International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

A New Semimetric for Interval Data

Irani Hazarika, Anjana Kakoti Mahanta

Abstract*: *Interval data, a special case of symbolic data, is becoming more and more frequent in different fields of applications including the field of Data Mining. Measuring the dissimilarity or similarity between two intervals is an important task in Data Mining. In this paper an analysis of ten desirable properties that should be fulfilled by the measures for interval data for making it suitable for applications like clustering and classification has been done. Also, it has been verified whether these properties are satisfied by three existing measures- L_1-norm, L_2-norm, L_∞-norm and also a new dissimilarity measure for interval data has also been proposed. The performance of all the existing distance measures are compared with the proposed measure by applying well known K-Means algorithm on 6 interval datasets. It is seen that proposed measure gives better clustering accuracy then the existing measures on most of the datasets.

Keywords*: *Distance measure, Interval data, Interval data clustering, Semimetric

I. INTRODUCTION

In many real-life applications, we come across interval data. Any interval-valued data generalizes a single-valued data because it represents a range of values between two bounds. Each interval can be represented uniquely by a lower bound l and upper bound r and denoted as $[l, r]$. These l and r are also called starting point and ending point of an interval. Defining a suitable distance measure between two intervals is very much different from defining the distance between other types of data such as numeric values, points in multidimensional space or data having numeric/categorical features. This is mainly because there will be an overlapped area between any two intervals, even though the overlapped area might be empty. Thus while defining distances, apart from the distance between the endpoints, the size of the overlapping area also needs to be considered.

Interval-valued data are found in many real-life data mining applications such as weather forecasting, medical data analysis, stock market analysis etc. In some cases, measuring precisely or finding the exact value of an attribute is difficult. In certain cases, uncertainty may exist in the observations. In those cases, values of the attributes are considered as intervals. Sometimes attribute values are converted to intervals to reduce large data volume or sometimes to preserve a level of confidentiality (e.g. converting actual salary values to salary slabs, converting marks obtained by students to grades). The data may also come from intrinsically interval-valued random elements (e.g. daily temperature/humidity range for a location, the daily fluctuation of the blood pressure of a patient).

There are 13 possible relationships that may occur between two intervals [1]. From the point of distance measure, the 13 possible relationships [1] can be reduced to only three relations before, overlap and contain [2]. In-spite of the difference between interval and numeric data, several distance measures used for numeric data are used in interval data [2-7]. These measures may or may not consider the relationship between the intervals.

The aim of this paper is to analyze some desirable properties of distance measures for interval data. Also it has been verified whether these properties are satisfied by a number of standard distance measures for interval data. After analyzing the properties for three different existing distance measures, a new distance measure for interval data has been proposed.

II. BACKGROUNDS ON INTERVAL DATA

In this section formal definition of an interval is given together with various representation methods. Arithmetic operators that are used in this work and relations that hold on interval data are also stated.

A. Definition of Interval

Given a totally ordered domain D, a non-empty subset I of D is called an interval iff for all I^-, $I^+ \in I$ and $c \in D$, $I^- \le c \le I^+$ implies $c \in I$. If for a given interval I, $I^- \le c \le I^+$ holds for all $c \in I$ where I^- and I^+ are two specific elements in I then I is denoted by $[I^-, I^+]$, and I^- is the left end point (or lower bound) and I^+ is called the right end point (or upper bound) of I. The intervals that we consider are all closed intervals because for any $c \in I$, $I^- \le c \le I^+$.

B. Representations of Intervals

Intervals can be represented in one or two dimensional space as follows-

- *Interval as Line:* An interval $I = [I^-, I^+]$ can be represented as a line segment (Fig.1) in one- dimensional space. The ends I^-, I^+ denote the ends of the line. Thus, Intervals from same domain can be represented as a set of lines along an axis. If the intersection between any two intervals is nonempty then an overlapping between the lines are seen. Two intervals from same domain can be mainly categorized as overlapping or non-overlapping.

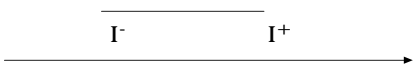

Fig.1. An interval $I = [I^-, I^+]$ as a line segment in one-dimensional space

Manuscript published on 30 September 2019
* Correspondence Author
Irani Hazarika*, Department of Computer Science, Gauhati University, Guwahati, India. Email:queensarathi@gmail.com
Anjana Kakoti Mahanta, Department of Computer Science, Gauhati University, Guwahati, India.. Email: anjana@gauhati.ac.in

- *Interval as Point:* An interval $I = [I^- I^+]$ can also be represented as points in two-dimensional space (shown in Fig. 2) and therefore a set of intervals from the same domain can be represented as a set of points in two-dimensional space. The disadvantage of this representation is that if any two intervals are overlapping in one place then it is not visible in its point-based representation.

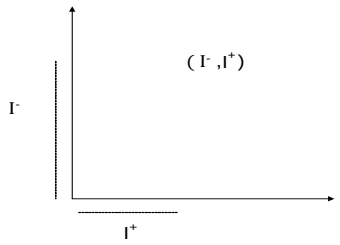

Fig.2. An interval $I = [I^-, I^+]$ as a point in two-dimensional spaces.

- *Interval as Circle:* Intervals can also be represented as circles (Fig. 3) in two-dimensional planes. Thus any interval I = [I⁻, I⁺] can be defined in terms of center and radius of its circle.

