16th World Congress of the International Fuzzy Systems Association (IFSA)
9th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT)

Modulo similarity in comparing histograms

Pasi Luukka[1] **Mikael Collan**[1]

[1] School of Business and Management
Lappeenranta University of Technology
P.O. Box 20, FIN-53851 Lappeenranta, Finland
pasi.luukka; mikael.collan@lut.fi

Abstract

Histograms are a tool for graphical representation of frequency data and thus helpful in creating a fast understanding of, e.g., contents of frequency data. Comparing histograms is topic of increasing importance due to an increase in the availability of data sets containing frequency information. Automatic data collection from “everywhere” has made collection of frequency data very common. As many different types of similarities exist, our focus is on Łukasiewicz logic-based similarity and we present two new measures, the “modulo similarity” measure and the “maximum pair assignment compatibility” measure. These measures do not use PDF conversion, or vector-based approaches, in the comparison of histograms, but concentrate on the data samples used to form histograms. We illustrate the usefulness of these measures with numerical examples.

Keywords: Histograms, comparison, Łukasiewicz logic, similarity, modulo similarity

1. Introduction

Histograms are used in a wide variety of applications and their usage is even likely to increase in the future, due to histograms being an intuitive and an easy to use way of visualizing frequency information. Histograms can be used in expanding data analysis capabilities and in visual presentation of big data related questions. One area of research related to histograms is the comparison of histograms. Comparison of histograms is by no means a trivial task, due to a wide variety of different types of histograms and the many possible ways of constructing methods of comparison.

One common approach is to compute a distance between histograms by using one of many different distance measures [1]. Another way is to transform histograms into a probability density functions (PDF) and to compare them to each other. This approach was one of the first ones introduced and is based on the assumption that a histogram created from measured values provides the basis for an empirical estimate of a PDF [2]. Computing the distance between two PDFs can be regarded as an operation similar to computing a Bayesian probability. Bhattacharyya distance (sometimes referred to as B-distance) is among the first created measures for the calculating the distance between two statistical populations [3]. Later on also other distance measures have been applied to the comparison of PDFs, e.g., the K-L distance [4] that also was among the first ones to appear.

In what can be called “vector type of approaches”, histograms are treated as fixed-dimensional vectors, between which a distance is computed. The usually applied distances include the Euclidean and the Manhattan distances, or generalizations of these, like the Minkowski distance, see Bandemer & Näther [5] for a listing of different types of distances and generalizations of standard Euclidean and Manhattan distances.

Later on, also other methods for comparing histograms have been introduced, these include, e.g., approaches that consider the overlapping / non overlapping parts of histograms that is, intersectional approaches. These methods also use distances, e.g., the earth mover’s distance [6] in the measurement of the non-overlapping parts – the idea in these approaches is that the distance is based on computing the minimal amount of work required to transform one histogram into the other by moving “distribution mass”.

In this paper we examine how *similarity measures* can be used in the comparing histograms and our main focus is on Łukasiewicz logic [7] based similarity. In the same way as a fuzzy subset generalizes a classical subset, the concept of similarity can be considered as a many-valued (MV) generalization of the classical notion of equivalence [8]. Equivalence relation is a familiar way to classify similar mathematical objects.

Jan Łukasiewicz [7] was the first researcher to systematically investigate many-valued logics in the 1920's. Chang introduced MV algebras in 1958 [9], and provided a proof of completeness to Łukasiewicz logic. Łukasiewicz logic was generalized in 1979, by Jan Pavelka [10]. In 1999, Turunen showed that the arithmetic mean of many similarities [11] is still a fuzzy similarity, which is a property that holds only, when we use Łukasiewicz logic. This property also holds in the generalized form of Łukasiewicz logic that is, in Łukasiewicz-Pavelka logic. Similarity based on Łukasiewicz logic has been applied in a variety of applications: in classification it has been applied to a similarity based classifier, see e.g. [12,13], in multi-criteria decision making (MCDM) problems it is has been applied to the TOPSIS method [14], in feature selection problems it has been applied to classification based problems [15] and MCDM based problems [16], and in control applications it has been applied to traffic signal control [17] and water reservoir control [18]. Furthermore, it has been applied to defining athletes' aerobic and anaerobic thresholds [19] and the maximal heart rate [20]. A survey of applications of the Łukasiewicz-Pavelka logic has been written by Turunen [21].

