ISSN 2007-9737

# Visualization of Similarity Measures for Binary Data and 2 x 2 Tables

Ildar Z. Batyrshin[1], Nailya Kubysheva[1], Valery Solovyev[2], Luis A. Villa-Vargas[1]

[1] Instituto Politécnico Nacional, México, Centro de Investigación en Computación (CIC), Mexico

[2] Kazan Federal University, Kazan, Russia

batyr1@gmail.com, maki.solovyev@mail.ru

**Abstract.** We propose methods of 3D visualization of the main similarity measures for binary data and 2 x 2 tables. We present the shapes of Jaccard, Dice, Sokal & Sneath, Roger & Tanimoto and other similarity measures. Such visualization of the similarity measures gives the direct, visual, method of comparison of these measures and helps to understand the similarity and the difference between them. Based on the visualization of the two known parametric families of similarity measures the paper proposes the new parametric family of measures generalizing these two families and giving the possibility to construct similarity measures occupying the intermediate position between them.

**Keywords.** Similarity measure, binary variable, contingency table, visualization.

## 1 Introduction

Similarity measures have numerous applications in computational linguistics, ecology, medicine, biology, social sciences, etc. They play important role in pattern recognition, machine learning, classification and statistics [1, 5-7, 10, 12-14, 16, 18, 19]. Dozens of similarity (or dissimilarity) measures for binary data have been proposed and the problem of their comparison and selection for specific application is studied in many works [2-10, 15, 17-20]. In different papers, such measures are referred to as association coefficients, similarity coefficients, resemblance measures etc. Different approaches for comparing similarity measures are based on: similarity of the properties of these measures, similarity of formulas, possibility of transformation of one measure into another one, ordering of the measures, distance between them etc. [2, 3, 6-12,18-20].

To the best of our knowledge, there are not works on 3D visualization of the binary similarity measures. Such visualization can be useful for comparing the shapes of similarity measures and selecting measure more suitable for specific applications. The paper proposes the methods of 3D visualization of the most popular similarity measures used for binary data and 2 x 2 tables. Such visualization of similarity measures gives the direct, visual, method of comparison of these measures and can help to understand the similarity and the difference between them.

Several authors have proposed different parametric families of similarity and dissimilarity measures [3, 9, 19, 20]. Based on the visualization of the two known parametric families of similarity measures the paper proposes the new parametric family of the similarity measures generalizing these two families and giving possibility to construct similarity measures occupying intermediate position between them.

The paper has the following structure. Section 2 considers some basic definition related with the similarity measures for binary data and describes the most popular similarity measures. Section 3 proposes the methods of 3D visualization of similarity measures for binary data and visualize the most popular measures. Section 4 proposes the new parametric family of similarity measures. The last section contains discussion and conclusion.

## 2 Basic Definitions

Consider objects described by *n* binary attributes, descriptors or properties. The object *x* is coded by the vector $x = (x_1, \dots, x_n)$ of *n* attribute values such that $x_k = 1$ if the object possesses the property $k$ and $x_k = 0$ otherwise. Such data are called also presence/absence data [8, 11]. For any two objects $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$ the following four numbers are calculated:

- $a$ is the number of attributes such that $x_k = 1$, $y_k = 1$;
- $b$ is the number of attributes such that $x_k = 1$, $y_k = 0$;
- $c$ is the number of attributes such that $x_k = 0$, $y_k = 1$;
- $d$ is the number of attributes such that $x_k = 0$, $y_k = 0$.

The numbers *a* and *d* also referred to as the numbers of positive and negative matches, correspondingly [9, 17].

Note that the following is fulfilled for these four number:

$$a + b + c + d = n, \tag{1}$$

where *n* is the number of binary attributes. These four numbers are represented in Table 1 also known as 2×2 contingency table [1].

Below there are presented some popular similarity measures defined for such tables [4, 5, 10].