Thus, center and radius of I will be –

Center, $c = \frac{I^- + I^+}{2}$ Radius, $r = \frac{I^+ - I^-}{2}$

Again, left end, $I^- = c - r$
right end, $I^+ = c + r$

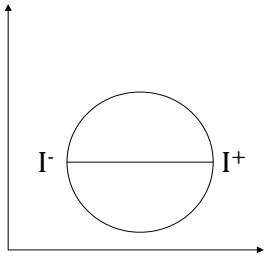

Fig.3. An interval $I = [I^-, I^+]$ as a circle in two-dimensional spaces

C. Arithmetic of Intervals

The following arithmetic concepts are defined for intervals [11]-

- The width and midpoint of an interval $I = [I^-, I^+]$ are defined as-
  i. Width,(| I |) = $I^+ - I^-$
  ii. Midpoint, (mid_I)= $\frac{(I^- + I^+)}{2}$
- Suppose, $I_1 = [I_1^-, I_1^+]$ and $I_2 = [I_2^-, I_2^+]$denote any two intervals then-
  i. Meet, $I_1 \cap I_2$: Meet of two intervals $I_1 = [I_1^-, I_1^+]$ and $I_2 = [I_2^-, I_2^+]$ is empty (i.e. $I_1 \cap I_2 = \phi$) if $I_1^+ > I_2^-$ or $I_2^+ > I_1^-$.
  Otherwise, Meet, $I_1 \cap I_2 = [\max(I_1^-, I_2^-), \min(I_1^+, I_2^+)]$
  ii. Join, $I_1 \oplus I_2 = [\min(I_1^-, I_2^-), \max(I_1^+, I_2^+)]$
  iii. Contain, $I_1 \subseteq I_2$ iff $I_1^- \leq I_2^-$ and $I_1^+ \geq I_2^+$
  iv. Properly Contain, $I_1 \subset I_2$ iff $I_1^- < I_2^-$, $I_1^+ \geq I_2^+$ or $I_1^- \leq I_2^-$, $I_1^+ > I_2^+$

- Suppose there are n intervals $I_1, I_2, \ldots, I_n$ in domain J. Then width of domain J will be

$$|Dom_j| = |I_1 \oplus I_1 \oplus I_3 \oplus \text{------} \oplus I_n|$$

D. Possible Relationship between Two Intervals

If we consider intervals as a set of lines along an axis, then the intervals can be categorized as non-overlapping and overlapping. Suppose $I_1 = [I_1^-, I_1^+]$, $I_2 = [I_2^-, I_2^+]$ are two intervals. If $I_1 \cap I_2 = \emptyset$ then they are called non-overlapping otherwise they are called overlapping intervals. Having seen the positions of the intervals along the axis, in the paper [1] Allen proposed 13 possible relationships that may occur between any two intervals as before, after, meets, meet by, overlap, overlap by, during, includes, starts, start by, finishes, finished by, equal. Among these relations, some relations are the mirror image of others. For example, '*X* before *Y*' and '*Y* after *X*' denotes the same relation. Thus, these 13 relationships can be reduced to only 7 relationships- before, meet, overlap, during, start, finishes, and equal. For the purpose of distance measure, in paper [2] authors have mentioned that only three possible relationships may be considered between two intervals. They defined these as Before, Overlap, Contains. The 'Before' relation holds when the two intervals are non-overlapping and 'Overlap', 'Contains' hold for overlapping intervals. The mirror relations are 'After', 'Overlapped by', 'Contained by' respectively.

Suppose, $I_1 \cap I_2 = \emptyset$ and I_1 before I_2 or I_1 after I_2 is true. Then, *Separation* $(I_1, I_2) = |I_1^+ - I_2^-|$, which increases when the I_1 and I_2 are more apart from each other.

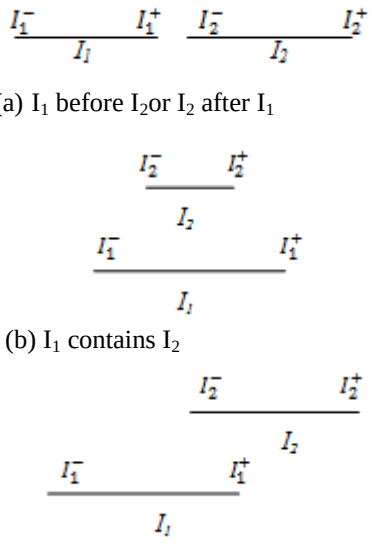

Fig.4. Relationships between two intervals I_1 and I_2

III. STANDARD DISTANCE MEASURES FOR INTERVAL DATA

As said earlier, intervals can be represented in one or two dimensional space as a line, as a point or as a circle. By considering interval as point in two dimensional spaces, it is possible to measure the distance between two intervals as the distance between two points. If intervals are considered as lines,

then distance between the intervals can be computed in terms of their boundary values [6]. Similarly, by considering the intervals as circles in two dimensional spaces, it is possible to express the distance between two intervals in terms of centre and radius.

Suppose $I_1 = [I_1^-, I_1^+]$, $I_2 = [I_2^-, I_2^+]$ are two intervals. Let, c_1, c_2 and r_1, r_2 denote the centre and radii of I_1 and I_2 respectively then,

$$\text{Center, } c_1 = \frac{I_1^- + I_1^+}{2} \qquad \text{Radius, } r_2 = \frac{I_2^+ - I_2^-}{2}$$

There are various existing distance measures to measure the distance between intervals I_1 and I_2. In this section a study of three existing distance measures of interval data has been reported.

