options: without considering the trend and assuming its presence. Here, the first of them will be used – in the investigated procedure, each trend will be treated as a nonstationary factor.

The test statistics, calculated on the basis of T values $X_1, X_2, \ldots, X_T$ takes the form

$$KPSS = \frac{\sum_{t=1}^{T} S_t^2}{\hat{\sigma}_c^2} \quad , \tag{13}$$

where S_t denotes the partial sum of the residuals of mean-square approximation of the series by a constant function (the optimal value here is equal to the arithmetic mean), i.e.

$$S_t = \sum_{l=1}^{t} R_l \tag{14}$$

$$R_l = X_l - \bar{X} \quad \text{for } l = 1, 2, \ldots, t \tag{15}$$

$$\bar{X} = \frac{1}{T}\sum_{l=1}^{T} X_l \quad , \tag{16}$$

and $\hat{\sigma}_c$ means the consistent estimator of a standard deviation, given by the formulas

$$\hat{\sigma}_c^2 = T\sum_{l=1}^{T} R_l^2 + 2T\sum_{s=1}^{L} W(s,L)\sum_{z=s+1}^{T} R_z R_{z-s} \tag{17}$$

$$W(s,L) = 1 - \frac{s}{L+1} \tag{18}$$

$$L = int[4 \cdot (0.01\,T)^{\frac{1}{4}}] \quad , \tag{19}$$

where int means rounding to an integer. To avoid $0/0$, define $KPSS = 0$ for $T = 1$.

The critical set takes the right-hand form, while the critical values for critical levels equal respectively

critical level	0.1	0.05	0.025	0.01
critical value	0.347	0.463	0.574	0.739

. (20)

Based on the fuzzy approach, a quantity with values from the interval $[0, 1]$ will now

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_43

be defined, characterizing the "degree of nonstationarity" of the data stream under research. Namely, the function KPSS will be subjected to a linear transformation and then covered by the sigmoid function $sgm : \mathbb{R} \to (0, 1)$ given as

$$sgm(x) = \frac{1}{1+e^{-x}} \ . \tag{21}$$

After determining the transfomation parameters one obtains

$$sgmKPSS = sgm(0.995\ KPSS\ -\ 2.932) \ . \tag{22}$$

The coefficients of the linear transformation, appearing in the formula (22) were fixed heuristically such that the biggest, used in practice, critical value 0.739 is transformed into the highest critical level 0.1, and the smallest critical value 0.347 into the level lower by 30%. The last value has been fixed through the inspiration of automatic control practice, in particular the Ziegler-Nichols method of tuning PID controllers [6]. Namely, the integral quality index L_1 was minimized in a response to the unique step in the time series $\{X_t\}_{t=1,2,\dots}$. Such a value generally seems to be the most favorable (Section 4). Using the classic automatic control language, one then obtains a course without or with small over-regulations.

For purposes of the procedure investigated here, we fixed by the same method $T = 600$. Its increase results in sensitivity improving, however, at the expense in a slower reaction; a reduction brings opposite efects. Naturaly, in the initial t steps when $t < 600$ we should emloy as many elements as we have; therefore

$$T = \begin{cases} t & \text{when} \quad t < 600 \\ 600 & \text{when} \quad t \geq 600 \end{cases} . \tag{23}$$

For simple unimodal distributions, the value $T = 600$ can be reduced to 500.

In the multidimensional case

$$sgmKPSS = \max_{i=1,2,\dots,n} sgmKPSS_i \ , \tag{24}$$

where $sgmKPSS_i$ denotes the quantity $sgmKPSS$ given by the formula (22) for the i-th continuous attribute. The maximum norm assumed in the formula (24) allow the strongest, among particular attributes, nonstationarity to be identified. Note that using smooth functions in the above formulas will result in relatively mild fluctuations in time of the estimated density. For further considerations recall also that $0 < sgmKPSS < 1$.