Jaccard (1908):

$$S_J(x, y) = \frac{a}{a+b+c}. \tag{2}$$

Dice (1945), Czekanowski (1913), Sorensen (1948):

$$S_{CDS}(x, y) = \frac{2a}{2a+b+c}. \tag{3}$$

Sokal & Sneath (1963):

$$S_{SS-\mathrm{I}}(x, y) = \frac{a}{a+2b+2c}. \tag{4}$$

Sokal & Michener (1958) or “simple matching”:

**Table 1.** 2 x 2 contingency table

| | | *y* | |
|---|---|---|---|
| | | **1** | **0** |
| **x** | **1** | *a* | *b* |
| | **0** | *c* | *d* |

$$S_{SM}(x, y) = \frac{a+d}{a+b+c+d} \tag{5}$$

Rogers & Tanimoto (1960):

$$S_{RT}(x, y) = \frac{a+d}{a+2b+2c+d} \tag{6}$$

Sokal & Sneath (1963):

$$S_{SS-\mathrm{II}}(x, y) = \frac{2a+2d}{2a+b+c+2d} \tag{7}$$

Rassel & Rao (1940):

$$S_{RR}(x, y) = \frac{a}{a+b+c+d} \tag{8}$$

Faith (1983):

$$S_F(x, y) = \frac{a+0.5d}{a+b+c+d} \tag{9}$$

## 3 Visualization of Similarity Measures

Let us consider parametric families of similarity measures that include the known similarity measures as particular cases [9, 20]. The similarity measures (2)-(4) can be generalized as follows:

$$T_\theta = \frac{a}{a+\theta(b+c)}, \tag{10}$$

where $\theta$ is some positive real number. The similarity measures (5)-(7) can be considered as the particular cases of the following parametric family of functions:

