Experimental Analysis of Crisp Similarity and Distance Measures

Leila Baccour
REGIM-Lab.: REsearch Groups in Intelligent Machines,
University of Sfax,
ENIS, BP 1173, Sfax, 3038, Tunisia.
Email: leila.baccour@ieee.org

Robert I. John
Automated Scheduling,
Optimisation and Planning (ASAP) group,
Computer Science
Jubilee Campus, Wollaton Road Nottingham, NG8 1BB, UK
Email: Robert.John@nottingham.ac.uk

***Abstract*—Distance measures are a requirement in many classification systems. Euclidean distance is the most commonly used in these systems. However, there exists a huge number of distance and similarity measures. In this paper we present an experimental analysis of crisp distance and similarity measures applied to shapes classification and Arabic sentences recognition.**

***Keywords*-Similarity measures, distance measures, crisp sets, classification of shapes, classification of Arabic sentences**

I. Introduction

In some applications such as recognition, classification and clustering, it is necessary to compare two objects described with vectors of features. This operation is based on a distance or a similarity measure. There are many such measures deployed in such diverse fields as psychology, sociology (e.g. comparing test results), medicine (e.g. comparing parameters of patients), economics (e.g. comparing balance sheet ratios), etc. The characteristics of the data can be very different. Essentially, data can be discrete (i.e. binary or real) or continuous, nominal or numeric. The application of a measure depends on the type of the objects to be compared and the data describing them. In this paper we are interested in similarity and distance measures used in classification and clustering of objects in general. Our intention is to show the importance of the choice of distance or similarity measures in classification systems. Thus, measures from literature are applied to classification of two data sets using the KNN classifier. The obtained results are analyzed and conclusions are given.
In the next section, we detail properties of distance and similarity measures between crisp data. In section three, we present measures between discrete data. In section four and five we apply respectively similarity and distance measures to shapes classification and Arabic sentences recognition.

II. Properties of Crisp Distance and Similarity Measures

A distance measure between two crisp vectors x and y, in a discourse universe E, is defined as a function $d : E^2 \rightarrow R^+$. For all $x, y, z \in E$, this function is required to satisfy the following conditions:

1) minimality: $\forall\ x \in X, d(x, x) = 0$
2) identity: $\forall\ x, y \in X, d(x, y) = 0 \Rightarrow x = y$
3) symmetry: $\forall\ x, y \in X, d(x, y) = d(y, x)$
4) triangular inequality: $\forall\ x, y, z \in X, d(x, y) \leq d(x, z) + d(z, y)$

This function is called distance, index of distance, index of dissimilarity, semi-metric distance, or distance ultra-metric according to one of the following cases:

- index of dissimilarity: if the function d satisfies the properties 1 and 2.
- index of distance or semi-metric distance: if the function d satisfies the properties 1, 2 and 3
- distance: if the function d satisfies the properties 1, 2, 3 and 4
- distance ultra-metric: if the function d verifies the property $\forall\ x, y, z \in E, d(x, y) \leq max\left(d\left(x, z\right), d\left(z, y\right)\right)$

Two patterns x and y are close if the distance $d(x, y)$ tends to 0, $d(x, y) = 0$ imply a complete similarity between x and y. Inversely, x and y are close, if the measure of similarity is high, $S(x, y) = 0$ imply a complete dissimilarity between x and y. The choice of distance or similarity measures depends on the application type and on the objects description.

III. Similarity and Distance Measures for Discrete Data

In the following we give the mathematical definitions of distances from literature to measure the closeness between two samples $x(x_1 \ldots x_n)$ and $y(y_1 \ldots y_n) \in E$ having n numeric attributes.

