Using the Characteristic of Sample Data Distribution Law in the Machine Learning Tasks

Lyudmyla Kirichenko [1], Oksana Pichugina [2] and Volodymyr Kobziev [1]

1 Kharkiv National University of Radio Electronics, 14 Nauki ave., 61166, Kharkiv, Ukraine
[2] National Aerospace University "Kharkiv Aviation Institute", 17 Chkalova Street, 61070 Kharkiv, Ukraine

Abstract
The paper introduces a new characteristic of a random variable with a bounded support titled a distribution indicator, defined as the ratio of the squared support length to the random variable variance. It is shown that this characteristic of the whole family of random variables of the same distribution law type. Computer simulation of random time series signals and evaluation of the distribution indicator had demonstrated the possibility of detecting the distribution law of the series observations by analyzing this characteristic only.

Keywords
Sample random variable, probability distribution, distribution law indicator, time series clustering

1. Introduction

In the development of intelligent measuring, control and diagnostic instruments, the tasks of automatic recognition of objects, represented by a time series, arise. In the case when the initial data are random by nature, the recognition may be reduced to just in identifying the probability distribution type for these data. Respectively, the knowledge of the distribution allows choosing optimal algorithms for modeling or information processing [1]-[3], in particular for infocommunication technologies of information transfer [4]-[6].

The quantitative characteristic of the distribution law acquires special significance in tasks related to the time series classification or clustering by machine learning [7]-[8]. In this case, a quantitative assessment can be used as a feature that is an input for the classifier or a metric in the case of clustering [9]-[12].

So, in this paper, we pose a problem of experimental detecting of the probabilistic distribution law of sample data, including the time series, based on the calculation of a quantitative value.

2. Related Work

The classical method for identifying the distribution law of a random variable (RV) is the construction of a histogram and it analysis [13], [14]. When using this method, human participation is required to determine the distribution law. Another method is the analysis of such numerical characteristics as the mathematical expectation, variance, mode and median by evaluating the corresponding sample statistics, such as mean, sample variance, etc. This stage is followed by validating statistical hypothesis on a particular distribution law with a specific expectation and a standard deviation. Respectively, detecting a type of distribution such as normal, uniform, Poisson etc. might require significant effort since it is closely related with rather precise estimates of the distribution parameters. These methods are described in detail in [1], [15].

Most of the considered tests for determining the distribution law of the observed data have been implemented in modern software tools [16]-[18]. We introduce a new numerical characteristic of a continuous RV with a bounded support, i.e.

EMAIL: lyudmyla.kirichenko@nure.ua (A 1);
oksanapichugina1@gmail.com (A. 2);
volodymyr.kobziev@nure.ua (A. 3)
ORCID: 0000-0002-2780-7993 (A.1); 0000-0002-7099-8967 (A.2); 0000-0002-8303-1595 (A. 3)

defined on a bounded interval $[a, b]$ titled a distribution indicator.

It will be shown that this characteristic depends only on the RV distribution law, it can be considered as an identifier of the type of distribution. Also, it can be qualitatively assessed by a sample from the distribution by a sample identifier.

3. Indicator of Distribution Law

3.1. Distribution type indicator

By the type of distribution we will mean a family of RVs ξ with variance σ^2 and mathematical expectation m obtained from a RV ξ_0 with zero mathematical expectation and unit variance by linear transformation

$$\xi = k\xi_0 + b, \text{ where } k \in R^1_{>0},\ m \in R^1.$$

If ξ_0 has a probability density function (PDF) $p_{\xi_0}(x)$ and is defined on $[a_{\xi_0}, b_{\xi_0}]$, then the PDF of the transformed RV ξ determined on the interval $[a,b] = [a_\xi, b_\xi]$ will be equal to

$$p_\xi(x) = \frac{1}{\sigma} p_{\xi_0}\left(\frac{x-m}{\sigma}\right).$$

It is easy to see that, in order the conditions

$$M[\xi] = m,\ Var[\xi] = \sigma^2$$

hold given that $\xi = k\xi_0 + b$,

$$M[\xi_0] = 0,\ Var[\xi_0] = 1,$$

the coefficients k, b should be taken as follows: $k = \sigma,\ b = m$. Thus

$$\xi = \sigma\xi_0 + k.$$

Definition. The distribution type indicator $Id[\xi]$ is calculated by the formula

$$Id[\xi] = \frac{(b-a)^2}{Var[\xi]},$$

where $[a,b]$ is the RV support, and $Var[\xi]$ is its variance.

The following statement can be proved.

Proposition. The indicator $Id[\xi]$ is constant for all RVs of the same type.