$$S_\theta = \frac{a+d}{a+d+\theta(b+c)}, \tag{11}$$

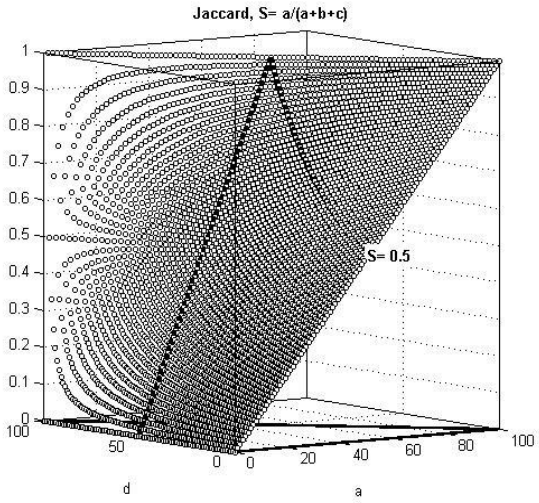


**Fig. 1(a).** Jaccard similarity measure (view 1)

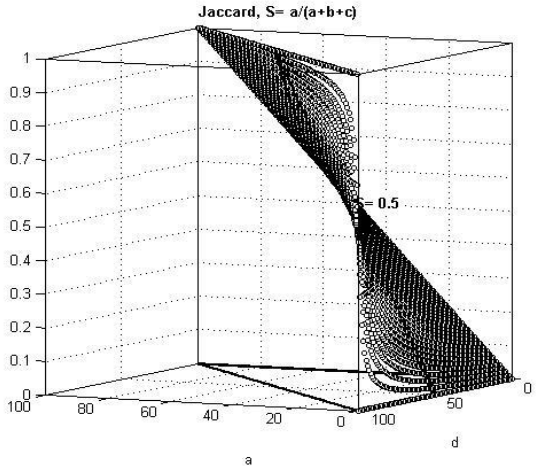


**Fig. 1(b).** Jaccard similarity measure (view 2)

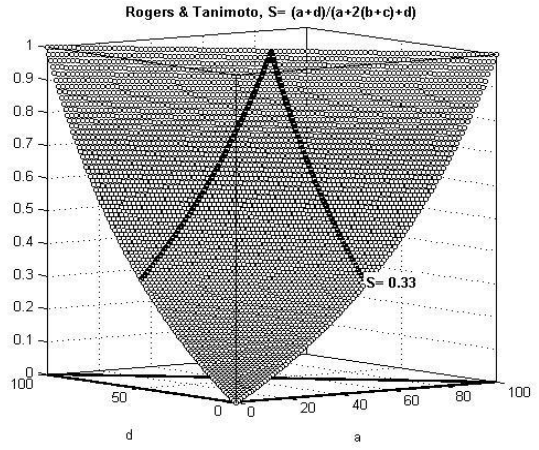


**Fig. 2(a).** Rogers & Tanimoto similarity measure (view 1)

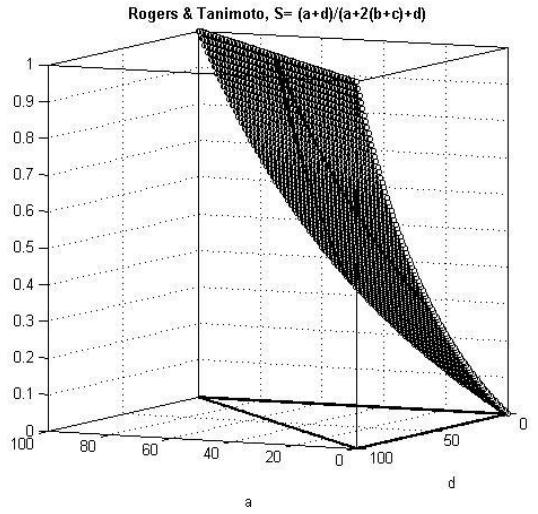


**Fig. 2(b).** Rogers & Tanimoto similarity measure (view 2)

For us it will be more convenient to use the following notation of these parametric families of similarity measures:

$$S_{(a)}(x,y) = \frac{a}{a+t(b+c)}, \qquad (12)$$

$$S_{(a+d)}(x,y) = \frac{a+d}{a+d+t(b+c)}, \qquad (13)$$

where $t$ is some positive real number. The similarity measures (2)-(4) are obtained from (12) for the parameter values $t$= 1, 0.5, 2, correspondingly. The similarity measures (5)-(7) are obtained from (13)

for the parameter values *t*= 1, 2, 0.5, correspondingly. Taking into account that from (1) it follows

$$b + c = n - (a + d), \tag{14}$$

the formulas (12) and (13) can be given in such form:

$$S_{(a)}(x, y) = \frac{a}{a+t(n-a-d)}, \tag{15}$$

$$S_{(a+d)}(x, y) = \frac{a+d}{a+d+t(n-a-d)}, \tag{16}$$

The parametric families of the similarity measures (15) and (16) have been considered in [20] in the following forms:

$$S_{(A)}(x, y) = \frac{A}{A+\theta(1-A-D)}, \tag{17}$$

$$S_{(A+D)}(x, y) = \frac{A+D}{A+D+\theta(1-A-D)}, \tag{18}$$

where $A = \frac{a}{n}$, $D = \frac{d}{n}$. Further we will use the formulas (15) and (16) for the considered parametric families of similarity measures that will be referred to as (*a*)-family and (*a*+*d*)-family of similarity measures, correspondingly.

We propose to use the relationship (14) for representation of other similarity measures. The similarity measures (8) and (9) do not belong to the considered families of measures, but, using the relation (1), they also can be written as the functions of *a* and *d*:

$$S_{RR}(x, y) = \frac{a}{n}, \tag{19}$$

$$S_F(x, y) = \frac{a+0.5d}{n}. \tag{20}$$

As it is clear from the formulas (15), (16), (19), (20) for fixed numbers *n* and *t* one can build all of these formulas in 3D space as the functions of 2 variables *a* and *d*. (The formula (19) will depend really only from a). From (1) and (14) we obtain:

$$0 \le a + d \le n. \tag{21}$$

This condition defines restrictions on the domain of the considered functions. In all figures below we use the value *n* = 100 and build the graphics of all functions for values *a* and *d* changing from 0 to 100 with the step 1, with the domain restriction (21).

Figures 1(a) and 1(b) show in two different projections Jaccard similarity measure obtained from the parametric formulas (12) and (15) for parameter value *t*=1 as follows:

$$S_J(x, y) = \frac{a}{n-d}. \tag{22}$$

The domain (21) is presented on the plane *S*=0 by triangle with bold sides. Two black lines show the profiles of the surface of the similarity measure: 1) for value *a*=50 and all values of *d;* 2) for value *d*=50 and all values of *a*. The value *S* = 0.5 depicts the value of the measure *S* for *a* = 50 and *d*= 0. When *d* = 0 we obtain in (22) *S*=*a*/*n* that corresponds on Figure 1(a) to the line increasing from 0 to 1 when *d*=0 and *a* is increasing from 0 to 100. Figure 1(b) is obtained from Figure 1(a) by rotation of the axis to show the profile of the surface for small values of *a* and large values of *d*. This situation corresponds to large number of negative matches *d* and hence to small values of nominator and denominator in (2). The similar comments can be done for the figures of other similarity measures shown later.