The following section 2 continues by presenting different types of similarities that can be applied to comparing histograms. A new "modulo similarity" for circular modulo-type problems is introduced, and the axiomatic properties of the new modulo similarity measure are examined. We prove that modulo similarity satisfies all three axioms required of a "true" similarity measure. Modulo similarity is a totally new concept and the proof that the three axioms required of a similarity measure hold is a new contribution. In section 3, another new concept the "maximum pair assignment compatibility" is introduced. These new concepts are numerically illustrated with an example. Section 4 closes the paper with concluding remarks.

2. Similarity of different types of histograms

Let us first start with the definition of a histogram and then move onto different types of histograms, and how to define similarity between histograms. We begin with the definition of a histogram:

Definition 1*: Let x be a feature having m different values given in a set* $X = \{x_1, \dots, x_m\}$. *Consider set of elements* $A = \{a_1, \dots, a_n\}$, *where* $a_j \in X$. *The histogram of the set A along with feature x is H(x,A) giving an ordered m-dimensional list consisting of the number of occurrences of the discrete values of x among* a_i.

Here we focus in the comparison of histograms of the same measurement *x*, notation *H(A)* will be used in place of *H(x,A)* without loss of generality. If $H_i(A), 1 \leq i \leq m$, denotes the number of elements of *A* that have values x_i, then $H(A) = \{H_1(A), H_2(A), \dots, H_m(A)\}$, where

$$H_i(A) = \sum_{j=1}^{n} b_{ij}, where\ b_{ij} = \begin{cases} 1 \ \ if \ \ a_j = x_i \\ 0, otherwise \end{cases} \tag{1}$$

Note: If $P_i(A)$ denotes the probability of samples in the *j*th value, then $P_i(A) = \frac{H_i(A)}{n}$. This is also sometimes used as a histogram measure, and is well suited for similarity measure-type comparison, since $P_i(A) \in [0,1]$. We simply denote this type of variation as $HP_i(A)$, formally:

$$HP_i(A) = \frac{\sum_j^n b_{ij}}{n} \tag{2}$$
$$where\ b_{ij} = \begin{cases} 1 \ \ if \ \ a_j = x_i \\ 0, otherwise \end{cases}$$

Example: Consider *n=10*, *m=6* and *A={1,6,5,1,1,2,5,5,1,1}*, *H(A)={5,1,0,0,3,1}*, and *HP(A)={0.5,0.1,0,0,0.3,0.1}*. If the ordering of the elements in the set *A* is not considered, then *H(A)* is a lossless representation of *A,* meaning that *A* can be fully reconstructed from *H(A).*

2.1. Different histogram types

Histograms can be divided into three different types, in connection with computing histogram similarities: 1) nominal, 2) ordinal, and 3) modulo. In nominal histograms each variable has a "name" that is, the variable "make of a car" can take nominal values such as "Ford", "Toyota", "Skoda", and so forth. Nominal type histograms can, e.g., consist of the frequency of cars manufactured by each car maker in a parking lot. In ordinal type histograms, the variables are (can be) ordered, e.g., the number of valves in a car can be quantified into 2 to 5 valves per cylinder, or the weight of the vehicle from 1 to 10 tons. In the third, modulo type histograms, the measured (or observed) variables form a circle in the same way as hours form a circle on a clock-face with arithmetic modulo 12, or a compass with degrees, arithmetic modulo 360. Graphical presentation of modulo-type histograms is available, e.g., in [1].

2.2. Similarity between samples of discrete measurement results

Given a set of samples, with each sample containing measured discrete values of a variable, a histogram represents the frequency of each discrete variable value measured. Considering three different types of measurements, nominal, ordinal, and modulo, we present three different *similarities* between two measurements (samples) $x_a, x_b \in X$. We normalize the sample values between unit intervals, by setting $x_{am} = \frac{x_a}{m}, x_{bm} = \frac{x_b}{m}$, where m denotes the largest variable value, or "bin" value (e.g., m=360° in a compass).