A. L_1-norm or City Block distance

The L_1-norm or City Block distance or Manhattan distance for a pair of interval is a suitable extension of the L_1 Minkowski distance [2]. By considering intervals as points, the L_1- norm between I_1 and I_2 can be expressed as the sum of the absolute distances between its ends.

$$D_{L_1}(I_1, I_2) = |I_1^- - I_2^-| + |I_1^+ - I_2^+| \text{ ---------- (1)}$$

Again, L_1-norm in (1) can be rewritten in terms of center and radius as-

$$D_{L_1}(I_1, I_2) = \begin{cases} 2|r_1 - r_2| & if\ I_1 \subseteq I_2\ or\ I_2 \subseteq I_1 \\ 2|c_1 - c_2| & otherwise \end{cases} \quad \text{---(2)}$$

B. L_2-norm or Euclidean distance

In mathematics, the Euclidean distance or L_2-norm is the "ordinary" straight-line distance between two points in Euclidean space. Using L_2- norm [7] the distance between I_1 and I_2 can be expressed as the square root of the sum of the square distances between its ends.

$$D_{L_2}^2(I_1, I_2) = |I_1^- - I_2^-|^2 + |I_1^+ - I_2^+|^2 \text{ ---------- (3)}$$

Again, L_2-norm can be rewritten in terms of center and radius as

$$D_{L_2}^2(I_1, I_2) = 2|c_1 - c_2|^2 + 2|r_2 - r_1|^2 \text{ --------- (4)}$$

C. L_∞- norm

Using the L_∞- norm [2] the distance between I_1 and I_2 can be expressed as the maximum of the absolute distances between their ends.

$$D_{L_\infty}(I_1, I_2) = \max(|I_1^- - I_2^-|, |I_1^+ - I_2^+|) \text{ -------- (5)}$$

L_∞-norm can be rewritten in terms of center and radius as-

$$D_{L_\infty}(I_1, I_2) = |c_1\text{-}c_2| + |r_1\text{-}r_2| \text{ ---------- (6)}$$

IV. Properties and Characteristics of Distance Measures for Interval Data

There are many applications in data mining like clustering, classification etc. in which distance measures are used. In many real life applications objects are described by means of intervals. Sometimes one or more features of objects may be of interval type. Also in many clustering algorithms clusters are represented by single representatives. For numeric data the representative may be the median, mode or other central tendencies of the objects in the cluster. For interval type objects a suitable representative will have to be defined. Keeping in mind applications of these types, in this section some desirable properties for distance measures of interval data are stated. Then it has been verified whether these properties are satisfied by L_1-norm, L_2-norm, L_∞-norm for interval data.

Suppose, I_1 and I_2 are two intervals and $D(I_1, I_2)$ denotes the distance between I_1 and I_2. Then the measure D is called a **metric** or distance measure if it satisfies the following properties [9]-

P1. $D(I_1, I_2) \geq 0$; non-negativity
P2. $D(I_1, I_2) = 0$; if and only if $I_1 = I_2$
P3. $D(I_1, I_2) = D(I_2, I_1)$; symmetric
P4. $D(I_1, I_3) \leq D(I_1, I_2) + D(I_2, I_3)$; triangular inequality.

A measure which partially satisfies these properties is called a **divergence or dissimilarity or semi-metric** measure.

All the measures **L_1-norm, L_2-norm, L_∞-norm are metric** as they satisfy all the properties P1, P2, P3, P4.

By analysing the relations before, contains, overlaps in the context of clustering of interval data, we had observed that distance measures satisfying the following properties, would be superior to the others.

P5. Let, I_1 and I_2 are two intervals. Assume that width of I_1 and I_2 are fixed. If I_1 and I_2 are disjoint, $D(I_1, I_2)$ should become larger, when I_1 and I_2 are farther apart from each other i.e. $D(I_1, I_2)$ should become larger when $Separation(I_1, I_2)$ becomes larger [2]. In Fig.5. $D(I_1, I_2)$ should become larger when $d = |I_2^- - I_1^+|$ becomes larger.

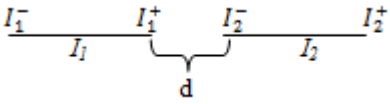

Fig.5. An example to illustrate P5

Suppose $|I_x|$ and $|I_y|$ are fixed. If $I_x \cap I_y = \phi$ then all the measures mentioned in Section III can be rewritten as follows-

- **L_1-Norm:** $D_{L_1}(I_1, I_2) = |I_1| + |I_2| + 2 \times Separation(I_1, I_2)$ ---- (7)
- **L_2-Norm:** $D_{L_2}^2(I_1, I_2) = [|I_1| + Separation(I_1, I_2)]^2 + [|I_2| + Separation(I_1, I_2)]^2$ ---- (8)
- **L_∞-Norm:** $D_{L_\infty}(I_1, I_2) = \max([|I_1| + \text{Separation}(I_1, I_2)], [|I_2| + \text{Separation}(I_1, I_2)])$ ------- (9)

From (7), (8) and (9), it can be said that if $Separation(I_1, I_2)$ increases then distance between I_1 and I_2 increases for all of the above measures. Thus all **L_1-norm, L_2-norm, L_∞-norm satisfy property P5.**

P6. Let, I_1 and I_2 are two intervals. Assume that width of $I_1$and I_2 are fixed. If I_1 and I_2 overlap, $D(I_1, I_2)$ should become smaller, when overlapped interval becomes larger i.e. when $I_1 \cap I_2$ becomes larger then $D(I_1, I_2)$ should become smaller [2]. In Fig.6. $D(I_1, I_2)$ should become smaller when $d = |I_1^+ - I_2^-|$ becomes larger.

