Considerations on the Information and Entropy of Ordinal Data

Iulian Petrila[1, 2 *], Florina Ungureanu[1], Vasile Manta[1]

[1] Faculty of Automatic Control and Computer Engineering, “Gheorghe Asachi” Technical University of Iasi,
Str. Dimitrie Mangeron, Nr. 27, 700050, Iasi, Romania
[2] Interdisciplinary Research Department and RAMTECH, “Alexandru Ioan Cuza” University of Iasi, Bd. Carol I, Nr. 11, Iasi, 700506, Romania

***Abstract* — This study attempts to establish methods for characterizing the complexity of ordinal data through the information and entropy parameters. In this respect, there were examined the methods for measuring the complexity of data with similar statistical characteristics and the parameters that can make the difference between them were established. For this purpose, the analysis was applied to three data sets with identical overall statistical characteristics but with different order of elements, respectively incremental, random and oscillatory incremental data. The analysis highlighted that it is necessary to include the information and entropy parameters for different variation orders of data elements. In this regard new parameters have been proposed, based on information and entropy expressions that describe, in an adequate way, the complexity of ordinal data.**

Keywords* — *information theory; data structures; data analysis; random processes; data mining

I. Introduction

Optimal description of data structures requires simple methods in quantitative evaluations of their complexity [1]. The complexity of data, which are not seen only as a collection, comes both from their values, and from their sequence. The entropy characterization of data is equivalent to the identification of patterns in data collections [2, 3]. This process may require some discretization operations in order to highlight the most significant attributes of the data [4]; the approach may allow a unitary description of the complexity and the entropy of various systems [5]. However, the entropy does not provide the relevant information regarding the order of elements in a data set, but only describes the information associated with the values of data elements [6-8]. In this respect, the extension of entropic description to ordinal data can lead to relevant description of the discrete data entropy with multiple applicative implications in various fields such as: nonlinear systems and analysis of signals [9], software systems analysis [10], data protection [11], computer cognition analysis [12], analysis of various human activities [13], data compression [14], multimedia codecs, hashing process optimizations etc.

This study aims to be an answer to the challenge of finding an entropic way to describe different data sets which contain the same elements but placed in different order. In the next sections, the general and specific characteristics of the relevant successive data sets are presented. A section dedicated to general statistical analysis of the reference data which includes also a spectral data characterization through Walsh-Hadamard transform is included in order to emphasize the essential statistical characteristics of the data. In the section dedicated to the entropy of data variations the differences in complexity of the analyzed data are marked. The sections dedicated to the entropy of all data successive and truncated variations present the relevant parameters which describe in an appropriate way the complexity of ordinal data.

II. Information and Entropy of Successive Data

The quantitative evaluation of some data from the information perspective implies their analysis through statistical parameters. Often, the usual statistical parameters do not count the correlations or order relationship between data elements because only the data elements values are evaluated [15]. According to the classical information theory [16, 17], an event $\mathrm{x_n}$ that occurs with the probability $\mathrm{p(x_n)}$ corresponds to an information $I(\mathrm{x_n}) = \log \mathrm{p(x_n)}^{-1} = -\log \mathrm{p(x_n)}$. For a collection of events, the total amount of information is defined by $I = \sum_n I(\mathrm{x_n}) = -\log \prod_n \mathrm{p(x_n)}$ as the information of all (independent) events and the Shannon entropy is defined as the mean of the information content by

$$H = \sum_n \mathrm{p(x_n)} I(\mathrm{p(x_n)}) = -\sum_n \mathrm{p(x_n)} \log_2 \mathrm{p(x_n)} \qquad (1)$$

Generally, the Shannon entropy cannot distinguish between an incremental and a random dataset, as shown the entropy analysis of different data sets like: the entire set of all pure octet incremental (PI) data represented in an octet

$$PI = \{n \mid n = 0..255\} = \{0,1, .., 255\} \qquad (2)$$

a set of pseudo-random (PR) data (octet representation)

$$\begin{cases} X_0 = 0, X_n = ((4p+1)X_{n-1} + 2k + 1) \bmod 256 \\ PR = \{X_0, X_1, .., X_{255}\} = \{0,1,6, .., 51\}, p = 1, k = 0 \end{cases} \qquad (3)$$