Figures 2(a) and 2(b) show two projections of Rogers & Tanimoto similarity measure. From (6), (13) and (16) we obtain for *t*=2:

$$S_{RT}(x, y) = \frac{a+d}{2n-a-d}. \tag{23}$$

Figure 3 shows the surfaces of the following similarity measures belonging to the parametric (*a*)-family of measures (from the left to the right): 1) Dice-Czekanowski-Sorensen, 2) Jaccard, 3) Sokal-Sneath-I.

Figure 4 shows the surfaces of the following similarity measures belonging to the parametric (*a*+*d*)-family of measures (from the left to the right): 1) Sokal-Sneath-II, 2) Sokal & Michener, 3) Rogers and Tanimoto.

For all of these similarity measures the formulas like (22) and (23) can be easily obtained from their original definitions by replacement *b*+*c* by *n*-*a*-*d*, see (1) and (14).

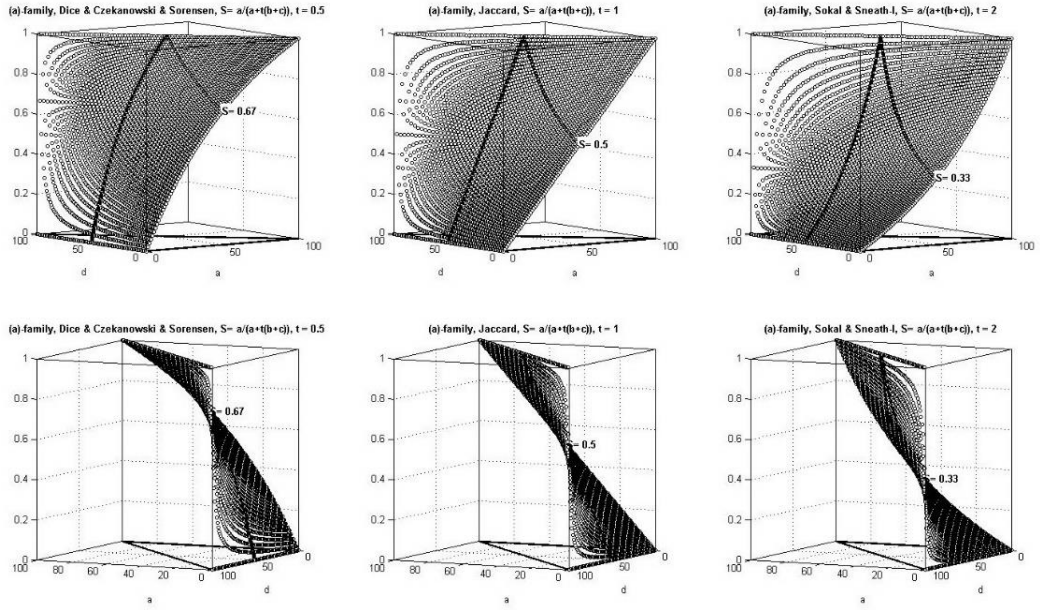


**Fig. 3.** (a)-family of similarity measures in 2 views

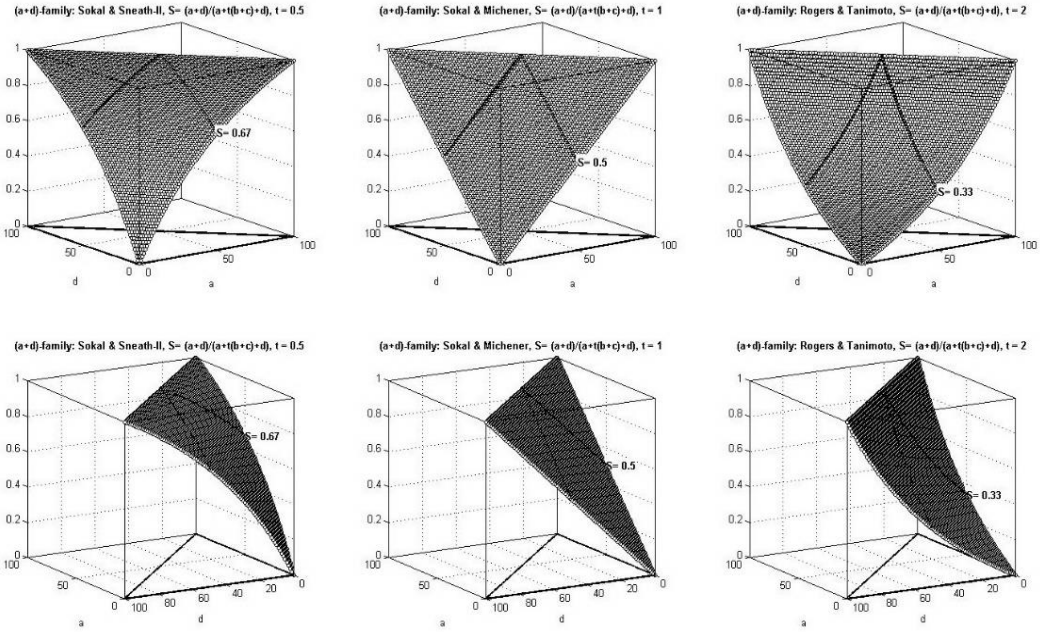


**Fig. 4.** (a+d)-family of similarity measures in 2 views

Figure 5 depicts the surfaces of Rassel & Rao and Faith measures in the same projection as the similarity measures shown on Figures 1(a) and 2(a). Rassel & Rao and Faith measures do not belong nor to (*a*)-family nor to (*a*+*d*)-family of similarity measures and one can see that they