1) Minkowski Distances: The Minkowski distance [1] is a generalized metric, also said L_p norm, between two samples x and y. It is defined by the following formula:

$$d_{MK}(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}} \quad (1)$$

where $p > 0$ is the order of the Minkowski metric.
This distance is the generalization of the following distances (2, 3) [2]. According to the value of the parameter p, we have one of the following distances:

- $p = 1$: Manhattan distance, also said L_1-Norm, Taxicab norm, rectilinear distance or City-Block distance. Manhattan Distance between x and y is defined as:

$$d_{Ma}(x, y) = \sum_{i=1}^{n} |x_i - y_i| \quad (2)$$

This distance degenerates to the Hamming distance [3].

978-1-4799-5934-1/14/$31.00 ©2014 IEEE

- $p = 2$: Euclidean distance , $L2$-Norm or Ruler distance is the "ordinary" distance between two points and has the following formula:

$$d_E(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{3}$$

If the square root is omitted the obtained formula will be that of Euclidean Squared distance.
- $p = \infty$: Chebyshev distance [4], [5] is also called maximum value distance, chessboard distance in 2D or the minimax approximation.

$$d_{Ch}(x,y) = max_i \left|x_i - y_i\right| \tag{4}$$

2) Spearman Distance: Spearman distance is the square of Euclidean distance between two rank vectors defined as:

$$d_S(x,y) = \sum_{i=1}^{n} (x_i - y_i)^2 \tag{5}$$

3) Hamming Distance: The Hamming distance [6] is a metric expressing the distance between two samples by the number of mismatches of their pairs of attributes. It is mainly used for nominal data, string and bitwise analyses, but can also be useful for numerical variables. This can be found by using the XOR operator for the corresponding bits or equivalent. The Hamming distance is an important measurement for the detection of errors in information transmission [6]. The corresponding formula is:

$$d_{HA}(x,y) = \sum_{i=1}^{n} x_i \neq y_i \tag{6}$$

4) L_1 Distances:

- Bray-Curtis or Sorensen distance: The non-metric Bray-Curtis dissimilarity [7] is one of the most commonly applied measurements to express relationships in ecology, environmental sciences and related fields.
  Bray-Curtis is a modified Manhattan measurement, where the summed differences between the attributes values of the samples x and y are standardized by their summed attributes values. The general equation of Bray-Curtis dissimilarity is:

$$d_{BC}(x,y) = \frac{\sum_{i=1}^{n} \left|x_i - y_i\right|}{\sum_{i=1}^{n} (x_i + y_i)} \tag{7}$$

Or equivalent

$$d_{BC}(x,y) = \frac{\sum_{i=1}^{n} \left|x_i - y_i\right|}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i} \tag{8}$$

The Bray-Curtis similarity d_{BC} is a slightly modified equation. It can be directly calculated from the dissimilarity value: $S_{BC} = 1 - d_{BC}$
The result is undefined, when the variables among two samples x and y are entirely 0. In this case the denominator becomes 0 and [8] suggest to use a zero-adjusted Bray-Curtis coefficient and proposed this modified formula:

$$d_{BC_M}(x,y) = \frac{\sum_{i=1}^{n} \left|x_i - y_i\right|}{2 + \sum_{i=1}^{k} (x_i + y_i)} \tag{9}$$

- Lorentzian Distance defined as [9]:

$$d_{Lo} = \sum_{i=1}^{n} \ln(1 + \left|x_i - y_i\right|) \tag{10}$$

1 is added to guarantee the non-negativity property and to avoid the log of zero.
- Canberra Distance:
  It was introduced in 1966 [10] and has the following formula:

$$d_C(x,y) = \sum_{i=1}^{n} \frac{\left|x_i - y_i\right|}{\left|x_i + y_i\right|} \tag{11}$$

d_C (11) is slightly modified by the following formula suggested in [11].

$$d_C(x,y) = \sum_{i=1}^{n} \frac{\left|x_i - y_i\right|}{\left|x_i\right| + \left|y_i\right|} \tag{12}$$

The Canberra metric is mainly used for positive values. This metric is very sensitive for values close to 0, where it is more sensitive to proportional than to absolute differences [11]. This characteristic becomes more apparent in higher dimensional space, and respectively with an increasing number of variables.