[*] **Corresponding author:** Iulian Petrila,
E-mail address: IulianPetrila@gmail.com

in which a linear generator is used to fulfill randomly entire byte range from 0 to 255 and a set of oscillatory incremental (OI) data

$$OI = \{n + (-1)^n \mid n = 0..255\} = \{1,0,3,2,5,..,255,254\} \quad (4)$$

Although the sequences of 256 values are different, as can be seen in Fig. 1, still, these data have the same Shannon entropy $H(PI) = H(PR) = H(OI) = 8$ and the same amount of information $I(PI) = I(PR) = I(OI) = 2048$. Rigorously, because the probabilities can be calculated for a large number of elements, the data can be extended through periodicity to a large number of bytes, with the same results.

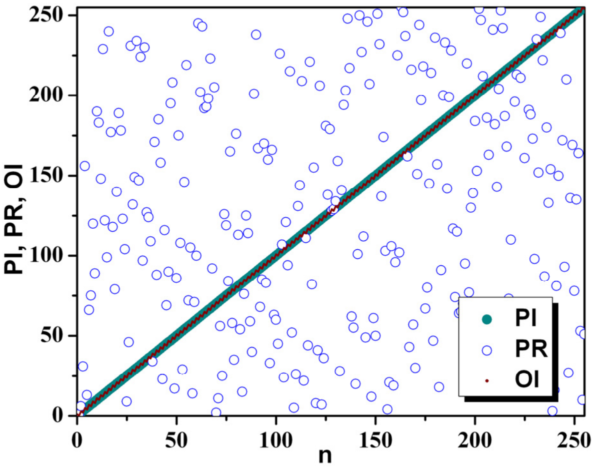

Fig. 1. Representation of Pure Incremental (PI), Pseudo-Random (PR) and Oscillatory Incremental (OI) octet data

These identical entropies are caused by the fact that the position data has no matter in Shannon entropy expression because it is taken into account only a data value and the probability of its presence in the list regardless the data element position in the list (the sequence of elements is not counted). It can be seen that, although each data set appears only once through the whole range values from 0 to 255, the information about the position in the data list (browsing sequence rule) is not counted by the entropy expression. Near to the entropy, from general statistical approach [18], the usual statistical parameters such as: average value of data $\mu = \frac{1}{n}\sum_{i=1..n} x_i$, standard deviation $\sigma = \sqrt{\frac{1}{n}\sum_{i=1..n}(x_i - \mu)^2}$ and variation coefficient $v = \sigma / \mu$ cannot perform a differentiation between the dataset because $\mu(PI) = \mu(PR) = \mu(OI) = 127.5$, $\sigma(PI) = \sigma(PR) = \sigma(OI) = 73.9$ and $v(PI) = v(PR) = v(OI) = 0.579$.

Therefore, how can be counted the differences in the information (or equivalent in the entropy) of these data sets? In the following sections we will try to establish a measure of the dataset complexity with these data to be properly distinguished.

Some differences between the analyzed data can be seen if we perform some transformations on them, such as Walsh-Hadamard defined by $W_{ij} = 2^{-m/2}(-1)^{i \cdot j}$, with $i \cdot j$ the bitwise dot product of the binary representations of the zero indexing numbers i and j, $n = 2^m$ is W matrix size, which in this form satisfies auto reverse relation $WW = I$ and Hadamard recursive relation $W^{(m)} = \frac{\sqrt{2}}{2}\begin{pmatrix} W^{(m-1)} & W^{(m-1)} \\ W^{(m-1)} & -W^{(m-1)} \end{pmatrix}$.