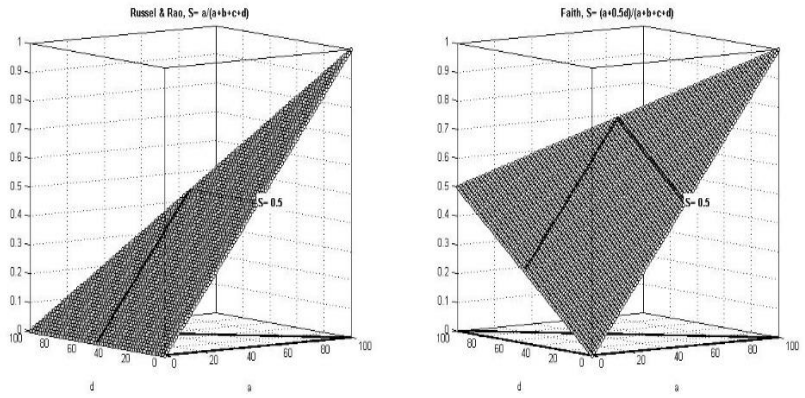


**Fig. 5.** Russel & Rao (on the left side) and Faith (on the right side) similarity measures

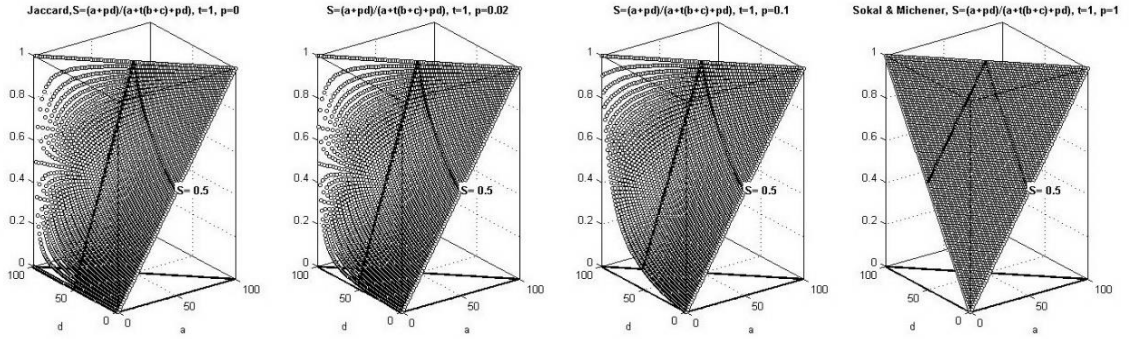


**Fig. 6.** (a+pd)-family of similarity measures: Jaccard (on the left side) and Sokal & Michener (on the right side)

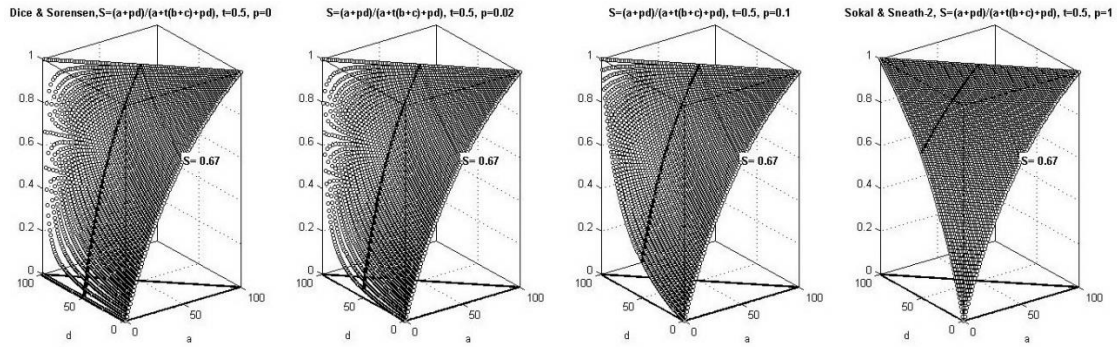


**Fig. 7.** (a+pd)-family of similarity measures: Dice & Czekanowski & Sorensen (on the left side) and Sokal & Sneath – II (on the right side)

have the shapes quite different from the shapes of similarity measures from these families shown on Figure 3 and 4.

The main problem with these two measures that they do not satisfy the reflexivity property $S(x,x)= 1$ that requires that reflexive similarity measure should have the value 1 on the border of the domain where $a+d= n$ and $b= c= 0$. One can see that the similarity measures both from ($a$)-family and from ($a+d$)-family are reflexive.

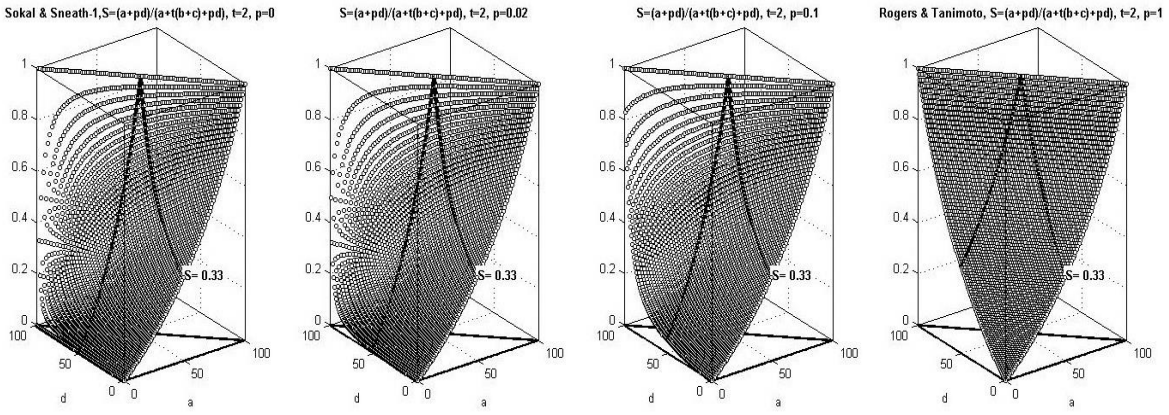


**Fig. 8.** (a+pd)-family of similarity measures: Sokal & Sneath – I (on the left side) and Rogers & Tanimoto (on the right side)