Proof. Let us use the properties of the variance of the RV:

a) if the random variables X, Y are independent, then
$$Var[X+Y] = Var[X] + Var[Y];$$

b) if k is a constant, then
$$Var[kX] = k^2 Var[X];$$

c) the variance of the constant random variable C is equal to zero –
$$Var[C] = 0.$$

By assumption $M[\xi_0] = 0,\ Var[\xi_0] = 1$ and $\xi_0 \in [a_{\xi_0}, b_{\xi_0}]$. From that,

$$[a,b] = [\sigma a_{\xi_0} + m, \sigma b_{\xi_0} + m],$$

wherefrom $b - a = \sigma(b_{\xi_0} - a_{\xi_0})$. Hence it follows that $Id[\xi_0] = (b_{\xi_0} - a_{\xi_0})^2$. On the other side,

$$Id[\xi] = Id[\sigma\xi_0 + \mathrm{m}] = \frac{(b-a)^2}{Var[\xi]} =$$

$$= \frac{(\sigma(b_{\xi_0} - a_{\xi_0}))^2}{\sigma^2} = (b_{\xi_0} - a_{\xi_0})^2 = Id[\xi_0].$$

Thus, the proposition is true.

Since ξ was chosen in an arbitrary way in the family of RVs of the same type, which follows from the linear dependence of ξ on ξ_0, this statement is true for an arbitrary random one of the same type.

Remark. Theoretically, it is possible to expand the family of RVs with a certain probability type indicator $Id[\xi_0]$. For example, with the same support and having mirror-symmetric PDF will have the same distribution type indicator. In this case, in order to distinguish the RV from the mirror one, it is necessary to analyze other characteristics, such as mode, coefficient of asymmetry and kurtosis. However, in practice, such situations are quite few.

Table 1 shows the values of the distribution type indicator $Id[\xi]$ for three common probability distributions [19] with support $[a,b]$ and the PDF $p_\xi(x)$:

$p(x) = \frac{1}{b-a}$ - Uniform;

$p(x) = \frac{1}{\pi\sqrt{(x-a)(b-x)}}$ - Arcsine;

$p(x) = \frac{2}{b-a} - \frac{2}{(b-a)^2}|a+b-2x|$ - Symmetric triangular.

Table 1

$Id[\xi]$ **values for the probability distributions**

Distribution type	$Var[\xi]$	$Id[\xi]$
Uniform	$\frac{(b-a)^2}{12}$	12
Arcsine	$\frac{(b-a)^2}{8}$	8
Symmetric triangular	$\frac{(b-a)^2}{24}$	24

3.2. Practical application and evaluation of the distribution type indicator

In practical applications, the distribution type indicator can be used to identify the type (law) of RV distribution by values $X^0 = \{x_i\}_{i=\overline{1,n}}$ of a time series of length n [20].

As an estimate of the indicator $Id[\xi]$, we propose to use a numerical characteristic (further referred to as the sample distribution type indicator of the RV ξ)

$$\hat{Id}_n = \frac{(X_{\max} - X_{\min})^2}{S^2},,$$

where $X_{\max}$ and $X_{\min}$ are the maximum and minimum values of the series X^0, $(X_{\max} - X_{\min})$ is the range of the sample X^0, S^2 is an unbiased sample variance. The latter means that S^2 is an unbiased and consistent estimate of $Var[\xi]$, i.e.

$$\lim_{n\to\infty} S^2 = Var[\xi], M[S^2] = Var[\xi].$$

It follows that the estimate $\hat{Id}_n$ is also consistent:

$$\lim_{n\to\infty} \hat{Id}_n = \frac{\lim_{n\to\infty}(X_{\max} - X_{\min})^2}{\lim_{n\to\infty} S^2} = \frac{(b-a)^2}{Var[\xi]} = Id[\xi].$$

However, it can be biased and shifted to the left:

$$M\left(\hat{Id}_n\right) = \frac{M(X_{\max} - X_{\min})^2}{M\left(S^2\right)} = \frac{M(X_{\max} - X_{\min})^2}{Var[\xi]} \le$$

$$\le \frac{(b-a)^2}{D[\xi]} = Id[\xi],$$

whence it can be seen that it is unbiased estimate if and only if the range of the sample coincides with the range of the random variable $X_{\max} - X_{\min} = b - a$, otherwise it will be biased to the left.

For an experimental evaluating the distribution type indicator $Id[\xi]$, a numerical simulation of values of the RV $\xi \in [a,b]$, where $[a,b]$ is bounded was done for the probability distributions listed in Table 1.

The RV values were presented as a time series of length n. Figure 1 shows plots of dependence of $\hat{Id}_n$ on n. The figure shows the convergence of the estimate $\hat{Id}_n$ of the distribution type indicator to the corresponding theoretical value $Id[\xi]$.