Thus, Walsh-Hadamard transforms of analyzed data become $PIW = W\,PI = \{2040, -8, 0, -32, .., 0\}$, $PRW = W\,PR = \{2040, -8, 0, 0, .., 0\}$ and $OIW = W\,OI = \{2040, 8, 0, -32, .., 0\}$; whose graphical representation can be seen in Fig. 2.

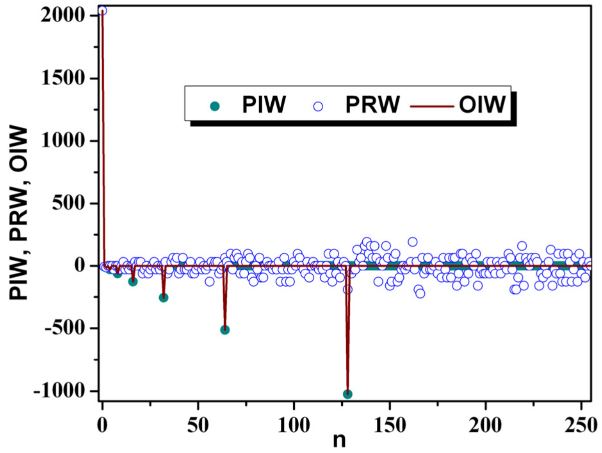

Fig. 2. Walsh-Hadamard transforms of Pure Incremental (PIW), Pseudo-Random (PRW) and Oscillatory Incremental (OIW) octet data.

Given the specificity of Walsh-Hadamard transform, the first term can be linked to the average value of data $PIW_1 = PRW_1 = OIW_1 = 2^{m/2}\mu$. The incremental difference between PI and OI data is metered by the second terms of Walsh-Hadamard transform ($PIW_2 = -OIW_2$). However, the entropy characteristics of transformed data ($I(PIW) = I(OIW) = 72.05$, $I(PRW) = 89.54$, $H(PIW) = H(OIW) = 0.331$, $H(PRW) = 3.277$) do not provide more relevant information about the specific differences of the analyzed data.

III. Information and Entropy of Data Variations

Instead of focusing on the entropy of data element values, the differences in complexity can be counted by the changes in the successive data elements. Practically, it is required to define the entropy of successive somehow related to the correlation order to build a relationship between data. A simple approach would be to take into account the variations of data elements (if it can define such an operation). In the case of (2)-(4) dataset, the set of variations data becomes: $\Delta PI = \{1,1,..,1\}$, $\Delta OI = \{-1,3,-1,3,..,-1\}$ and

$\Delta PR = \{1, 5, 25, 125, -143, 53, .., 41\}$; which can be seen in Fig. 3.

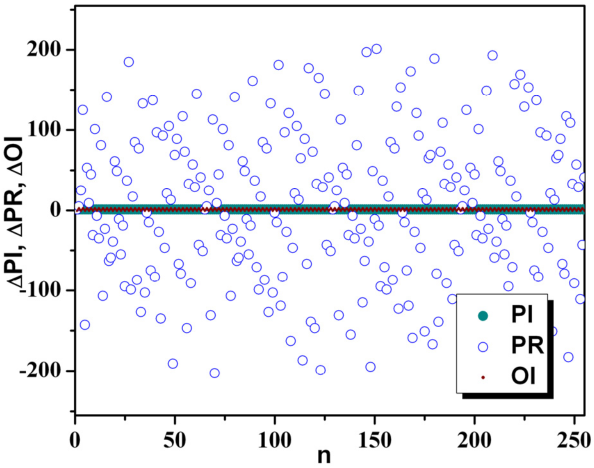

Fig. 3. Variation of Pure Incremental (ΔPI), Pseudo-Random (ΔPR) and Oscillatory Incremental (ΔOI) octet data.