## 4 New Parametric Family of Similarity Measures

As one can see from Figures 3 and 4 the shapes of the similarity measures from (*a*)-family and (*a*+*d*)-family are sufficiently different. The similarity measures *S(x,y)* from (*a*)-family are based on the positive matches of binary attributes in *x* and *y*. The similarity measures from (*a*+*d*)-family are based both on positive and on negative matches. Discussions pro and contra of these two types of similarities measures can be found for example in [5, 10, 17, 19]. We propose the new parametric family of binary similarity measures formally generalizing both these families and giving the possibility to build the similarity measures intermediate between these two families. Below are the two equivalent forms of the new parametric family of measures called (*a*+*pd*)-family:

$$S_{(a+pd)}(x,y) = \frac{a+pd}{a+pd+t(b+c)}, \tag{24}$$

$$S_{(a+pd)}(x,y) = \frac{a+pd}{a+pd+t(n-a-d)}, \tag{25}$$

where *t* is the positive real number and *p* is the number from the interval [0,1]. When *p* = 0 we obtain the (*a*)-family of similarity measures and when *p* = 1 we obtain the (*a*+*d*) family of similarity measures. Changing parameter *p* between 0 and 1 one can move similarity measure from (*a*)-family to (*a*+*d*) family. Generally, the parameters *p* and *t* can be tuned in some procedure of selection of suitable similarity measure for specific application. The selected value of the parameter *p* can reflect the trade-off or relative importance of positive and negative matches in the constructed similarity measure.

Figures 6, 7, 8 show the shapes of binary similarity measures from (*a*+*pd*)-family when parameter *p* is changed from 0 (on the left sides) to 1 (on the right sides) such that on the left sides we have similarity measures from (*a*)-family and on the right sides the measures from (*a*+*d*)-family. The parameter *t* has the values 1, 0.5 and 2 on Figures 6, 7 and 8, respectively. On Figure 6. the similaty measures are changed from Jaccard (on the left side) to Sokal & Michener (on the right side). On Figure 7. the similaty measures are changed from Dice & Czekanowski & Sorensen (on the left side) and Sokal & Sneath – II (on the right side). On Figure 8. the similaty measures are changed from Sokal & Sneath – I (on the left side) and Rogers & Tanimoto (on the right side).

## 5 Discussion and Conclusion

The paper proposes the methods of visualization of the popular similarity measures for binary data and contingency 2 x 2 tables. Such visualization

helps to understand the relationships between these measures and can explain why these similarity measures joined in clusters of similar measures obtained in different works where the clustering of these measures is applied [6,12]. The new parametric family of the similarity measures is proposed. This family generalizes the two known parametric families of similarity measures and gives the possibility to construct similarity measures intermediate between these two families. Such intermediate position can reflect the trade-off or relative importance of positive and negative matches in the construction of similarity measures from the new parametric class of similarity measures. The proposed methodology of visualization of binary similarity measures can be extended on other binary similarity and association measures considered in literature.

## Acknowledgements

The work is partially supported by the projects SIP 20162204 of IPN, 240844 of CONACYT, Mexico, by RFBR project 15-29-01173 and by the Russian Government Program of competitive growth of Kazan Federal University.