2.4 Forecasting, Exponential Smoothing

If a nonstationarity is detected, the possibility appears of identification of a potential trend of the changes that have occurred, and regarding in the algorithm the values related with it. In this paper the exponential smoothing forecasting method [10] will be applied with the assumption of linear form of the trend. This method enables effective updating of the prediction model after receiving the subsequent value of the time series $\{X_t\}_{t=1,2,\dots}$.

The identified trend is assumed in the form $a_2 t + a_1$; denote the coefficients existing here in the form of a line vector, additionally denoting they dependence on t:

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_43

$$A_t = [a_{1,t}, a_{2,t}] \ . \tag{25}$$

The prognosis calculated at the moment t, with the anticipation $p \in \mathbb{N}\backslash\{0\}$ is given by:

$$X_t^p = A_t \begin{bmatrix} 1 \\ p \end{bmatrix} , \tag{26}$$

and then $a_{2,t}$ characterizes the velocity of changes.

In order to determine the matrix A_t, first define the following matrixes:

$$L = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \tag{27}$$

$$B_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} , \quad B_{t+1} = B_t + v^t \begin{bmatrix} 1 & -t \\ -t & t^2 \end{bmatrix} \quad \text{for } t = 1, 2, \ldots \tag{28}$$

$$b_1 = \begin{bmatrix} X_1 \\ 0 \end{bmatrix} , \quad b_{t+1} = vL^{-1}b_t + \begin{bmatrix} 1 \\ 0 \end{bmatrix} X_{t+1} \quad \text{for } t = 1, 2, \ldots , \tag{29}$$

where the parameter $v \in [0, 1]$ determines the intensity of adaptation of the forecasting model, fitting it to the changing reality. The possible increase in its value reduces the speed of reaction for forecasting errors, while decrease intensifies this reaction but threatens instability. The parameter v value will be determined in the following.

On the basis of the values successively obtained in the examined time series $X_1, X_2, \ldots$ one can calculate the matrixes (28)-(29), and finally

$$A_t = B_t^{-1} b_t \quad \text{for } t = 1, 2, \ldots \ . \tag{30}$$

Its second element $a_{2,t}$ will be used in the next chapter for the procedure designed there.

Detailed information on the exponential smoothing method can be found in the monograph [10] and the classic textbook [1].

3 Procedure for Predicted Distribution Density Estimation

The distribution density of streaming data will be determined using the moving window concept. Assume three parameters $m_{min}, m, m_0 \in \mathbb{N}\backslash\{0\}$ such that $m_{min} \leq m \leq m_0$. They represent a minimal, current and standard (in practice also maximal) number of elements, on the basis of which the kernel estimator $\hat{f}$ will be calculated. A reservoir consisting of m_0 last elements of the data stream under research will be created and successively updated. The elements of the reservoir are stored with the order of currency, from the newest x_1 to the oldest x_{m_0}.

The parameters m_{min}, m_0 are constant, while m changes depending on the current behavior of the data stream (see section 3.1). They are assigned weights resulting from the outdatedness with intensity depending on the nature of the stream under research (section 3.2). Following its characteristics, the procedure will be supported by forecasting methods (section 3.3). Atypical elements are accordingly amplified or reduced depending on whether it represents new or diminishing tendencies (section 3.4). Each of the above concepts reduces the estimation error, while these gains are independent and

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_43

cumulative (chapter 4).

3.1 Variable Reservoir Size

The reservoir size m, on the basis of which the kernel estimator is calculated, has a fundamental meaning for the quality of estimation. In the stationary case, when the characteristics of the data stream do not change, the higher value of this parameter gives more accurate results. However in the case of nonstationarity, smaller values of m enable us to more effectively keep up with changes.

We have assumed the following heuristic evaluation concerning the accuracy of the basic one-dimensional ($n = 1$) estimator:

$$\begin{aligned} &m = 100 \text{ - acceptable quality} \\ &m = 1{,}000 \text{ - good quality} \\ &m = 5{,}000 \text{ - very good quality .} \end{aligned} \tag{31}$$

The accuracies obtained experimentally for exemplary one-, two-, and three-modals distributions are shown in Tab. 1. Of course, all intermediate values as well as outside of the above range are possible. Enlarging the data dimension by one, requires about 4-fold increase in the size m to maintain quality.