The expression of entropy that can discard between correlated and non-correlated data may link at least the information between two neighboring elements. For instance, we can define the entropy of variations of successive elements by

$$H_{\Delta} = -\sum_{n} \mathrm{p}(\Delta \mathrm{x}_{\mathrm{n}}) \log_2 \mathrm{p}(\Delta \mathrm{x}_{\mathrm{n}}) \qquad (5)$$

where $\Delta \mathrm{x}_{\mathrm{n}} = \mathrm{x}_{\mathrm{n}} - \mathrm{x}_{\mathrm{n-1}}$ is the algebraic variation (if can be defined for the $\{\mathrm{x}_{\mathrm{n}}\}$ dataset). In this case, the differences between the entropies of variations are obviously: $H_{\Delta}(PI) = 0.00$, $H_{\Delta}(PR) = 6.52$ and $H_{\Delta}(OI) = 1.00$. Entropy values are in concordance with the complexity or with the degree of data randomness.

TABLE I. Statistical parameters of Pure Incremental (PI), Pseudo-Random (PR) and Oscillatory Incremental (OI) data variations

Data	Statistical parameters				
	μ	σ	ν	I	H
ΔPI	1.00	0.00	0.00	0.00	0.00
ΔPR	0.20	93.3	466.6	698. 2	6.52
ΔOI	1.00	2.00	2.00	2.00	1.00

It should be noted that for various values of p and k in (3) it is obtained byte random data with different entropy of variations $H_{\Delta}(PR)$ and this parameter can be used to identify the optimal parameters p and k for the pseudo-random generators defined through the expression (3). However, given that PR numbers were generated through an analytical expression, the entropy has no maximum value ($H_{\Delta}(PR) = 6.52 < 8$) but its high values show however that its complexity is higher than the incremental expressions (2) and (4). Thus, it can be considered that the proposed expression (5) can be a measured for the degree of dataset randomization.

IV. Information and Entropy of All Successive Variations

It is clear that the order of data elements in a data set must be taken into account for a relevant description of the successive data. In this respect, it can define the variation of a particular order m of the data D by the recurrence relation $\Delta^{(m)}(D) = \Delta(\Delta^{(m-1)}(D))$ with $\Delta^{(0)}(D) = D$. The total amount of information for an arbitrary order of variations m is given by

$$I_{\Delta^{(m)}} = -\sum_{n} \log_2 \mathrm{p}(\Delta^{(m)} \mathrm{x}_{\mathrm{n}}) \qquad (6)$$

The total information $I_{\Delta^{(m)}}$ decreases with increasing of order variation m, as can be seen in Fig. 4 for data taken in the analysis.

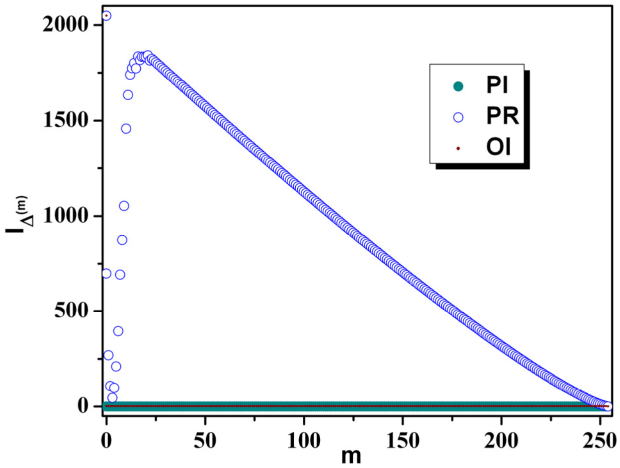

Fig. 4. Characteristics of the total information $I_{\Delta^{(m)}}$ versus variation order m for Pure Incremental (PI), Pseudo-Random (PR) and Oscillatory Incremental (OI) data

However, in order to discern between data complexity, we define a parameter that includes the total information associated with any order of variations by

$$S_I^{(\Delta)} = \sum_{m=0}^{N-1} I_{\Delta^{(m)}} \;. \qquad (7)$$

Through this parameter a relevant quantitative differentiation of analyzed data can be performed because $S_I^{(\Delta)}(PI) = 2048$, $S_I^{(\Delta)}(PR) = 222448.37$ and $S_I^{(\Delta)}(OI) = 2556.34$. The entropy of variations of order m can be evaluated in the same manner through

$$H_{\Delta^{(m)}} = -\sum_{n} \mathrm{p}(\Delta^{(m)} \mathrm{x}_{\mathrm{n}}) \log_2 \mathrm{p}(\Delta^{(m)} \mathrm{x}_{\mathrm{n}}) \qquad (8)$$

The entropy characteristics for different variation order m are represented in Fig. 5.