5) Inner Product Similarities:

- Inner product similarity:

$$S_{IP} = X \cdot Y = \sum_{i=1}^{n} x_i y_i \tag{13}$$

where $X \cdot Y$ is the scalar product between X and Y. The inner product is sometimes called the scalar product or dot product [12]. If it used for binary vectors, it is called the number of matches or the overlaps.
- Harmonic mean similarity:

$$S_{HM} = 2 \sum_{i=1}^{n} \frac{x_i y_i}{x_i + y_i} \tag{14}$$

- Cosine similarity is the normalized inner product, called the cosine coefficient because it measures the angle between two vectors. Sometimes called angular metric [9]. Other names for the cosine coefficient include Ochiai [9], [13] and Carbo [13].

$$S_{Cos} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}} \tag{15}$$

- Jaccard Similarity: The Jaccard index, known as the Jaccard similarity coefficient [14] is widely used in various fields such as ecology and biology. It is defined as:

$$S_{Jac} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} x_i y_i} \tag{16}$$

Jaccard distance is computed as $1 - S_{Jac}$, has the following formula:

$$d_{Jac} = \frac{\sum_{i=1}^{n}(x_i - y_i)^2}{\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} x_i y_i} \quad (17)$$

Jaccard similarity is the same as Kumar-Hassebrook similarity.

- Dice Similarity is defined in [15] as:

$$S_{Dice} = \frac{2\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2} \quad (18)$$

known as the Dice coefficient [16]. The distance of Dice is computed as $1 - S_{Dice}$, does not validate the property of triangle inequality and has the following formula:

$$d_{Dice} = \frac{\sum_{i=1}^{n}(x_i - y_i)^2}{\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2} \quad (19)$$

It is identical to the Sorensen similarity index [17], and is said the Sorensen-Dice coefficient.

The measures (13, 15, 17, 18) are frequently encountered in information retrieval and biological taxonomy for the binary feature vector comparisons/

6) Squared Chi-Squared Distance: This distance is defined as:

$$d_{Scs}(x, y) = \sum_{i=1}^{n} \frac{(x_i - y_i)^2}{|x_i + y_i|} \quad (20)$$

7) Squared chord Distance: The Squared chord distance is defined in [18], [19] as:

$$d_{Sc}(x, y) = \sum_{i=1}^{n} \left(\sqrt{x}_i - \sqrt{y}_i\right)^2 \quad (21)$$

This distance measure is popular with paleontologists and in studies on pollen. However, it can be applied in comparative evaluation [20].

8) Hausdorff Distance: The Hausdorff metric is defined in [21], [22] as follows:

$$d_{Hau}(x, y) = max\left(h\left(x, y\right), h\left(y, x\right)\right) \quad (22)$$

where

$$h(x, y) = \max_{x_i \in x} \min_{y_i \in y} ||x_i - y_i|| \quad (23)$$

and $||.||$ is the vector norm (e.g. L_2 norm, $L_2(x) = \sqrt{x_1^2 + \cdots + x_n^2}$).

The function $h(x, y)$ is called the directed Hausdorff distance from x to y. The Hausdorff distance $d_{Hau}(x, y)$ measures the degree of mismatch between the sets x and y by measuring the remoteness between each point x_i and y_i and vice versa. Thus, the measure of resemblance is interested to measure the closeness of each x_i with y_i and inversely. The Hausdorff Distance is modified in [23] and was proven to be more efficient for object matching than the original measure. The proposed modified Hausdorff Distance does not verify the triangular inequality, so it is an index of distance. The corresponding formula for modified Hausdorff Distance concerns the distance h (22) which is substituted by a distance g as follows:

$$g(x, y) = \frac{1}{|x|} \sum_{x_i \in x} \min_{y_i \in y} d_1(x_i, y_i) \quad (24)$$

where d_1 is any distance and $|x|$ the cardinality of the set x. Thus, modified Hausdorff distance becomes:

$$d_{MHd}(x, y) = max\left(g\left(x, y\right), g\left(y, x\right)\right) \quad (25)$$

9) Bhattacharyya Distance: The Bhattacharyya distance [24] is defined by the following formula:

$$d_{Bh}(x, y) = -\ln \sum_{i=1}^{n} \sqrt{x_i y_i} \quad (26)$$

The Bhattacharyya distance has a value between 0 and 1.