Table 1. Accuracy of estimation of the exemplary distributions
one-modal N(0,1) with formula (7),
two-modals 60% N(0,1)+40% N(5,1) with plug-in of degree 2,
three-modals 30% N(-5,1)+40% N(0,1)+30%N (5,1) with plug-in of degree 3,
for L^1, L^2 and *sup* norms.

Accuracy	One-modal	Two-modals	Three-modals
$m = 50$	0.184, 0.300, 0.080	0.246, 0.340, 0.056	0.254, 0.350, 0.044
$m = 100$	0.141, 0.224, 0.062	0.187, 0.264, 0.046	0.205, 0.283, 0.037
$m = 1{,}000$	0.060, 0.098, 0.030	0.076, 0.112, 0.023	0.091, 0.127, 0.018
$m = 5{,}000$	0.032, 0.054, 0.018	0.041, 0.062, 0.013	0.050, 0.071, 0.011
$m = 10{,}000$	0.024, 0.041, 0.014	0.031, 0.047, 0.011	0.038, 0.055, 0.009

Therefore let m_0 constitute the assumed reservoir size for the conditions of stationarity, as well as m_{min} its minimal permissible level. Then define the value on the basis of which the kernel estimator $\hat{f}$ will be calculated as

$$m = \begin{cases} m_{min} & \text{when} \quad m_* < m_{min} \\ m_* & \text{when} \quad m_{min} \le m_* \le m_0 \\ m_0 & \text{when} \quad m_* > m_0 \end{cases}, \tag{32}$$

while

$$m_* = int(1.1\, m_0\, (1 - sgmKPSS)) \ , \tag{33}$$

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_43

where $sgmKPSS$ is given by the formula (22) substituting (13)-(19) and (21)-(24). Therefore, if one is dealing with a stationary process, then $sgmKPSS \cong 0$ and in consequence $m \cong m_0$. In turn, in the case of distinct nonstationarity $sgmKPSS \cong 1$, and then $m \cong m_{min}$. The intermediate values consecutively fluctuate in a continuous manner, as the term $sgmKPSS$ successively changes its value. Multiplication of the parameter m_0 by 1.1 (i.e. increase by 10%) in the formula (33) and then restriction m to m_0 in (33) eliminates possible fluctuations of m near m_0, resulting from the "tail" of the KPSS test statistics. To justify the level of 10%, see the determination of the parameters of the linear transformation in the equality (22). Note also that for purposes of the KPSS test (and only here) described in section 2.4, the elements should be provided in the opposite order, from the oldest with the index 1 to the newest with m_0.

Using the classic automatic control methods, in the basic one-dimensional case $n = 1$, the value $m_0 = 1{,}000$ has been obtained as a standard (compare the formula (31) and Tab. 1). In the case of complex, significantly multimodal distributions, it can be increased by 100 for each additional mode.

The parameter m_{min} value should be dependent on the biggest speed of changes. In particular, we propose

$$m_{min} = int(\tfrac{m_0}{10}) \ , \tag{34}$$

The value $m_{min} = 100$ can be treated as a standard. Such a value enables an effective tracking of changes not faster than $0.01\hat{\sigma}$ per step. For slower changes, the bottom boundary by m_{min} will be simply inactive. For faster alternations $m_{min} = 50$ is possible, however, runs can excessive fluctuating in time. Further decreasing of this value is not recommended (compare the formula (31) and Tab. 1).

3.2 Outdatedness

Particular elements used to calculate the kernel estimator will undergo outdatedness over time. This function will be performed by appropriate definition of values of the coefficients w_t, introduced in the definition (5). The linear formula will be applied

$$w_i^* = 2\left[1 - \tfrac{\alpha(i-1)}{m}\right] \quad \text{for } i = 1, 2, \dots, m \ , \tag{35}$$

where $\alpha \in [0, 1]$ specifies the intensity of outdatedness. In particular $\alpha = 0$ means its absence; all the reservoir elements then have the same weight. In contrast, if $\alpha = 1$, the weights successively decreased from 2 for the newest element with the index 1, to $2/m$ for the oldest with the index m, with the step $2/m$.