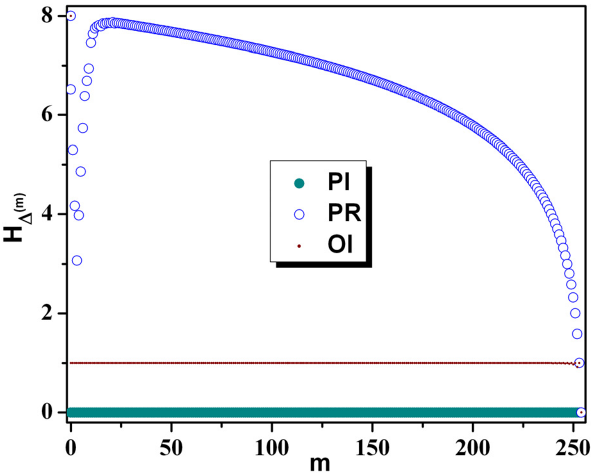

Fig. 5. Characteristics of the entropy $H_{\Delta^{(m)}}$ versus variation order m for Pure Incremental (PI), Pseudo-Random (PR) and Oscillatory Incremental (OI) data

Even for small variations orders m the entropy $H_{\Delta^{(m)}}$ may have some fluctuations, but from a particular order, m the entropy characteristics are clearly separable. The cumulative entropy for all orders m defined by

$$S_H^{(\Delta)} = \sum_{m=0}^{N-1} H_{\Delta^{(m)}} \tag{9}$$

can be used as a differentiation parameter for analyzed data. Thus, the pure incremental data have the lowest complexity $S_H^{(\Delta)}(PI) = 8$, the pseudo-random data have the highest complexity $S_H^{(\Delta)}(PR) = 1656.42$ and the oscillatory incremental data have an intermediate complexity $S_H^{(\Delta)}(OI) = 261.83$.

V. Information and Entropy of Truncated Successive Variations

For large amount of data, the successive variations can generate large numbers of states, which will be difficult to be evaluated. Moreover, when the data are disturbed by external factors (some noise for instance) the number of distinct states increases and the evaluation of the entropy becomes problematical and less relevant [19]. In this case, in order to diminish the variety of states (variations or events) and to perform a noise filtration [20], it can analyze the data variation through threshold limit by

$$\Delta_\varepsilon x = \begin{Bmatrix} 0, \, if \left| x_n - x_{n-1} \right| < \varepsilon \\ x_n - x_{n-1}, \text{otherwise} \end{Bmatrix} \tag{10}$$

where through ε it reduces the number of variations and it can provide different levels of filtering. The total amount of information for ε limited variations can be calculated by

$$I^{(\varepsilon)}{}_\Delta = -\sum_n \log_2 p(\Delta_\varepsilon x_n) \tag{11}$$

As expected, the total information $I^{(\varepsilon)}{}_\Delta$ decreases with increasing of threshold ε, as represented in Fig. 6.