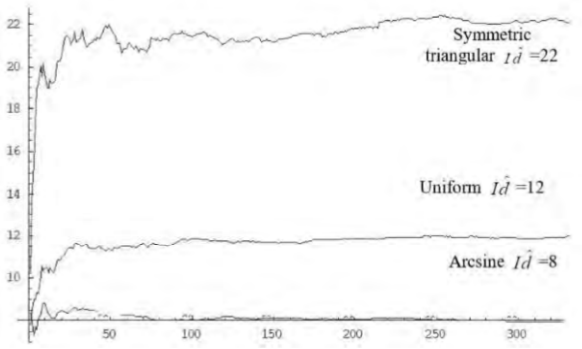


Figure 1: Dependence $\hat{Id}$ on n

The consistency of the estimate $\hat{Id}$ implies the mean square convergence of the series $\hat{Id}_i, i = 1,2,...$ to $Id[\xi]$. The average value of n for which the condition is satisfied

$$\frac{1}{n-n_1}\sum_{i=n_1}^{n}\left(Id[\xi] - \hat{Id}_i\right)^2 < \varepsilon$$

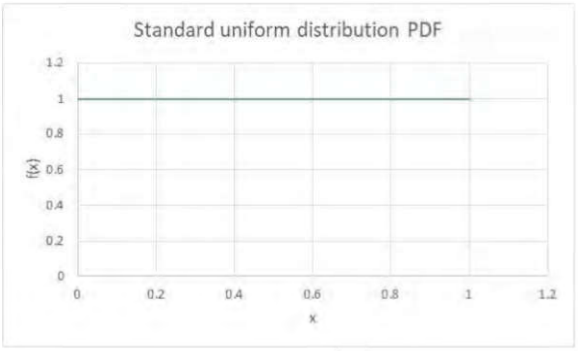
for an arbitrary $\varepsilon > 0$ on the interval $[n_1, n]$, $n_1 < n$ characterizes the rate of this convergence. The values n_1 calculated for the different values of ε are given in Table 2.

Table 2

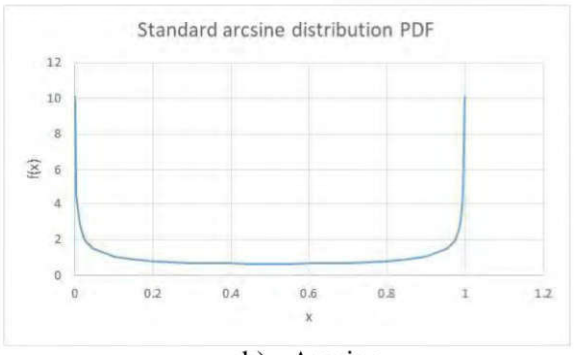
Convergence of the estimates

Probability distribution	ε =0,3	ε =0,5	ε =1
uniform	53	48	30
arcsine	19	16	12
Symmetric triangular	351	315	204

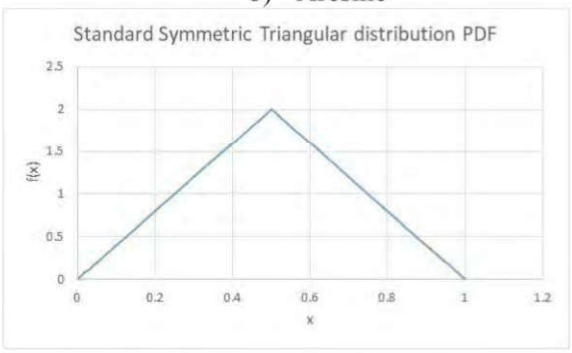
The convergence rate is closely related to the values of the PDF in the vicinity of the boundaries of the RV support $[a,b]$.



a) Uniform



b) Arcsine



c) Symmetric triangular

Figure 2: PDF of the standard distributions with support $[0,1]$

Figure 2 shows that the probability of falling out of numbers near the ends of the interval $[0,1]$ for the standard arcsine distribution is greater than that of the standard uniform distribution, so the range $\left(X_{\max} - X_{\min}\right)$ approaches its true value faster as n increases. For the symmetric triangular distribution, the probability of falling close to the bounds of the support $[0,1]$ is relatively small compared to the one of falling near the mean value 0.5, so the convergence to is slow. The values given in Table 2 can be considered as the size of the time series required for identifying the probability distribution type with the accuracy ε .

Also in the work, a numerical experiment was carried out on the clustering of model data. Three samples of independent random variables, which were discussed above, were generated. Clustering was performed using the k-means method; sample lengths were different: 20, 30, and 3 values. Figure 3 shows the results of clustering.