Nominal similarity:

$$S_{nom}(x_{am}, x_{bm}) = \begin{cases} 1 \ \ if \ \ x_{am} = x_{bm} \\ \ 0, otherwise \end{cases} \tag{3}$$

Ordinal similarity:

$$S_{ord}(x_{am}, x_{bm}) = 1 - |x_{am} - x_{bm}| \tag{4}$$

Modulo similarity:

$$S_{mod}(x_{am}, x_{bm}) = \begin{cases} 1 - |x_{am} - x_{bm}| \ if \ |x_a - x_b| \leq \frac{m}{2} \\ |x_{am} - x_{bm}| \qquad\qquad otherwise \end{cases} \tag{5}$$

In the first similarity measure, for the similarity of two nominal type sample values we either have a match, or we don´t, in line with classical equivalence. In the ordinal type similarity, the element values´ similarity is defined in the same way as the original Łukasiewicz similarity (see [1] and [2]). In the third case, with modulo similarity, the values form a circle, an issue that must be taken into consideration. For example, for the compass situation, the angular values between 0° to 360° $(355^o, 13^o) = \left|\frac{355}{360} - \frac{13}{360}\right| = 0.95 \neq 0.05 = 1 - \left|\frac{355}{360} - \frac{13}{360}\right|$.

2.3. Similarity in the Łukasiewicz structure

Since, in (4) and (5) we are using similarity based on Łukasiewicz logic, let us first briefly review axiomatic properties required for this similarity. Also note that our first equation for similarity (3) is same as the standard crisp equivalence relation. A Łukasiewicz similarity [2] measure needs to satisfy reflexivity, symmetricity, and transitivity conditions. We next first shortly introduce the Łukasiewicz logic and then go through the needed axioms, in order to examine, whether our new modulo similarity is a similarity in the mathematical sense.

Definition 2: *A lattice is partially ordered set in which $x \wedge y$ (infimum) and $x \vee y$(supremum) exists in L for all elements $x, y \in L$. A lattice is often denoted by $\langle L, \leq, \wedge, \vee \rangle$.*

Definition 3: *A lattice is called residuated, if it contains the greatest element 1, and binary operations $\odot$ (called multiplication) and $\rightarrow$ (called residuum) such that following conditions hold*

1. *$\odot$ is associative, commutative, and isotone.*
2. *$a \odot 1 = a$ for all elements $a \in L$ and*
3. *for all elements $a, b, c \in L$, $a \odot b \leq b$ if and only if $a \leq b \rightarrow c$*

Definition 4: *Letting L be the real unit interval [0,1] endowed with the usual order relation, we may construct the following usual residuated lattice: Łukasiewicz structure: $a \odot b = max\{a + b - 1{,}0\}$, $a \rightarrow b = min\{1{,}1 - a + b\}$.*

Definition 5: *Let L be a residuated lattice and X is a non empty set. L-valued binary relation S, defined in X is a similarity, if it fulfills the following conditions (Turunen, 1999):*

1. $\forall x \in X: S(x, x) = 1$
2. $\forall x_1, x_2, \in X: S(x_1, x_2) = S(x_2, x_1)$
3. $\forall x_1, x_2, x_3, \in X: S(x_1, x_2) \odot S(x_2, x_3) \leq S(x_1, x_3)$

Notice that in case we let L be the two element set {0,1}, similarity coincides with the usual equivalence relation. In Łukasiewicz-logic equivalence relation (or similarity relation) is defined as $1 - \max\{x_1, x_2\} + min\{x_1, x_2\}$, or equivalently as $S(x_1, x_2) = 1 - |x_1 - x_2|$ (see [2]).

2.4. Axiomatic properties of modulo similarity

We prove that modulo similarity is transitive (condition 3 in definition 5) and then continue by proving that it is also reflexive and symmetric (conditions 1 and 2 in definition 5).