IV. Application of Crisp Similarity Measures to Shapes Classification

Each of measures from literature is applied on SQUID data set, which contains about 1100 images of marine creatures (fish) using KNN classifier. The data set is described with features detailed in [25], [26], [27], [28], [29]. Results are presented in percent in the table I.

Table I
Results of Shapes Classification Using Crisp Similarity Measures

Similarity Measures	1-Best	2-Best	10-Best	Error Rate
d_{Bh} (26)	24.31	39.50	77.62	22.38
S_{HM} (14)	24.03	43.09	76.24	23.76
d_{Scs} (20)	22.93	38.67	80.66	19.34
S_{IP} (13)	18.23	37.29	85.36	14.64
d_{MHd_1} (25)	16.85	27.62	69.06	30.94
d_{Ch} (4)	15.47	26.52	66.57	33.43
d_E (3)	14.36	27.35	65.19	34.81
d_{MHd_2} (25)	13.81	30.94	71.82	28.18
d_{Sc} (21)	13.81	20.72	37.85	62.15
d_{BC} (7)	13.54	21.27	53.59	46.41
d_{Lo} (10)	12.71	22.38	59.94	40.06
d_{Ma} (2)	12.15	22.65	60.5	39.5
d_{Hau} (22)	11.6	19.61	54.7	45.3
d_C (11)	8.29	16.57	51.1	48.9
S_{Jac} (16)	4.14	8.01	52.21	47.79
S_{Dice} (18)	4.14	8.01	51.93	48.07
S_{Cos} (15)	0.28	0.28	0.55	99.45

According to one-best classification, the best results are produced with Bhattacharyya distance d_{Bh} followed by the results of harmonic mean similarity S_{HM} and squared Chi Squared Distance $d_{S_{cs}}$. The latter present lower error rate than S_{HM} according to 10-Best. In the fourth rank, we find results of inner product similarity S_{IP} with the lowest error rate and best results according to 10-Best. In the fifth, sixth and seventh rank, we find results of modified Hausdorff distance

using the norm of vectors d_{MHd_1}, Chebyshev distance d_{Ch} and Euclidean distance with close results. Next results are produced with the measures modified Hausdorff distance using squared chi squared distance d_{MHd_2}, Squared chord distance d_{Sc}, Bray Curtis distance d_{BC}, Lorentzian distance d_{Lo} and Manhattan distance d_{Ma}. Results decrease using Hausdorff distance d_{Hau} and Canberra distance d_C. The worst results are obtained with Jaccard, Dice and cosine coefficient similarities. We note that results of modified Hausdorff distance (d_{MHd_1}, d_{MHd_2}) are better than the initial one d_{Hau}.

To give more idea about distance and similarity measures results we apply them to another data set, using KNN classifier, presented in next section.

V. Handwritten Arabic Sentences Classification

The crisp measures are applied to a data set of 6537 images of 824 handwritten Tunisian town/village names extracted from the IFN/ENIT data set [30]. The data set is defined with features detailed in [25], [26], [28], [27].

In the table II below, the results in percent of crisp measures applied to handwritten Arabic sentences classification are exposed. According to one-best in classification, the best results

Table II
Arabic Sentences Classification Results Obtained with Each Fuzzy Similarity Measure