In the case of stationarity, it is worth assuming the value $\alpha = 0$, successively growing it as nonstationarity increases, to the maximum permissible value 1. As a natural consequence, it has been accepted that

$$\alpha = sgmKPSS \ , \tag{36}$$

where $sgmKPSS$ is given by the formula (22) substituting (13)-(19) and (21)-(24).

Finally, to take account of the above outdatedness procedure, for the purposes of

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_43

constructing the estimator (5) one should assume $w_i = w_i^*$ for $i = 1, 2, \ldots, m$, where w_i^* are given by the formulas (35)-(36).

3.3 Prediction

In the case when nonstationarity of the data stream under research results from a formed trend, it is worthwhile suitably introducing elements of forecasting methods, described in section 2.4, to the model.

For each new reservoir element x_i, sequentially, from the moment of its receiving, one builds a forecasting model, following the material presented in section 2.4, for which the consecutive quantities $\hat{f}(x_i)$, where $\hat{f}$ is the kernel estimator calculated on the basis of section 3.1 are treated as successive values of the observed time series. Thanks to this we have the vector (25) and in particular its second component $a_{2,t}$, which for the element x_i can be naturally denoted as $a_{2,t,i}$. Note also that the forecasting model is assigned to the specific element x_i and when its index i changes over time within the reservoir, this model moves with it for $i = 1, 2, \ldots, m_0$.

Now define the function representing changes of the kernel estimator (5). Let, therefore, for the fixed t, the function $g_t: \mathbb{R}^n \to \mathbb{R}$ be given by the formula

$$g_t(x) = \frac{1}{m}\sum_{i=1}^{m} a_{2,t,i} K(x, x_i, h) \quad , \tag{37}$$

where $a_{2,t,i}$ is the second element of the vector A_t (25), at the moment t, for the element x_i; the function K remains unchanged (4), while the parameter h value is the same as in the estimator $\hat{f}$ calculated on the base of section 3.1.

Introduce now the coefficients

$$w_i^{**} = 1 + \beta_i \, sgmKPSS \quad \text{for } i = 1, 2, \ldots, m \quad , \tag{38}$$

where $sgmKPSS$ is given by the formula (22) substituting (13)-(19) and (21)-(23), while $\beta_i \in [-1, 1]$. The presence in the above dependence (38) of the factor $sgmKPSS$ causes that in the case when the data stream is stationary, the coefficients w_i^{**} are close to 1, while in the nonstationary case the influence of the parameters β_i is manifested accordingly. Define their values as

$$\beta_i = \beta_0 \cdot \frac{g_t(x_i)}{\bar{g}_t} \quad \text{for } i = 1, 2, \ldots, m \quad , \tag{39}$$

where

$$\bar{g}_1 = 1 \quad , \quad \bar{g}_t = \max_{i=1,2,\ldots,m_t} |g_t(x_i)| \quad \text{for } t = 2, 3, \ldots \quad , \tag{40}$$

m_t denotes the size of the reservoir in the moment t and the constant $\beta_0 \in [0, 1]$ indicates the intensity of the forecasting function constructed herein. For the stationary conditions the value $\beta_0 = 0$ is natural. In the case of nonstationarity, initially consider $\beta_0 = 0.5$ as a standard. Generally the values from the range [1/3, 2/3] are satisfactory, while for the slow changes smaller values are preferable (also because of the function $\hat{f}$ fluidity over time) and for fast − bigger. For the nonstationarity, values smaller than 1/3 result in

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_43

too weak prediction, larger than 2/3 seem to be somewhat extreme (in particular, for $\beta_0 = 1$ some kernels could be removed, which unintentionally reduces a sample size assumed in section 3.1). Finally we propose

$$\beta_0 = \frac{2}{3}\, sgmKPSS \ . \tag{41}$$

Note that the condition $\beta_i \in [-1, 1]$ is fulfilled only with accuracy of determining the maximum of the function g_t only on the finite set $\{x_i\}$ as assumed in the formula (40). It has no meaning from the applicational point of view, because these parameters are multiplied in the dependence (38) by $sgmKPSS$, which is strictly less than 1, what in practice ensures the meaningful inequality $w_i^{**} \geq 0$.