Theorem 1. S_{mod} satisfies the transitivity property

Proof: Since we know that $\forall x_1, x_2, \in X: S(x_1, x_2) = 1 - |x_1 - x_2|$ satisfies condition 3 (see e.g. [2]) the proof reduces to a study of the cases, where also $|x_1 - x_2| > \frac{m}{2}$ is applied. Here we have three possible different cases:

1) $|x_1 - x_2| > \frac{m}{2}$
2) $|x_1 - x_2| > \frac{m}{2}$ and $|x_1 - x_3| > \frac{m}{2}$
3) $|x_1 - x_3| > \frac{m}{2}$

For 1) we have $x_1 \leq x_3 \leq x_2$ and
$S(x_1, x_2) \odot S(x_2, x_3) \leq S(x_1, x_3)$
$\Leftrightarrow max\left\{\left|\frac{x_1}{m} - \frac{x_2}{m}\right| + 1 - \left|\frac{x_2}{m} - \frac{x_3}{m}\right| - 1{,}0\right\} \leq 1 - \left|\frac{x_1}{m} - \frac{x_3}{m}\right|$
$\Leftrightarrow max\left\{\left|\frac{x_1}{m} - \frac{x_2}{m}\right| - \left|\frac{x_2}{m} - \frac{x_3}{m}\right|, 0\right\} \leq 1 - \left|\frac{x_1}{m} - \frac{x_3}{m}\right|$
Now $max\left\{\left|\frac{x_1}{m} - \frac{x_2}{m}\right| - \left|\frac{x_2}{m} - \frac{x_3}{m}\right|, 0\right\} = \left|\frac{x_1}{m} - \frac{x_2}{m}\right| - \left|\frac{x_2}{m} - \frac{x_3}{m}\right|$ due to $x_1 \leq x_3 \leq x_2$ so we get
$max\left\{\left|\frac{x_1}{m} - \frac{x_2}{m}\right| - \left|\frac{x_2}{m} - \frac{x_3}{m}\right|, 0\right\} \leq 1 - \left|\frac{x_1}{m} - \frac{x_3}{m}\right|$

$\Leftrightarrow \left|\frac{x_1}{m} - \frac{x_2}{m}\right| - \left|\frac{x_2}{m} - \frac{x_3}{m}\right| \leq 1 - \left|\frac{x_1}{m} - \frac{x_3}{m}\right|$
$\Leftrightarrow \left|\frac{x_1}{m} - \frac{x_2}{m}\right| - \left|\frac{x_2}{m} - \frac{x_3}{m}\right| + \left|\frac{x_1}{m} - \frac{x_3}{m}\right| \leq 1$
$\Leftrightarrow max\left\{\frac{x_1}{m}, \frac{x_2}{m}\right\} - min\left\{\frac{x_1}{m}, \frac{x_2}{m}\right\} - \left(max\left\{\frac{x_2}{m}, \frac{x_3}{m}\right\} - min\left\{\frac{x_2}{m}, \frac{x_3}{m}\right\}\right) + max\left\{\frac{x_1}{m}, \frac{x_3}{m}\right\} - min\left\{\frac{x_1}{m}, \frac{x_3}{m}\right\} \leq 1$
since we know that $x_1 \leq x_3 \leq x_2$, we get
$\Leftrightarrow \frac{x_2}{m} - \frac{x_1}{m} - \left(\frac{x_2}{m} - \frac{x_3}{m}\right) + \frac{x_3}{m} - \frac{x_1}{m} \leq 1$
$\Leftrightarrow -\frac{2x_1}{m} + \frac{2x_3}{m} \leq 1$
$\Leftrightarrow \frac{2}{m}(x_3 - x_1) \leq 1$
and $(x_3 - x_1) \leq 0 < 1$■