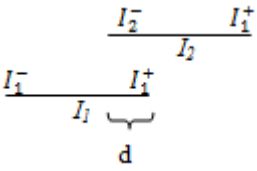

Fig.6. An example to illustrate P6

Suppose $|I_1|$ and $|I_2|$ are fixed. If $I_1 \cap I_2 \neq \phi$ then the measures L_1-Norm, L_2-Norm, L_∞-Norm can be rewritten as follows-

- **L_1-Norm:** $D_{L_1}(I_1, I_2) = |I_1| + |I_2| - 2|I_1 \cap I_2|$ ---------- (10)
- **L_2-Norm:** $D^2_{L_2}(I_1, I_2) = [|I_1| - |I_1 \cap I_2|]^2 + [|I_2| - |I_1 \cap I_2|]^2$ ---------- (11)
- **L_∞-Norm:** $D_{L_\infty}(I_1, I_2) = \max([|I_1| - |I_1 \cap I_2|], [|I_2| - |I_1 \cap I_2|])$ ---------- (12)

From (10), (11) and (12), it can be said that if overlapping area between I_1 and I_2 decreases then distance between I_1 and I_2 increases. Thus, all **L_1-norm, L_2-norm, L_∞-norm satisfy property P6.**

P7. Let, I_1 and I_2 are two intervals. If I_1 encloses I_2, $D(I_1, I_2)$ should be invariant, regardless of position of I_2 within I_1 [2]**.**

- **L_1-Norm:** From (2), it is seen that when relation between the intervals is 'Contain', then distance between I_1 and I_2 is twice the absolute differences between its radii r_1 and r_2 i.e. if I_1 contains I_2 then $D_{L_1}(I_1, I_2) = 2(r_1 - r_2)$.When lengths of I_1 and I_2 are fixed then length values of r_1 and r_2 are also fixed. Thus, changing the position of one interval does not affect the distance between them. Thus, **the propertyP7 holds for L_1-Norm**.

- **L_2-Norm:** From (4), it is seen that for each relation the distance between I_1 and I_2 is dependent on the distance between their centers. Thus, the distance between I_1, I_2 will change when position of I_2 within I_1 changes. Thus, **the propertyP7 does not hold for L_2-Norm**.

- **L_∞-Norm:** From (6), it is seen that for each relation the distance between I_1 and I_2 is dependent on their relative positions (distance between their centres) with respect to each other. Thus when I_1 contains I_2 (or I_2 contains I_1) then distance between I_1 and I_2 will change if position of I_2 changes within I_1. Thus, **the propertyP7 does not hold for L_∞-Norm**.

P8. Let, I_1, I_2 and I_3 are three intervals. If $I_3 \subset I_2$or $I_2 \subset I_3$ then distances $D(I_1, I_2)$ and $D(I_1, I_3)$ should not be same. (Fig.7) [18] .

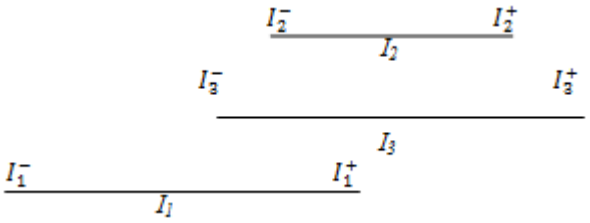

Fig.7. An example to illustrate P8. $D(I_1, I_2)$ and $D(I_1, I_3)$ should not be same

- **L_1-Norm: The property P8 does not hold for L_1-norm.**

  For example-Suppose, I_2=[12, 18], I_3=[14, 16] are two intervals such that $I_3 \subset I_2$..

  Now, if I_1=[4, 15] is an another interval then from (10),
  $D_{L_1}(I_1, I_2) = D_{L_1}(I_1, I_3) = 11$

- **L_2-Norm: The property P8 does not hold for L_2-norm.**

  For example- Suppose, I_2=[5, 21], I_3=[7, 19] are two intervals such that $I_3 \subset I_2$..

  Now, if I_1=[2, 16] is an another interval then from (11),
  $D_{L_2}(I_1, I_2) = D_{L_2}(I_1, I_3) = \sqrt{34}$

- **L_∞-Norm: The property P8 does not hold for L_∞-norm.**

  For example- Suppose, I_2= [7, 11], I_3= [8, 10] are two intervals such that $I_3 \subset I_2$.

  Now if I_1= [6, 9] then from (12), then
  $D_{L_\infty}(I_1, I_2) = D_{L_\infty}(I_1, I_3) = 2$

P9: Let, I_1, I_2 and I_3 are three intervals. If $|I_1 \cap I_2| = |I_1 \cap I_3| \neq \phi$ then distances $D(I_1, I_2)$ and $D(I_1, I_3)$ should not be same when $|I_2| \neq |I_3|$ (Fig.8.) [10].

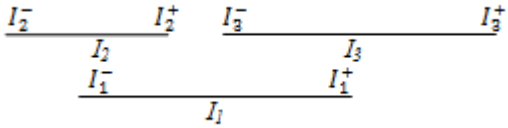

Fig.8. An example to illustrate P9. $D(I_1, I_2)$ and $D(I_1, I_3)$ should not be same. Actually $D(I_1, I_2) < D(I_1, I_3)$ should be true.