The parameter v introduced in the formulas (28)-(29), defining the intensity of adaptation of the forecasting model, can be determined by the natural dependence:

$$v = 1 - \frac{1}{m} \ . \tag{42}$$

The intensity of adaptation of the forecasting model is therefore proportional to information provided by every new element of the data stream with the current reservoir size m.

Finally, if the prediction is used without the outdatedness procedure, it should be assumed that $w_i = w_i^{**}$, where w_i^{**} was defined above by the formulas (37)-(42), while if the both concepts are implemented, then $w_i = w_i^{*} \cdot w_i^{**}$, for $i = 1, 2, \ldots, m$.

3.4 Atypical (rare) elements

By calculating the distribution density, one can easily detect, separately in every moment t, atypical elements in the sense of rarely occurring. As previously, introduce the coefficients

$$w_i^{***} = 1 + \gamma_i\, sgmKPSS \quad \text{for } i = 1, 2, \ldots, m \ , \tag{43}$$

where $sgmKPSS$ is given by the formula (22) substituting (13)-(19), (21) and (23), and moreover $\gamma_i \in [-1, 1]$ are defined as

$$\gamma_i = \begin{cases} \frac{g_t(x_i)}{\bar{g}_t} & \text{when} \quad x_i \text{ is atypical element} \\ 0 & \text{when} \quad x_i \text{ is typical element} \end{cases} \quad \text{for } i = 1, 2, \ldots, m \ , \tag{44}$$

while g_t and $\bar{g}_t$ were specified in the previous section 3.3; see formulas (37) and (40). Thus, in the case of stationarity, the values of the coefficients w_i^{***} will be close to 1, while the data stream is nonstationary, the more amplified will be atypical elements which represents a rising tendency (thanks to $\gamma_i > 0$), and reduced recessive elements (due to $\gamma_i < 0$).

The procedure presented in section 2.3. can be used to identify atypical elements. Based on suggestions from the formula (10), the value of the parameter r, determining the procedure sensitivity, will be assumed as

$$r = 0.01 + 0.09\, sgmKPSS \ . \tag{45}$$

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_43

In the stationary case, a few (around 1%) atypical elements are specified, with an indication which are connected with new trends ($\gamma_i > 0$) and which with diminishing ($\gamma_i < 0$). It may be a valuable suggestion in the fundamental analysis of the results obtained. In turn, in conditions of strong nonstationarity, when $sgmKPSS \cong 1$, almost 10% of elements are recognized as atypical, which introduces an additional forecasting factor, as the importance of elements of increasing significance grows ($\gamma_i > 0$) and of decreasing meaning shrinks ($\gamma_i < 0$), which generally improves estimation quality. (If the results are given in the graphical form, it is worthwhile to mark on the graph the atypical elements by different color whose with positive γ_i values and other with negative.)

Finally, if the procedure described in sections 3.2-3.4 are used, then the coefficients introduced in the definition (5) should be taken in the form

$$w_i = w_i^* \cdot w_i^{**} \cdot w_i^{***} \quad \text{for } i = 1, 2, \dots, m \ . \tag{46}$$

If any of these procedures, outdatedness, prediction or detection of atypical elements, should be omitted, then the appropriate element w_i^*, w_i^{**} or w_i^{***} should be removed from the above formula. For clarity of interpretation, each of them varies in the same range from 0 to 2. All of them change continuously, which smooths fluctuations of the density $\hat{f}$.