In case $|x_1 - x_2| \geq \frac{m}{2}$ and $|x_1 - x_3| \geq \frac{m}{2}$ we have two possible cases: $x_1 \leq x_2 \leq x_3$ and $x_1 \leq x_3 \leq x_2$. In case $x_1 \leq x_2 \leq x_3$, we have
$S(x_1, x_2) \odot S(x_2, x_3) \leq S(x_1, x_3)$
$\Leftrightarrow max\left\{\left|\frac{x_1}{m} - \frac{x_2}{m}\right| + 1 - \left|\frac{x_2}{m} - \frac{x_3}{m}\right| - 1{,}0\right\} \leq \left|\frac{x_1}{m} - \frac{x_3}{m}\right|$
$\Leftrightarrow max\left\{\left|\frac{x_1}{m} - \frac{x_2}{m}\right| - \left|\frac{x_2}{m} - \frac{x_3}{m}\right|, 0\right\} \leq \left|\frac{x_1}{m} - \frac{x_3}{m}\right|$
Case $\left|\frac{x_1}{m} - \frac{x_2}{m}\right| - \left|\frac{x_2}{m} - \frac{x_3}{m}\right| < 0$ is not possible, since $|x_1 - x_2| \geq \frac{m}{2}$ and $|x_1 - x_3| \geq \frac{m}{2}$. This leads to having
$\left|\frac{x_1}{m} - \frac{x_2}{m}\right| - \left|\frac{x_2}{m} - \frac{x_3}{m}\right| \leq \left|\frac{x_1}{m} - \frac{x_3}{m}\right|$
$\Leftrightarrow \left|\frac{x_1}{m} - \frac{x_2}{m}\right| - \left|\frac{x_2}{m} - \frac{x_3}{m}\right| - \left|\frac{x_1}{m} - \frac{x_3}{m}\right| \leq 0$
$\Leftrightarrow max\left\{\frac{x_1}{m}, \frac{x_2}{m}\right\} - min\left\{\frac{x_1}{m}, \frac{x_2}{m}\right\} - \left(max\left\{\frac{x_2}{m}, \frac{x_3}{m}\right\} - min\left\{\frac{x_2}{m}, \frac{x_3}{m}\right\}\right) - \left(max\left\{\frac{x_1}{m}, \frac{x_3}{m}\right\} - min\left\{\frac{x_1}{m}, \frac{x_3}{m}\right\}\right) \leq 0$
$\Leftrightarrow \frac{x_2}{m} - \frac{x_1}{m} + \frac{x_2}{m} - \frac{x_3}{m} - \frac{x_3}{m} + \frac{x_1}{m} \leq 0$
$\Leftrightarrow \frac{2}{m}(x_2 - x_3) \leq 0$
Since $x_2 \leq x_3$, $x_2 - x_3 \leq 0$, and we get 0≤ 0■

In case that we have $x_1 \leq x_3 \leq x_2$, we get

$$S(x_1, x_2) \odot S(x_2, x_3) \leq S(x_1, x_3)$$

$\Leftrightarrow max\left\{\left|\frac{x_1}{m} - \frac{x_2}{m}\right| + 1 - \left|\frac{x_2}{m} - \frac{x_3}{m}\right| - 1{,}0\right\} \leq \left|\frac{x_1}{m} - \frac{x_3}{m}\right|$

$\Leftrightarrow max\left\{\left|\frac{x_1}{m} - \frac{x_2}{m}\right| - \left|\frac{x_2}{m} - \frac{x_3}{m}\right|, 0\right\} \leq \left|\frac{x_1}{m} - \frac{x_3}{m}\right|$

Case $\left|\frac{x_1}{m}-\frac{x_2}{m}\right|-\left|\frac{x_2}{m}-\frac{x_3}{m}\right|<0$ gives $0\le\left|\frac{x_1}{m}-\frac{x_3}{m}\right|$ which is obviously valid. In case $\left|\frac{x_1}{m}-\frac{x_2}{m}\right|-\left|\frac{x_2}{m}-\frac{x_3}{m}\right|>0$, we have