- **L_1-Norm: The property P9 holds for L_1-norm.**

  Suppose $I_1 \cap I_2 \neq \emptyset$ and $I_1 \cap I_3 \neq \emptyset$, then from (10),

  $D_{L_1}(I_1, I_2) = |I_1| + |I_2| - 2|I_1 \cap I_2|$ ---------- (13)
  $D_{L_1}(I_1, I_3) = |I_1| + |I_3| - 2|I_1 \cap I_3|$ ---------- (14)

From (13) and (14), it can be concluded that if $|I_1 \cap I_2| = |I_1 \cap I_3|$ then $D_{L_1}(I_1, I_2)$ and $D_{L_1}(I_1, I_3)$ will never be same when $|I_2| \neq |I_3|$.

- **L_2-Norm: The property P9 holds for L_2-norm.**

Suppose $I_1 \cap I_2 \neq \emptyset$ and $I_1 \cap I_3 \neq \emptyset$, then from (8),

$D^2_{L_2}(I_1, I_2) = [|I_1| - |I_1 \cap I_2|]^2 + [|I_2| - |I_1 \cap I_2|]^2$--(15)
$D^2_{L_2}(I_1, I_3) = [|I_1| - |I_1 \cap I_3|]^2 + [|I_3| - |I_1 \cap I_3|]^2$-- (16)

Thus, from (15) and (16) it can be concluded that if $|I_1 \cap I_2| = |I_1 \cap I_3|$ then $D^2_{L_2}(I_1, I_2)$ and $D^2_{L_2}(I_1, I_3)$ will never be same when $|I_2| \neq |I_3|$.

- **L_∞-Norm: The property P9 does not hold for L_∞-norm.**

For example- Suppose, I_1= [10, 20], I_2= [15, 22], I_3= [15, 24], then $|I_1 \cap I_2| = |I_1 \cap I_3| = 5$ and $|I_2| \neq |I_3|$.

Now, $D_{L_\infty}(I_1, I_2)$ = max [(15-10), (22-20)] =5
$D_{L_\infty}(I_1, I_3)$ = max [(15-10), (24-20)] =5

Thus, $D_{L_\infty}(I_1, I_2) = D_{L_\infty}(I_1, I_3)$

P10: Let, I_1, I_2 and I_3 are three intervals. If *Separation*(I_1, I_2) = *Separation*(I_1, I_3)then distances $D(I_1, I_2)$ and $D(I_1, I_3)$ should not be same when $|I_2| \neq |I_3|$ (Fig.9.) [10].

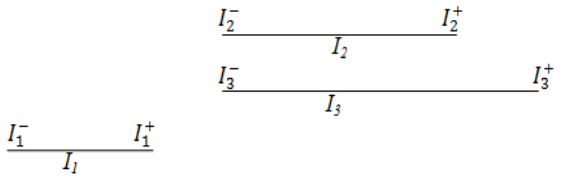

Fig.9. An example to illustrate P10. $D(I_1, I_2)$ and $D(I_1, I_3)$ should not be same.

- **L_1-Norm: The property P10 holds for L_1-norm.**

Suppose, $I_1 \cap I_2 = \emptyset$ and $I_1 \cap I_3 = \emptyset$ then from (7),

$D_{L_1}(I_1, I_2) = |I_1| + |I_2| + 2 \times$ *Separation* (I_1, I_2) ---------- (17)
$D_{L_1}(I_1, I_3) = |I_1| + |I_3| + 2 \times$ *Separation* (I_1, I_3) ---------- (18)

Thus, from (17) and (18), it can be concluded that if *Separation* (I_1, I_2)=*Separation*(I_1, I_3) then $D_{L_1}(I_1, I_2)$ and $D_{L_1}(I_1, I_3)$ will never be same when $|I_2| \neq |I_3|$.

- **L_2-Norm: The property P10 holds for L_2-norm.**

Suppose, $I_1 \cap I_2 = \emptyset$ and $I_1 \cap I_3 = \emptyset$ then from (8),

$D^2_{L_2}(I_1, I_2) = [|I_1| + \text{Separation}(I_1, I_2)]^2 + [|I_2| + \text{Separation}(I_1, I_2)]^2$ ---------- (19)
$D^2_{L_2}(I_1, I_3) = [|I_1| + \text{Separation}(I_1, I_3)]^2 + [|I_3| + \text{Separation}(I_1, I_3)]^2$ ---------- (20)

From (19) and (20) it can be conclude that if *Separation* (I_1, I_2)=*Separation* (I_1, I_3) then $D^2_{L_2}(I_1, I_2)$ and $D^2_{L_2}(I_1, I_3)$ will never be same when $|I_2| \neq |I_3|$.

- **L_∞-Norm: The property P10 does not hold for L_∞-norm.**

For example- Suppose, I_1=[10, 15], I_2=[20, 22], I_3=[20, 24] then *Separation* (I_1, I_2) = *Separation* (I_1, I_3) = 10 and $|I_2| \neq |I_3|$.

Now, $D_{L_\infty}(I_1, I_2)$ = max [(20-10), (22-15)] =10
$D_{L_\infty}(I_1, I_3)$ = max [(20-10), (24-15)] =10

Thus, $D_{L_\infty}(I_1, I_2) = D_{L_\infty}(I_1, I_3)$

V. PROPOSED MEASURE

From the above analysis it is seen that the distance measures L_1-norm, L_2-norm, L_∞-norm do not satisfy all properties. Thus, we propose a new dissimilarity measure or divergence for interval data which satisfies all the above properties except triangular inequality.