Similarity Measures	1-Best	2-Best	10-Best	Error Rate
d_C (11)	67.16	76.38	89.86	10.14
d_{Scs} (20)	65.05	74.91	90.09	9.91
d_{Ma} (2)	64.82	75.18	89.82	10.18
S_{Cos} (15)	64.5	73.26	89.63	10.37
d_E (3)	64.45	74.91	89.86	10.14
S_{Dice} (18)	63.94	74.08	88.81	11.19
d_{Sc} (21)	63.39	73.07	89.36	10.64
S_{Jac} (16)	62.89	71.83	81.33	18.67
d_{Lo} (10)	56.7	66.56	83.35	16.65
d_{Ch} (4)	53.39	63.81	84.17	15.83
S_{HM} (14)	50.92	62.11	83.85	16.15
d_{Bh} (26)	50.28	61.51	83.12	16.88
S_{IP} (13)	43.07	56.06	81.01	18.99
d_{BC} (7)	28.17	39.91	68.07	31.93
d_{MHd_2} (25)	27.75	38.17	67.98	32.02
d_{Hau} (22)	26.33	38.03	69.36	30.64
d_{MHd_1} (25)	23.76	35.28	64.95	35.05

are obtained with d_C, followed by results of d_{Scs}, d_{Ma}, S_{Cos}, d_E, S_{Dice}, d_{Sc}, S_{Jac} which have close results with differences of 1% or less. Results decrease using d_{Lo} with a difference of 6.19% from its previous and continue decreasing using d_{Ch} with a diffrence of 3.31% from d_{Lo}. Next results are those of S_{HM}, d_{Bh} and S_{IP} and worst results are obtained using measures d_{BC}, d_{MHd_2}, d_{Hau} and d_{MHd_1}.

The obtained results for classification of shapes and Arabic sentences are different. So, results of d_{Bh} are the best in the first application and modest in second application. As well, results of d_C are the best in second application and low in first application. S_{Cos} has the worst results with the last rank in the first application and has the forth rank in second application. Generally, there is no common rank of results of measures in the two applications. This shows that the choice of a measure between crisp data depends on application and on data.

VI. Conclusion

Crisp similarity and distance measures are applied for classification of shapes and handwritten Arabic sentences. The measures produced different results in both applications, which makes difficult the choice of a measure for such applications. However, Euclidean distance did not produce best results. Thus, the importance in classification can not be assigned only to features, the choice of distance is also important and can affect results. Other studies can be found in [27], [31] about fuzzy similarity and intuitionistic fuzzy similarity measures.

Acknowledgment

The authors would like to acknowledge the financial support of this work by grants from General Direction of Scientific Research (DGRST), Tunisia, under the ARUB program.

References

[1] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a non metric hypothesis," *Psychometrika*, pp. 1–27, 1964.

[2] E. Krause, *Taxicab Geometry An Adventure in Non-Euclidean Geometry*, 1975.

[3] B. S. Manjunath, J. R. Ohm, V. V. Vasudevan, and A. Yamada, "Color and texture descriptors," *In special Issue on MPEG-7. IEEE transactions on circuits and systems for video technology*, vol. 11(6), pp. 703–715, June 2001.

[4] C. D. Cantrell, *Modern Mathematical Methods for Physicists and Engineers*. Cambridge University Press, 2000.

[5] J. M. Abello, P. M. Pardalos, and M. G. C. Resende, *Handbook of Massive Data Sets*. Springer, 2002.

[6] R. W. Hamming, "Error detecting and error correcting codes," *Bell System Technical Journal*, vol. 2, pp. 147–160, 1950.

[7] J. R. Bray and J. T. Curtis, "An ordination of the upland forest communities of southern wiscons," *Ecological Monographies*, vol. 27, pp. 325–349, 1957.

[8] K. R. Clarke, P. J. Somerfield, and M. G. Chapman, "On resemblance measures for ecological studies, including taxonomic dissimilarities and a zero-adjusted bray-curtis coefficient for denuded assemblages," *Journal of Experimental Marine Biology and Ecology*, vol. 330, pp. 55–80, 2006.

[9] E. Deza and M. Deza, *Dictionary of Distances*. Elsevier Science, 2006.

[10] G. N. Lance and W. T. Williams, "Computer programs for hierarchical polythetic classification (similarity analysis)," *Computer journal*, vol. 9, pp. 60–64, 1966.

[11] ——, "Mixed-data classificatory programs," *Agglomerative systems. australian computer journal*, vol. 1, pp. 15–20, 1967.