$\left|\frac{x_1}{m}-\frac{x_2}{m}\right|-\left|\frac{x_2}{m}-\frac{x_3}{m}\right|\le\left|\frac{x_1}{m}-\frac{x_3}{m}\right|$
$\Leftrightarrow max\left\{\frac{x_1}{m},\frac{x_2}{m}\right\}-min\left\{\frac{x_1}{m},\frac{x_2}{m}\right\}-\left(max\left\{\frac{x_2}{m},\frac{x_3}{m}\right\}-min\left\{\frac{x_2}{m},\frac{x_3}{m}\right\}\right)-\left(max\left\{\frac{x_1}{m},\frac{x_3}{m}\right\}-min\left\{\frac{x_1}{m},\frac{x_3}{m}\right\}\right)\le 0$
$\Leftrightarrow\frac{x_2}{m}-\frac{x_1}{m}-\frac{x_2}{m}+\frac{x_3}{m}-\frac{x_3}{m}+\frac{x_1}{m}\le 0$
$\Leftrightarrow 0\le 0$■

In the last case we have $|x_1-x_3|\ge\frac{m}{2}$ and now $x_1\le x_2\le x_3$, which gives us
$S(x_1,x_2)\odot S(x_2,x_3)\le S(x_1,x_3)$
$\Leftrightarrow max\left\{1-\left|\frac{x_1}{m}-\frac{x_2}{m}\right|+1-\left|\frac{x_2}{m}-\frac{x_3}{m}\right|-1,0\right\}\le\left|\frac{x_1}{m}-\frac{x_3}{m}\right|$
$\Leftrightarrow max\left\{1-\left|\frac{x_1}{m}-\frac{x_2}{m}\right|-\left|\frac{x_2}{m}-\frac{x_3}{m}\right|,0\right\}\le\left|\frac{x_1}{m}-\frac{x_3}{m}\right|$
in case $1-\left|\frac{x_1}{m}-\frac{x_2}{m}\right|-\left|\frac{x_2}{m}-\frac{x_3}{m}\right|<0$ we again get $0\le\left|\frac{x_1}{m}-\frac{x_3}{m}\right|=S(x_1,x_3)\in[0,1]$ giving $0\le[0,1]$
in case $1-\left|\frac{x_1}{m}-\frac{x_2}{m}\right|-\left|\frac{x_2}{m}-\frac{x_3}{m}\right|>0$ we have
$1-\left|\frac{x_1}{m}-\frac{x_2}{m}\right|-\left|\frac{x_2}{m}-\frac{x_3}{m}\right|\le\left|\frac{x_1}{m}-\frac{x_3}{m}\right|$
$\Leftrightarrow 1-\left|\frac{x_1}{m}-\frac{x_2}{m}\right|-\left|\frac{x_2}{m}-\frac{x_3}{m}\right|-\left|\frac{x_1}{m}-\frac{x_3}{m}\right|\le 0$
$\Leftrightarrow 1-\left(\frac{x_2}{m}-\frac{x_1}{m}\right)-\left(\frac{x_3}{m}-\frac{x_2}{m}\right)-\left(\frac{x_3}{m}-\frac{x_1}{m}\right)\le 0$
$\Leftrightarrow 1-2\left(\frac{x_3}{m}-\frac{x_1}{m}\right)\le 0$
$\Leftrightarrow\left(\frac{x_3}{m}-\frac{x_1}{m}\right)\le\frac{1}{2}$, which is valid, since we have $|x_1-x_3|\ge\frac{m}{2}$ and $x_1\le x_2\le x_3$
Since there are no other cases this concludes the proof■

Theorem 2 Modulo similarity S_{mod} satisfies reflexivity and symmetricity

Proof: $S_{mod}(x_a,x_a)=1-|x_a-x_a|=1-0=1$■

if $|x_a-x_b|\le\frac{m}{2}$ we have

$S_{mod}(x_a,x_b)=1-|x_a-x_b|=1-|x_b-x_a|=S_{mod}(x_b,x_a)$

If $|x_a-x_b|>\frac{m}{2}$ we have

$S_{mod}(x_a,x_b)=|x_a-x_b|=|x_b-x_a|=S_{mod}(x_b,x_a)$■

Because all three axioms hold we conclude that modulo similarity is a similarity measure in the sense defined by Łukasiewicz [10] and Zadeh [8].