Suppose I_1 and I_2 are two intervals of domain J (Dom_j) then-

$$D_{proposed}(I_1, I_2) = D_{L_1}(I_1, I_2).\left(1 - \frac{|I_1 \cap I_2|}{|Dom_j|}\right) \text{ ----- (21)}$$

When $|I_1 \cap I_2| = 0$ then the measure is same as L_1-norm. If the intervals are not disjoint then the ratio of the size of overlapping to the total size of domain is computed. Larger the ratio value, lesser is the distance between the intervals.

Since the domain size is taken into account, this measure will be applicable only for finite domains. In real database of data mining applications underlying domains are in general finite. The use of domain size is meaningful here because the same amount of overlapping in two different size domains will mean different degrees of similarity.

The analysis of properties P1-P10 for the proposed measure has been given below-

P1-P4: Proposed measure is a semi-metric

The proposed measure is a dissimilarity or divergence or semi-metric because it satisfies only the properties P1, P2 and P3.

- **P1:** $D_{proposed}(I_1, I_2) \geq 0$; non-negative
- **P2:** $D_{proposed}(I_1, I_2) = 0$; if and only if $I_1 = I_2$
  It may be noted that, $1 - \frac{|I_1 \cap I_2|}{|Dom_j|}$ is zero,
  only if $|I_1 \cap I_2| = |Dom_j|$
  i.e. $|I_1| = |I_2| = |Dom_j|$
  i.e. if $I_1 = I_2$
- **P3:** $D_{proposed}(I_1, I_2) = D_{proposed}(I_2, I_1)$; symmetric
- **P4:** But, the proposed measure does not satisfy the property triangular inequality (property P4).

For example: Suppose, I_1=[10, 30], I_2=[20, 40], I_3=[100, 120], then $|I_1 \cap I_2|$= 10.

$D_{proposed}(I_1, I_2)=D_{L_1}(I_1, I_2).(1-\frac{I_1 \cap I_2}{|Dom_j|})$
$= 20 \text{ x } (1-0.09) = 18.2$
$D_{proposed}(I_2, I_3)=D_{L_1}(I_1, I_2).(1-\frac{I_2 \cap I_3}{|Dom_j|})$
$= 160 \text{ x} 1 = 160$
$D_{proposed}(I_1, I_3)=D_{L_1}(I_1, I_2).(1-\frac{I_1 \cap I_3}{|Dom_j|})$
$= 180 \text{ x } 1 = 180$

Now, $D_{proposed}(I_1, I_2) + D_{proposed}(I_2, I_3) = 178.2$

Thus,
$D_{proposed}(I_1, I_3) \leq D_{proposed}(I_1, I_2) + D_{proposed}(I_2, I_3)$ is not true i.e. the proposed measure does not satisfy triangular inequality.

- **P5-P6: Properties P5 and P6 hold for proposed measure**

If $I_1 \cap I_2 = \emptyset$, then the proposed distance in (21) can be reduced to L_1-norm. Thus the proposed measure satisfies property P5.

If $I_1 \cap I_2 \neq \emptyset$, then (21) can be rewritten as-

$$D_{proposed}(I_1, I_2) = (|I_1|+|I_2| - 2 \text{ x } |I_1 \cap I_2|).(1 - \frac{|I_1 \cap I_2|}{|Dom_j|}) \quad \text{---------- (22)}$$

From (22), it is seen that when $|I_1 \cap I_2|$ increases then distance $D_{proposed}(I_1, I_2)$ decreases. So the proposed measure satisfies property P6.

- **P7: Property P7 holds for proposed measure**

Suppose, $I_2 \subseteq I_1$ then $|I_1 \cap I_2| = |I_2|$.
Thus, from (22),

$$D_{proposed}(I_1, I_2) = (|I_1|-|I_2|).(1 - \frac{|I_2|}{|Dom_j|}) \quad \text{---------- (23)}$$

In (23) it is seen that this distance is not dependent on the position of I_2 inside I_1. Thus, the proposed measure satisfies property P7 also.

- **P8: Property P8 holds for proposed measure**

Suppose, I_1, I_2 and I_3 are three intervals. Then from (21),

$$D_{proposed}(I_1, I_2) = D_{L_1}(I_1, I_2).(1 - \frac{|I_1 \cap I_2|}{|Dom_j|})$$
$$D_{proposed}(I_1, I_3) = D_{L_1}(I_1, I_3).(1 - \frac{|I_1 \cap I_3|}{|Dom_j|})$$

Suppose, $I_3 \subset I_2$, then $|I_1 \cap I_3| \leq |I_1 \cap I_2|$.

(a) Let, $|I_1 \cap I_3| = |I_1 \cap I_2|$
Then $D_{L_1}(I_1, I_2) > D_{L_1}(I_1, I_3)$ and $(1-\frac{|I_1 \cap I_2|}{|Dom_j|}) = (1-\frac{|I_1 \cap I_3|}{|Dom_j|})$
Thus, $D_{proposed}(I_1, I_2) > D_{proposed}(I_1, I_3)$ ---------- (24)

(b) Let, $|I_1 \cap I_3| < |I_1 \cap I_2|$ then $(1-\frac{|I_1 \cap I_2|}{|Dom_j|}) < (1-\frac{|I_1 \cap I_3|}{|Dom_j|})$
Thus, $D_{proposed}(I_1, I_2) < D_{proposed}(I_1, I_3)$ ---------- (25)
Hence from (24) and (25), it can be concluded that P8 holds for the proposed measure.