## References

1. **Agresti, A. (2002).** *Categorical data analysis*. Wiley and Sons.
2. **Batagelj, V. & Bren, M. (1995).** Comparing resemblance measures. *Journal of Classification,* Vol. 12, No. 1, pp. 73–90. DOI: 10.1007/BF01202268.
3. **Baulieu, F.B. (1989).** A classification of presence/absence based dissimilarity coefficients. *Journal of Classification*, Vol. 6, No. 1, pp. 233–246. DOI: 10.1007/BF01908601.
4. **Choi, S.S., Cha, S.H., & Charles, C.T. (2010).** A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics,* pp. 43–48.
5. **Clifford, H.T. & Stephenson, W. (1975).** *An introduction to numerical classification* (Vol. 229). New York: Academic Press.
6. **Duarte, J.M., Santos, J.B.D., & Melo, L.C. (1999).** Comparison of similarity coefficients based on RAPD markers in the common bean. *Genetics and Molecular Biology,* Vol. 22, No. 3, pp. 427–432. DOI: 10.1590/S1415-47571999000300024.
7. **Goodman, L.A. & Kruskal, W.H. (1954).** Measures of association for cross classifications. *Journal of the American Statistical Association,* Vol. 49, pp. 732–764. DOI: 10.1007/978-1-4612-9995-0_1.
8. **Gower, J.C. (1971).** A general coefficient of similarity and some of its properties. *Biometrics*, pp. 857–871.
9. **Gower, J.C. & Legendre, P. (1986).** Metric and Euclidean properties of dissimilarity coefficients. *Journal of classification*, Vol. 3, No. 1, pp. 5–48. DOI: 10.1007/BF01896809.
10. **Legendre, P. & Legendre, L.F. (1998).** *Numerical ecology,* 2nd English Ed., Elsevier.
11. **Lesot, M-J., Rifqi, M., & Benhadda, H. (2009)** Similarity measures for binary and numerical data: a survey. *Int. J. Knowledge Engineering and Soft Data Paradigms,* Vol. 1, No. 1, pp. 63–84. DOI: 10.1504/IJKESDP.2009.021985.
12. **Meyer, A.D.S., Garcia, A.A.F., Souza, A.P.D., & Souza Jr, C.L.D. (2004).** Comparison of similarity coefficients used for cluster analysis with dominant markers in maize (Zea mays L). *Genetics and Molecular Biology*, Vol. 27, No. 1, pp. 83–91. DOI: 10.1590/S1415-47572004000100014.
13. **Poria, S., Cambria, E., & Gelbukh, A. (2015).** Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. *Proceedings of EMNLP*, pp. 2539–2544.
14. **Poria, S., Gelbukh, A., Hussain, A., Howard, N., Das, D., & Bandyopadhyay, S. (2013).** Enhanced SenticNet with affective labels for concept-based opinion mining. *IEEE Intelligent Systems*, Vol. 28, No. 2, pp. 31–38.
15. **Rodríguez-Salazar, M.E., Álvarez-Hernández, S., & Bravo-Núñez, E. (2001).** *Coeficientes de asociación.* Plaza y Valdés Editores, México.
16. **Sidorov, G., Gelbukh, A., Gómez-Adorno, H., & Pinto, D. (2014).** Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computación y Sistemas,* Vol. 18, No. 3, pp. 491–504. DOI: 10.13053/CyS-18-3-2043.
17. **Sokal, R.R. & Sneath, P.H.A. (1963).** *Principles of Numerical Taxonomy.* WH Freeman.
18. **Tan, P.N., Kumar, V., & Srivastava, J. (2002).** Selecting the right interestingness measure for association patterns. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining,* pp. 32–41. DOI: 10.1145/775047.775053.

19. **Tversky, A. (1977).** Features of similarity. *Psychological review*, Vol. 84, No. 4, pp. 327–352. DOI: 10.1037/0033-295X.84.4.327.
20. **Warrens, M.J. (2013).** A comparison of multi-way similarity coefficients for binary sequences. *International Journal of Research and Reviews in Applied Sciences*, Vol. 16, No. 1, p. 12.

**Ildar Z. Batyrshin** graduated from the Moscow Physical-Technical Institute. He received PhD from the Moscow Power Engineering Institute and Dr. Sci. (habilitation) degree from the Highest Attestation Committee of Russia. He is currently the Titular Professor "C" of CIC IPN. He served as a Co-Chair of 10 International Conferences on Soft Computing, Artificial Intelligence and Computational Intelligence. He is an author and editor of 20 books and special volumes of journals and an author of more than 200 papers in journals and conference proceedings.

**Nailya Kubysheva** recieved her PhD from Lobachevsky State University of Nizhny Novgorod, Russia, and Dr. Sci. (habilitation) degree from the Highest Attestation Committee of Russia. She is currently with IPN, Mexico, as Postdoctoral Researcher.

**Valery Solovyev** did his research with the Higher School of Information Technologies and Information Systems, University of Kazan, Russia. He graduated with a Doctor of Science in Computer Science degree at the Russian Academy of Sciences in 1995. He is currently working as a Senior Researcher and a head of Laboratory of Medical Informatics. His research interests include data mining, computational linguistics, cognitive science. He is the President of The Association of Cognitive Science (Russia).

**Luis A. Villa-Vargas** received his Ph.D. in computer science from the Polytechnic University of Catalonia in 1999. From December 1999 to February 2001 he was with the Laboratory for Computer Science as a Postdoctoral Fellow at the Massachusetts Institute of Technology. From October 2001 to January 2007 he was with the Mexican Petroleum Institute. Since January 2007 he has been with the Center for Computer Research at The National Polytechnic Institute in Mexico, where he was director from 2010 to 2016.

*Article received on 15/07/2016; accepted 10/09/2016.*
*Corresponding author is Ildar Z. Batyrshin.*