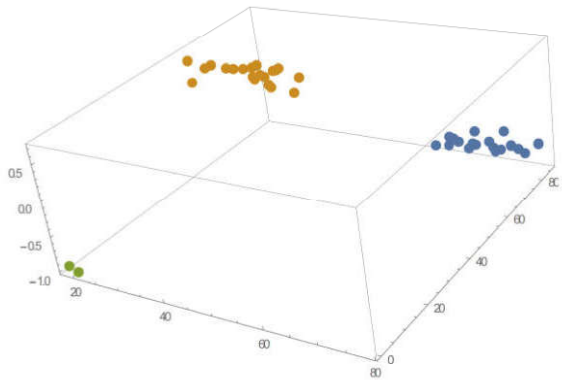


Figure 3: Results of clustering various independent random variables

4. Conclusions

The paper proposes and theoretically substantiates a method for identifying the probability distribution type of a continuous RD with a bounded support. The distribution type indicator $Id[\xi]$ introduced for this purpose provides a numerical characteristic of the distribution. Numerical studies have shown that the proposed method makes it possible to identify the distribution type of a RV using a statistical characteristic $\hat{Id}_n$ playing a role of an estimate of $Id[\xi]$ value and called the sample distribution indicator is the latter is evaluated on samples of a relatively small size. The contributions can be used for the development of measuring instruments, control, diagnostics and many tasks of object recognition, where speed and quality of the recognition is important.

5. References

[1] S. Brandt, Data Analysis: Statistical and Computational Methods for Scientists and Engineers. Cham Switzerland ; New York: Springer, 2014

[2] O. Pichugina, New Approaches to Modelling Covering Problems in Monitoring Optimization, in: 2019 IEEE

International Scientific-Practical Conference Problems of Infocommunications, Science and Technology (PICST), 2019, pp. 300–304. doi:10.1109/PICST47496.2019.9061386

[3] A. Srivastava and E. P. Klassen, Functional and Shape Data Analysis. New York: Springer-Verlag, 2016. doi: 10.1007/978-1-4939-4020-2

[4] L. Kirichenko and T. Radivilova, "Analyzes of the distributed system load with multifractal input data flows," 2017 14th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM), Lviv, 2017, pp. 260-264, doi: 10.1109/CADSM.2017.7916130.

[5] I. Ivanisenko, L. Kirichenko and T. Radivilova, "Investigation of multifractal properties of additive data stream," 2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP), Lviv, 2016, pp. 305-308, doi: 10.1109/DSMP.2016.7583564.

[6] O. Pichugina, N. Muravyova, Data Batch Processing: Modelling and Applications, in: 2020 IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S T), 2020, pp. 765–770. doi: 10.1109/PICST51311.2020.9467928.

[7] Esling, P., Agon, C.: Time series data mining. ACM Computing Surveys 46(1) (2012).

[8] Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., Muller, P. A.: Deep learning for time series classification: a review. Data Mining and Knowledge Discovery 33(4), 917-963 (2019).

[9] Ben D.: Feature-based time-series analysis, https://arxiv.org/abs/1709.08055

[10] Kirichenko, L., Radivilova, T., Bulakh, V. (2018) Machine Learning in Classification Time Series with Fractal Properties. Data 4(1):5. doi: 10.3390/data4010005.

[11] Trovero M. A., Leonard M. J. (2018) Time Series Feature Extraction. Paper SAS2020-2018. SAS Institute Inc 1-18. URL: https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2018/2020-2018.pdf

[12] Kirichenko, L., Radivilova, T., Tkachenko, Λ. (2019) Comparative Λnalysis of Noisy Time Series Clusterin. CEUR Workshop Proceedings 2362:184-196. http://ceur-ws.org/Vol-2362/paper17.pdf

[13] Greenwood, P.E.; Nikulin, M.S. (1996). *A guide to chi-squared testing*. New York: Wiley.

[14] Montgomery, Douglas C.; Runger, George C. (2003). Applied Statistics and Probability for Engineers . John Wiley & Sons, Inc. p. 104.

[15] G. Cowan, Statistical Data Analysis, 1st edition. Oxford : New York: Clarendon Press, 1998.

[16] Deisenroth, Marc Peter; Faisal, A. Aldo; Ong, Cheng Soon (2020). Mathematics for Machine Learning. Cambridge University Press. p. 181.

[17] P. S. P. Cowpertwait and A. V. Metcalfe, Introductory time series with R. New York : Springer-Verlag, 2009.

[18] C. M. Mastrangelo, J. R. Simpson, and D. C. Montgomery, 'Time Series Analysis', in Encyclopedia of Operations Research and Management Science, S. I. Gass and M. C. Fu, Eds. Springer US, 2013, pp. 1546–1552. doi: 10.1007/978-1-4419-1153-7_1045

[19] S. Kotz and J. R. V. Dorp, Beyond Beta: Other Continuous Families of Distributions with Bounded Support and Applications. Singapore ; Hackensack, NJ: World Scientific Publishing Company, 2004.

[20] Time Series Analysis: With Applications in R, 2nd edition. New York: Springer, 2010.