[12] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. Wiley, 2001.

[13] V. Monev, "Introduction to similarity searching in chemistry," *Match commun. math. comput. chem.*, vol. 51, pp. 7–38, 2004.

[14] P. Jaccard, "Distribution de la flore alpine dans le bassin des dranses et dans quelques rgions voisines," *Bulletin de la socit vaudoise des sciences naturelles*, vol. 37, pp. 241–272, 1901.

[15] A. R. Leach and V. J. Gillet, *An Introduction to Chemoinformatics*. Springer, 2007.

[16] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26 (3), pp. 297–302, 1945.

[17] T. Sorensen, "A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons," *Videnski Selsk. Biol. Skr.*, vol. 5, pp. 1–34, 1948.

[18] I. C. Prentice, "Multidimensional scaling as a research tool in quaternary palynology: a review of theory and methods," *Review of Palaeobotany and Palynology*, vol. 31, pp. 71–104, 1980.

[19] J. T. Overpeck, T. Webb III, and I. C. Prentice, "Quantitative interpretation of fossil pollen spectra: dissimilarity coefficients and the method of modern analogs," *Quaternary Research*, vol. 23, pp. 87–108, 1985.

[20] R. Hu, S. Rger, D. Song, H. Liu, and Z. Huang, "Dissimilarity measures for content-based image retrieval," in *IEEE International Conference on Multimedia and Expo*, 2008.

[21] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the hausdorff distance," *Ieee transactions on pattern analysis and machine intelligence*, vol. 15 (9), pp. 850–863, 1993.

[22] J. Henrikson, "Completeness and total boundedness of the hausdorff metric," *MIT Undergraduate Journal of Mathematics*, pp. 69–79, 1999.

[23] M. P. Dubuisson and A. K. Jain, "A modified hausdorff distance for object matching," in *proc. 12th International Conference on Pattern Recognition, pp 566–568, Jerusalem, Israel, october 1994*, 1994.

[24] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bulletin of the Calcutta Mathematical Society*, vol. 35, pp. 99–109, 1943.

[25] L. Baccour, S. Kanoun, V. Maergner, and A. M. Alimi, "An application of intuitionistic fuzzy information for handwritten arabic word recognition," in *Proc. Notes on IFS, the twelfth International Conference on Intuitionistic Fuzzy Sets, ICIFS'2008, Sofia Bulgaria, vol. 14 (2), pp. 67–72*, 2008.

[26] L. Baccour and A. M. Alimi, "A comparison of some intuitionistic fuzzy similarity measures applied to handwritten arabic sentences recognition," in *Proc.IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2009, ICC Jeju Island, Korea, pp. 1389–1392.*, 2009.

[27] L. Baccour, A. M. Alimi, and R. I. John, "Some notes on fuzzy similarity measures and application to classification of shapes, recognition of arabic sentences and mosaic," *IAENG International Journal of Computer Science*, vol. 41 (2), pp. 81–90, 2014.

[28] L. Baccour and A. M. Alimi, "Applications and comparisons of fuzzy similarity measures," in *Proc. IEEE World Congress on Computational Intelligence, WCCI 2010, FUZZ-IEEE 2010, Barcelone, pp. 1–7*, 2010.

[29] L. Baccour, A. M. Alimi, and R. I. John, "Relationship between intuitionistic fuzzy similarity measures," in *Proc. IEEE International Conference on Fuzzy Systems, FUZZ-IEEE'2011, Taipei, Taiwan, pp. 971–975*, 2011.

[30] M. Pechwitz and al., "Ifn/enit – data set of handwritten arabic words," in *Proc. of CIFED 2002, pp. 129-136, Hammamat, Tunisia*, 2002.

[31] L. Baccour, A. M. Alimi, and R. I. John, "Similarity measures for intuitionistic fuzzy sets : State of the art," *Journal of Intelligent & Fuzzy Systems*, vol. 24 (1), pp. 37–49, 2013.