4 Conclusion, Additional Aspects, and Numerical Evaluation

This paper investigates the concept of calculation of the current distribution density of the streaming data. The function $\hat{f}$ is defined by the formula (5), whereas standard quantities are associated with kernel estimators are presented in section 2.2, while the determination of the reservoir size and the construction of the coefficients w_i are given in the particular sections of chapter 3.

In the first step (to avoid zeros in the denominator in the formula (8)) it is arbitrarily assumed that $h = 1$. If in the initial steps, the number of the elements received is insufficient to fill the reservoir of the size obtained in section 3.1 or the average $\bar{g}_t$ from the formula (40), then this size should be reduced naturally to the number we have, similarly to the formula (23). Up to the step $t = m_{min}$, the results obtained are only indicative and do not give ground for further analysis at the assumed accuracy level. Only after the moment $t = m_0$ does the procedure work under appropriate sufficient stabilized conditions.

Similarly, during an increase in the parameter m value, one should add to the reservoir only new elements, even if they come slower than the m grow speed.

The procedure investigated has been comprehensively verified using both, simulated and real, data streams. For the basic illustration, consider a single continuous attribute, when the testing stochastic process is given in the form

$$X_t = p\ t + 0.6\ N(0,1) + 0.4\ N(5,1) \quad \text{for } t = 1, 2, \dots \ , \tag{47}$$

where $p\ t$ represents a deterministic trend, while

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_43

$$p = \begin{cases} 0.000{,}5 & \text{when} \quad t < 2{,}000 \\ 0.01 & \text{when} \quad 2{,}000 \le t < 5{,}000 \\ 0.005 & \text{when} \quad 5{,}000 \le t < 8{,}000 \\ +1 & \text{when} \quad t = 8{,}000 \\ 0 & \text{when} \quad 8{,}000 < t \le 10{,}000 \end{cases}, \tag{48}$$

whereas +1 denotes a unit step. Therefore during the initial period with very slow changes, a consolidation of the algorithm occurs, and then the data stream increases firstly very fast and then with medium speed, and finally, after a unit step, the process becomes stationary. Such changes in dynamics pose a big challenge for the worked out method.

Three performance indexes were used; averaged over time differences between the real density resulting from the formulas (47)-(48) and the estimator, in the senses of the L^1, L^2, sup norms. The presented results were obtained on the basis of 20 averaged runs. Each calculation cycle was performed in a few seconds.

In the stationary case the results were close (with accuracy to 1%) to those obtained simply on the basis of the last m_0 elements; the KPSS test correctly classified the stationarity of the data stream. The above simple strategy of the last m_0 elements was generally treated as the reference. The introduction of the variable m, as indicated in section 4.1, resulted in a decrease in the value of these indexes of 56%, 67% and 64%, respectively for particular indexes. The addition of outdatedness (section 4.2) improved the indexes by further 23%, 24% and 9%, while the addition of forecasting (section 4.3) reduces their values by 21%, 30% and 16%, and finally, adding atypical elements detection (section 4.4) by further 3%, 2% and 1%. In total, all four factors (sections 4.1-4.4) improved quality by about 63%, 83% and 65%.

Atypical elements detection does not introduce significant improvement of indexes. It works in those areas, in which the distribution density values are small and is also their influence of numerical indicators. This factor, however, captures even insignificant changes but often very important in the fundamental analysis of datasets, and also extraordinary errors and situations, not covered by the above research scheme. Note that forecasting as well as atypical elements detection work on characteristics which were already significantly improved by the modification of the reservoir size and outdatedness.

Similar results were obtained for multidimensional cases, also in the presence of categorical attributes [4], and with a noise correlated in time. Detailed experimental studies are the subject of the comprehensive paper [17], where a comparative analysis with the other methods quoted at the end of chapter 1, is also contained. Generally, the procedure presented here gives much better results, where the more clear and ambiguous formed trend is present.