- **P9: Property P9 holds for proposed measure**

Suppose, $I_1 \cap I_2 \neq \emptyset$ and $I_1 \cap I_3 \neq \emptyset$, then the proposed measure can be rewritten as-

$$D_{proposed}(I_1, I_2) = (|I_1|+|I_2| - 2 \text{ x } |I_1 \cap I_2|).(1 - \frac{|I_x \cap I_y|}{|Dom_j|})$$
$$D_{proposed}(I_1, I_3) = (|I_1|+|I_3| - 2 \text{ x } |I_1 \cap I_3|).(1 - \frac{|I_x \cap I_z|}{|Dom_j|})$$

From above it is seen that if $|I_1 \cap I_2| = |I_1 \cap I_3|$ then $D_{proposed}(I_1, I_2) \neq D_{proposed}(I_1, I_3)$ when $|I_2| \neq |I_3|$. Thus, property P9 holds for the proposed measure.

- **P10: Property P10 holds for proposed measure**

Suppose $I_1 \cap I_2 = \emptyset$ and $I_1 \cap I_3 = \emptyset$, then the proposed measure can be rewritten as-

$$D_{proposed}(I_1, I_2) = D_{L_1}(I_1, I_2)$$
$$= |I_1|+|I_2| + 2 \text{ x } Separation(I_1, I_2)$$
$$D_{proposed}(I_1, I_3) = D_{L_1}(I_1, I_3)$$
$$= |I_1|+|I_3| + 2 \text{ x } Separation(I_1, I_3)$$

From above it is seen that if $Separation(I_1, I_2) = Separation(I_1, I_3)$ then $D_{proposed}(I_1, I_2) \neq D_{proposed}(I_1, I_3)$ when $|I_2| \neq |I_3|$.

VI. Experimental Results

A. Dataset description

For the purpose of experimentation we consider 3 real life interval datasets- Car dataset (https://lhedjazi.jimdo.com/useful-links), Fish dataset (https://lhedjazi.jimdo.com/useful-links), Water dataset (http://cpcb.nic.in/data2005.php).

- The Car dataset consists of 33 samples with 8 interval features. The samples are spread across the 4 different classes with 10, 8, 8, and 7 samples per class respectively.
- The Fish dataset consists of 12 samples with 13 interval features. The samples are spread across 4 different classes, each with 4, 2, 4, and 2 samples respectively.
- The water dataset has water quality information from 32 states of India. It has 1600 samples with 8 interval features.

Apart from these real-life data sets, 3 synthetic interval data sets are generated online (http://www.freedatagenerator.com/csv-data-generator) as follows-

- ***Synthetic Interval Dataset 1:*** The first synthetic dataset generated has 300 samples with 3 interval features and 3 non-overlapping classes. Each class has 100 samples. For each feature, lower and upper values of the intervals are generated in the range as shown below-

Classes	Feature 1		Feature 2		Feature 3	
	Lower	Upper	Lower	Upper	Lower	Upper
Class 1	0-50	51-100	20-80	100-180	50-200	205-300
Class 2	101-15(	200-300	300-350	400-600	600-900	950-1000
Class 3	180-200	310-350	380-400	650-700	950-980	1050-1100

- ***Synthetic Interval Dataset 2:*** The second synthetic dataset is generated by adding 60 random samples (20 for each class) to the first Synthetic Interval Dataset 1. For these 60 samples, lower and upper values of the intervals are generated in the range as shown below-

Classe	Feature 1		Feature 2		Feature 3	
	Lower	Upper	Lower	Upper	Lower	Upper
Class 1	0-200	51-350	20-400	100-700	50-980	205-1100
Class 2						
Class 3						

- ***Synthetic Interval Dataset 3:*** The third synthetic dataset generated has 600 samples with 3 interval features and 3 overlapping classes. Each class has 200 samples. For each feature, lower and upper values of the intervals are generated in the range as shown below-

Classes	Feature 1		Feature 2		Feature 3	
	Lower	Upper	Lower	Upper	Lower	Upper
Class 1	1-40	51-80	10-30	50-70	20-80	90-100
Class 2	30-60	65-90	25-40	65-90	75-90	95-120
Class 3	38-70	77-100	27-50	68-95	79-100	96-130

B. Experimental procedure

Again, as mentioned earlier one of the applications of distance measure is clustering. Thus, to evaluate the performance of all the measures (i.e. L_1-norm, L_2-norm, L_∞-norm and proposed measure) the popular K-Means clustering method has been applied on the 6 interval datasets discussed above by using these measures. As we know that the K-Means is a numeric data clustering method, so in this work K-Means method has been adopted for interval data.

If there are *m* interval features then each sample and cluster centers can be represented as a hyper-rectangle in *m*-dimensional spaces. Suppose, a cluster *C* contains *n* samples with *m* interval features $\{([I_{11}^- I_{11}^+], [I_{12}^- I_{12}^+], \ldots\ldots\ldots, [I_{1m}^- I_{1m}^+]), \ldots\ldots\ldots, ([I_{n1}^- I_{n1}^+], [I_{n2}^- I_{n2}^+], \ldots\ldots\ldots, [I_{nm}^- I_{nm}^+])\}$ and $(\rho_1, \rho_2, \ldots\ldots, \rho_m)$ denotes the cluster representative for the cluster *C*.