3. Similarities in comparing histograms

The similarity between any two histograms can be given in terms of sample value similarities. Given two samples of n elements, A and B we approach this problem by considering maximum compatibility of pair assignments between the two samples. The problem is to determine the best one-to-one assignment between the two samples, such that the mean of all similarities between two individual elements in a pair is maximized. Maximum pair compatibility is a new concept and a new contribution and therefore we first start with a definition and then clarify its usefulness together with modulo similarity. Given m elements $a_i\in A$, and m elements $b_i\in B$, we define the maximum pair assignment compatibility as:

***Definition 6**: Given $A=\{a_1,\dots,a_n\}$ and $B=\{b_1,\dots,b_n\}$ and bin number value m. Normalized values for A and B are $A_m=\frac{\{a_1,\dots,a_n\}}{m}$, $B_m=\frac{\{b_1,\dots,b_n\}}{m}$, and maximum pair assignment compatibility*

$$S(A_m,B_m)=\frac{1}{n}\overset{max}{A,B}\left(\sum_{i,j=1}^{n}s(a_i,b_j)\right)\qquad(6)$$

where S and s are designated as S_{nom} and s_{nom}, S_{ord} and s_{ord} and S_{mod} and s_{mod} respectively.

Example: Consider the following three samples with m=8 and n=10: *A={1,1,1,1,2,3,7,7,7,8}*, *B={1,2,2,2,2,3,7,7,7,8}* , *C={1,1,2,3,7,7,7,8,8,8}*. Corresponding histograms would be *H(A)={4,1,1,0,0,0,3,1}*, *H(B)={1,4,1,0,0,0,3,1}* and *H(C)={2,1,1,0,0,0,3,3}*. Now, applying maximum pair assignment similarity to these cases we get $A_m=\frac{A}{m}, B_m=\frac{B}{m}, C_m=\frac{C}{m}$, and $S_{nom}(A_m,C_m)=0.8$, $S_{ord}(A_m,C_m)=0.825$, $S_{mod}(A_m,C_m)=0.975$. A summary of the results by maximum pair similarity assignment is visible in Table 1.

Pairs:	S_{nom}	S_{ord}	S_{mod}
A,B	0.7	0.963	0.963
A,C	0.8	0.825	0.975
B,C	0.7	0.838	0.938

Table 1: Maximum pair similarity assignments.

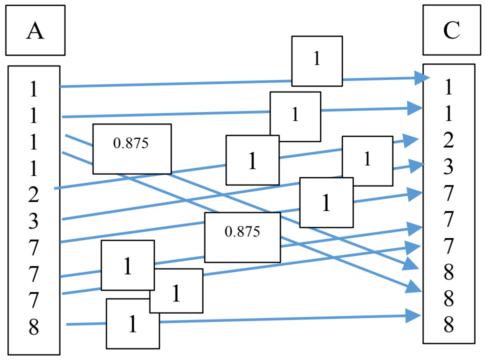

Figure 1: Similarities between H(A) and H(C), when modulo similarity is applied. With ordinal similarity the pair wise similarities would be the same, but the values "0.875" would be "0.125". With nominal similarity they would be" 0".

The procedure is illustrated in Figure 1. Note that if we deal with the modulo type histograms without acknowledging the modulo type and by using an ordinal type histogram, we would have the result that the pair A,B is the closest match. but when we use the more suitable modulo similarity to find the maximum pair similarity we get the result that the pair A, C is the closest match.

4. Conclusions

In this paper we have presented a new similarity measure, the "modulo similarity". We have proven that it fulfills the three axioms, reflexivity, symmetricity, and transitivity, required by a "true" similarity measure in the Łukasiewicz structure. We have shown its usefulness by applying it in comparison of histograms together with a new concept "maximum pair assignment compatibility" measure for histograms. We have demonstrated that using modulo similarity with modulo-type histograms results in distinctly different results than using, e.g., ordinal similarity, which is an important observation from the point of view of practical application.