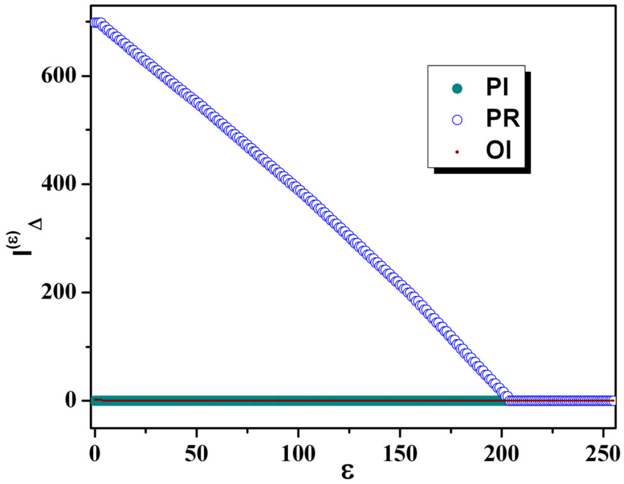

Fig. 6. Characteristics of the total information $I^{(\varepsilon)}{}_\Delta$ versus variation limit ε for Pure Incremental (PI), Pseudo-Random (PR) and Oscillatory Incremental (OI) data

Different variation limit ε can capture different information layers and their summation

$$S_{I^{(\varepsilon)}_\Delta} = \sum_\varepsilon I^{(\varepsilon)}{}_\Delta \tag{12}$$

provides quantitative relevant integrative information of analyzed data. The parameter defined by (12) gives a relevant hierarchy of data complexity because $S_{I^{(\varepsilon)}_\Delta}(PI) = 0$, $S_{I^{(\varepsilon)}_\Delta}(PR) = 76359.72$ and $S_{I^{(\varepsilon)}_\Delta}(OI) = 8$.

Similarly, we can define the entropy under variation limit ε by

$$H^{(\varepsilon)}{}_\Delta = -\sum_n p(\Delta_\varepsilon x_n) \log_2 p(\Delta_\varepsilon x_n) \tag{13}$$

this also decreases with increasing of ε (see Fig. 7).

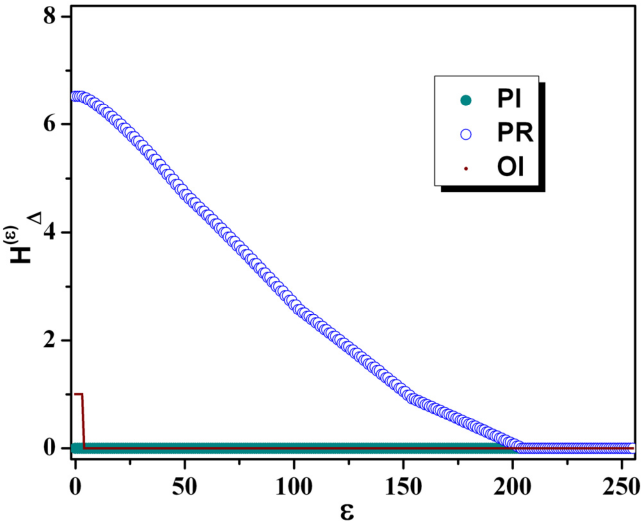

Fig. 7. Characteristics of the entropy $H^{(\varepsilon)}{}_\Delta$ versus variation limit ε for Pure Incremental (PI), Pseudo-Random (PR) and Oscillatory Incremental (OI) data

The cumulative entropy under variation limit ε can be defined by

$$S_{H^{(\varepsilon)}_{\Delta}} = \sum_{\varepsilon} H^{(\varepsilon)}{}_{\Delta} \tag{14}$$

which, for analyzed data, become $S_{H^{(\varepsilon)}_{\Delta}}(PI) = 0$, $S_{H^{(\varepsilon)}_{\Delta}}(PR) = 598.35$ and $S_{H^{(\varepsilon)}_{\Delta}}(OI) = 4$, confirming data complexity hierarchy.

One may mention that, for a more depth analysis, a combination between different order of variation m and different variation limit ε can be used. Also, for general data elements, the operations defined above can be used at each element sublevel and the contribution of each element can be accounted through an appropriate norm associated to data elements.