Future researches will lead to the substitution of the removal of the oldest elements of the reservoir by sampling with probabilities dependent on the current size of the reservoir, outdatedness, prognosis and atypical elements detection presented in the successive sections of chapter 3 of this paper. It prevents the complete omission of phenomena manifested by elements older than the current reservoir size as it is in the case of the moving window method applied here. Thanks to forecasting, this goal can be achieved without, both qualitative and quantitative, deterioration of a quality.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_43

References

1. Abraham B., Ledolter J. (2005) *Statistical Methods for Forecasting*, Wiley.
2. Aggarwal C.C. (2013) *Outlier Analysis*, Springer.
3. Aggarwal C.C. – eds. (2007) *Data Streams. Models and Algorithms*, Springer.
4. Agresti A. (2002) *Categorical Data Analysis*, Wiley.
5. Amiri A., Dabo-Niang S. (2018) Density estimation over spatio-temporal data streams, *Econometrics and Statistics*, vol. 5, pp. 148-170.
6. Bequette B.W. (2010) Process Control: Modeling, Design, and Simulation, Prentice Hall.
7. Boedihardjo A.P., Lu C.-T., Chen F. (2015) Fast adaptive kernel density estimator for data streams, *Knowledge and Information Systems* vol. 42, pp. 285-317.
8. Cao Y., He H., Man H. (2012) SOMKE: Kernel Density Estimation Over Data Streams by Sequences of Self-Organizing Maps, *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, pp. 1254-1268.
9. Heinz C., Seeger B. (2008) Cluster Kernels: Resource-Aware Kernel Density Estimators over Streaming Data, *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, pp. 880-893.
10. Hyndman R.J., Koehler A., Ord J.K., Snyder R.D. (2009) *Forecasting with Exponential Smoothing: The State Space Approach*, Springer.
11. Kulczycki P. (2005) *Estymatory jadrowe w analizie systemowej*, WNT.
12. Kulczycki P. (2020) Methodically Unified Procedures for Outlier Detection, Clustering and Classification; in *Proceedings of the Future Technologies Conference* (FTC) 2019", Springer, pp. 460-474.
13. Kulczycki P., Franus K. (2021) Methodically Unified Procedures for a Conditional Approach to Outlier Detection, Clustering, and Classification, *Information Sciences*, in press.
14. Kulczycki P., Kacprzyk J., Kóczy L.., Mesiar R., Wisniewski R. (2020) *Information Technology, Systems Research, and Computational Physics*, Springer.
15. Kulczycki P., Korbicz J., Kacprzyk J. (2021) *Automatic Control, Robotics, and Information Processing*, Springer.
16. Kulczycki P., Kruszewski D. (2017) Identification of Atypical Elements by Transforming Task to Supervised Form with Fuzzy and Intuitionistic Fuzzy Evaluations, *Applied Soft Computing*, vol. 60, pp. 623-633.
17. Kulczycki P., Rybotycki T., Kus M. (2021) Predicted Kernel Estimator for Streaming Data with Continuous and Categorical Attributes, in press.
18. Kwiatkowski D., Phillips P.C.B., Schmidt P., Shin Y. (1992), Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root, *Journal of Econometrics*, vol. 54, pp. 159-178.
19. Muthukrishnan S., Strauss M., Zheng X. (2005) Workload-Optimal Histograms on Streams. Annual European Symposium, in *Algorithms – ESA 2005*, Springer, pp 734-745.
20. Qahtan A., Wang S., Zhang X. (2017) KDE-Track: An Efficient Dynamic Density Estimator for Data Streams, *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, pp. 642-655.
21. Rutkowski L., Jaworski M., Duda P. (2019) *Stream Data Mining: Algorithms and Their Probabilistic Properties*, Springer.
22. Trevino E.S.G., Hameed M.Z., Barria J.A. (2018) Data Stream Evolution Diagnosis Using Recursive Wavelet Density Estimators, *ACM Transactions on Knowledge Discovery from Data*, vol. 12, article 14.
23. Wand M.P., Jones M.C. (1995) *Kernel Smoothing*, Chapman and Hall.
24. Zhou Z., Matterson D.S. (2015) *Predicting Ambulance Demand: a Spatio-Temporal Kernel Approach*, arXiv:1507.00364.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_43