There has been very little in terms of academic research into this type of problems so far. Histograms are an interesting way of visualizing, e.g., frequency information and modulo histograms a rather new way of presenting histograms.

Future research into this topic will include testing histograms with multiple concentrations of frequency (multiple peaks) and research into how the new methods may be used together with the histogram ranking method [14] to compare the results from different parametric MCDM decision-support methods.

References

[1] S-.H. Cha and S.N. Srihari, On measuring the distance between histograms, *Pattern Recognition* 35:1355-1370, 2002.

[2] R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*, 1st Edition, Wiley, New York, 1973.

[3] T. Kailath, The divergence and Bhattacharyya distance measures in signal selection, *IEEE Trans. Commun. Technol. COM-15,* 1:52–60, 1967.

[4] S. Kullback and R.A. Leibler, On information and sufficiency, *Ann. Math. Statist*. 22:79-86, 1951.

[5] H. Bandemer and W. Näther, *Fuzzy data analysis*, Kluwer academic publishers, Dordrecht, 1992.

[6] Y. Rubner, Tomasi, C., and L.J. Guibas, A metric for distributions with applications to image data base, proceedings of the *International Conference on Computer Vision,* pages 59-66, IEEE, 1998.

[7] J. Łukasiewicz, *Selected Works, North-Holland Publishing co.,* Amsterdam, 1970.

[8] L. Zadeh, Similarity Relations and Fuzzy Orderings. *Inform Sci*, 3, 1971.

[9] C.C. Chang, Algebraic analysis of many-valued logics, *Trans. Amer. Math. Soc*., 88:467-490, 1958.

[10] J. Pavelka. On Fuzzy logic I, II, III**.** *Zeitschr f. math. Logik und Grundlagen d. Math*., 25:45-52; 119.134; 447-464, 1979.

[11] E. Turunen, *Mathematics behind Fuzzy Logic*. Advances in Soft Computing, Physica-Verlag, Heidelberg, 1999.

[12] P. Luukka, K. Saastamoinen, and V. Könönen, A classifier based on the maximal fuzzy similarity in the generalized Łukasiewicz-structure. *In proceedings of the FUZZ-IEEE 2001 conference*, Melbourne, Australia.

[13] P. Luukka and T. Leppälampi, Similarity classifier with generalized mean applied to medical data. *Computers in Biology and Medicine*, 36:1026–1040.

[14] P. Luukka and M. Collan, Histogram ranking with generalized similarity-based TOPSIS applied to patent ranking, *Int. J. Operational Research*, In press.

[15] P. Luukka, Feature selection using fuzzy entropy measures with similarity classifier. *Expert Systems with Applications*, 38:4600-4607, 2011.

[16] S. Bray, L. Caggiani, M. Dell'Orco, and M. Ottomanelli, Feature selection based on fuzzy entropy for data envelopment analysis applied to transport systems, *Transportation Research Procedia*, 3:602-610, 2014.

[17] J. Niittymäki and E. Turunen, Traffic signal control on similarity logic reasoning. *Fuzzy Sets and Systems,* 133:109-131, 2003.

[18] T. Dubrovin, A. Jolma, and E. Turunen, Fuzzy model for real-time reservoir operation. *Journal of Water Resources Planning and Management* 128:66-73, 2002.

[19] K. Saastamoinen, J. Ketola, and E. Turunen, Defining Athletes′ Aerobic and Anaerobic Thresholds by Using Similarity Measures and Differential Evolution, proceedings of the *2004 IEEE Int. Conference of Systems, Man and Cybernetics,* 1331-1335, IEEE, 2004.

[20] A. Mänttäri, P. Luukka, and E. Turunen, Predicting maximal heart rate from the UKK 2 km walk test results: soft computing and linear regression methods, proceedings of the *8th annual congress of the European College of Sports Science*, pages 9-12, Austria, 2003.

[21] E. Turunen, Survey of Theory and Applications of Łukasiewicz-Pavelka Fuzzy Logic. in *Lectures on Soft Computing and Fuzzy Logic. Advances in Soft Computing,* 313-337, Physica-Verlag, Heidelberg, 2001.