VI. Conclusions

The complexity of ordinal data can be described through the information and entropy parameters. The analysis of different data sets which contain the same elements placed in different orders highlighted that it is required to include the information and entropy parameters for both: different variations orders of data elements and truncated successive variations. In this respect the proposed information and entropy parameters have been performed the differentiation in a relevant of incremental, random and oscillatory incremental data. The proposed parameters are useful in determining the correlation between the ordinal data elements or in measuring the complexity of data.

Acknowledgment

This research was supported by TRAVEE PN-II-PT-PCCA-2013-4-1580 project of Romanian National Authority for Scientific Research (UEFISCDI).

References

[1] R. Gray, “Entropy and information theory”, Springer, New York, 2011.

[2] M. Baena-Garcia, and R. Morales-Bueno, Mining interestingness measures for string pattern mining, Knowledge-Based Systems, vol. 25, 2012, pp. 45-50.

[3] M. Ye, X. Wu, X. Hu, and D. Hu, “Anonymizing classification data using rough set theory”, Knowledge-Based Systems, vol. 43, 2013, pp. 82-94.

[4] J. W. Grzymala-Busse, “Discretization Based on Entropy and Multiple Scanning”, Entropy, vol. 15, 2013, pp. 1486-1502.

[5] D. Ke, “Unifying Complexity and Information”, Scientific Reports, vol. 3, 2013, p.1585.

[6] N. Nagaraj, K. Balasubramanian, and S. Dey, “A new complexity measure for time series analysis and classification”, The European Physical Journal Special Topics, vol. 222, 2013, pp. 847-860.

[7] F. Musella, “A PC algorithm variation for ordinal variables”, Computational Statistics, vol. 28, 2013, pp. 2749-2759.

[8] V. A. Unakafova and K. Keller, “Efficiently Measuring Complexity on the Basis of Real-World Data”, Entropy, vol. 15, 2013, pp. 4392-4415.

[9] J. Chen, H. Li, Y. Wang, R. Xie, and X. Liu, “A Novel Approach to Extracting Casing Status Features Using Data Mining”, Entropy, vol. 16, 2014, pp. 389-404.

[10] G. Canfora, L. Cerulo, M. Cimitile, and. M. Di Penta, “How changes affect software entropy: an empirical study”, Empirical Software Engineering, vol. 19, 2014, pp. 1-38.

[11] B. Espinoza and G. Smith, “Min-entropy as a resource”, Information and Computation, vol. 226, 2013, pp. 57-75.

[12] S.K. Pal and R. Banerjee, “Context granulation and subjective-information quantification”, Theoretical Computer Science, vol. 488, 2013, pp. 2-14.

[13] C. Moreira and A. Wichert, “Finding academic experts on a multisensor approach using Shannon’s entropy”, Expert Systems with Applications, vol. 40, 2013, pp. 5740-5754.

[14] Z. Ren, P. Su, and J. Ma, “Information Content Compression and Zero-Order Elimination of Computer-Generated Hologram Based on Discrete Cosine Transform”, Optical Review, vol. 20, 2013, pp. 469-473.

[15] S. Cagnone and P. Monari, “Latent variable models for ordinal data by using the adaptive quadrature approximation”, Computational Statistics, vol. 28, 2013, pp. 597-619.

[16] T. Cover and J. Thomas, “Elements of Information Theory”, John Wiley&Sons, Hoboken, New Jersey, 2006.

[17] P. Seibt, “Algorithmic Information Theory - Mathematics of Digital Information Processing”, Springer-Verlag Berlin Heidelberg, 2006.

[18] V.J. Rayward-Smith, “Statistics to measure correlation for data mining applications”, Computational Statistics & Data Analysis, vol. 51, 2007, pp. 3968-3982.

[19] U. Bodenhofer, M. Krone, F. Klawonn, “Testing noisy numerical data for monotonic association”, Information Sciences 245 (2013) 21-37.

[20] J.A. Saez, J. Luengo, F. Herrera, “Predicting noise filtering efficacy with data complexity measures for nearest neighbor classification”, Pattern Recognition 46 (2013) 355-364.