Then, $\rho_i\ (1 \le i \le m)$ can be represented as an interval $[\rho_i^- \rho_i^+]$, where,

$$\rho_i^- = \sum_{j=1}^{n} I_{ji}^- / n \quad \text{and} \quad \rho_i^+ = \sum_{j=1}^{n} I_{ji}^+ / n$$

Again, the distance between a cluster center ρ and a sample I is calculated as-

$$D(\rho, I) = \sum_{i=1}^{m} d_{dist}(\rho_i, I_i).$$

Here, $d_{dist}(\rho_i, I_i)$ denotes the distance between i^{th} components of ρ and I and it can be measured using any distance measures available for interval data. In this work L_1-norm, L_2-norm, L_∞-norm and the proposed measure have been used to calculate $d_{dist}(\rho_i, I_i)$.

The performance of the data clustering is evaluated using clustering accuracy. If there are *k* number of clusters then the accuracy *R* of the clusters are measured as –

$$R = \frac{\sum_{i=1}^{k} a_i}{n}$$

Where *n* is the number of data points in the classified dataset and a_i is the total number of data points from the dominating class in i^{th} cluster. In a cluster, the dominating class is that class to which maximum data points in the cluster belong to.

C. Results and Discussion

In Table 1, Table 2, Table 3 and Table 4 the clustering accuracies obtained by the K-means method on the 6 interval datasets have been shown. On each dataset, the K-Means has been applied by setting the different parameters for *K* and distance measure function. In the distance measure function L_1-norm, L_2-norm, L_∞-norm and the proposed measure have been used. In all the tables '***Max***' denotes the maximum accuracy among all values of *K*.

Table 1: Clustering accuracy of K-Means method on Car Dataset

Distance Measures Used in K-Means	Accuracy for different values of *K*				*Max*
	K=3	*K*=5	*K*=7	*K*=9	
L_1-Norm	0.67	0.69	0.69	0.69	0.69
L_2-Norm	0.67	0.61	0.55	0.69	0.69
L_∞-Norm	0.30	0.30	0.30	0.30	0.3
Proposed	0.52	0.67	0.69	0.76	0.76

Table 2: Clustering accuracy of K-Means method on Fish Dataset

Distance Measures Used in K-Means	Accuracy for different values of *K*				*Max*
	K=3	*K*=4	*K*=5	*K*=6	
L_1-Norm	0.56	0.56	0.67	0.67	0.67
L_2-Norm	0.56	0.67	0.58	0.58	0.67
L_∞-Norm	0.56	0.56	0.33	0.56	0.56
Proposed	0.33	0.58	0.67	0.58	0.67

Table 3: Clustering accuracy of K-Means method on Water Dataset

Distance Measures Used in K-Means	Accuracy for different values of K				*Max*
	K=21	K=22	K=26	K=30	
L_1-Norm	0.37	0.37	0.38	0.41	0.41
L_2-Norm	0.34	0.33	0.35	0.35	0.35
L_∞-Norm	0.19	0.19	0.19	0.19	0.19
Proposed	0.38	0.37	0.39	0.41	0.41

Table 4: Accuracy of K-Means method on synthetic interval datasets SD1, SD2, SD3.

Datasets	Distance Measures Used in K-Means	Accuracy for different values of K				*Max*
		K=3	K=5	K=7	K=9	
SD1	L_1-Norm	0.67	0.85	0.85	0.87	0.87
	L_2-Norm	0.67	0.92	1	1	1
	L_∞-Norm	0.33	0.44	0.51	0.48	0.51
	Proposed	0.92	0.91	1	1	1
SD2	L_1-Norm	0.77	0.89	0.89	0.89	0.89
	L_2-Norm	0.64	0.88	0.88	0.88	0.88
	L_∞-Norm	0.44	0.51	0.33	0.33	0.51
	Proposed	0.61	0.88	0.89	0.89	0.89
SD3	L_1-Norm	0.8	0.87	0.86	0.85	0.87
	L_2-Norm	0.71	0.69	0.66	0.67	0.71
	L_∞-Norm	0.45	0.5	0.48	0.43	0.48
	Proposed	0.73	0.77	0.8	0.89	0.89

From Table 1, it is seen that on Car dataset, maximum accuracy obtained by the K-Means method is 0.76 and it is obtained by using proposed distance measure. From Table 2, it is seen that on Fish dataset, the maximum accuracy obtained by the K-Means method is 0.67. This accuracy is obtained by using the measures L_1-norm, L_2-norm and proposed measure. From Table 3, it can be concluded that using L_1-norm and proposed measure the K-Means method gives maximum accuracy 0.41 on the water dataset. On synthetic dataset SD1 (Table 4), the L_2-norm and proposed method give maximum accuracy 1. Again, using the proposed measure the K-Means method obtained maximum accuracy 0.89 on both SD2 and SD3 datasets (Table 4).

VI. CONCLUSION

In this paper, an analysis of ten desirable properties that should be fulfilled by the measures for interval data for making it suitable for applications like clustering and classification has been done. Also, it has been verified whether these properties are satisfied by the existing measures- L_1-norm, L_2-norm, L_∞-norm and it is seen that none of the measures satisfy all these properties. Thus, a new dissimilarity measure for interval data has also been proposed which satisfies all the desired properties except the triangular inequality. The performance of all the existing distance measures are compared with the proposed measure by applying well known K-Means algorithm on 6 interval datasets. It is seen that proposed measure gives better clustering accuracy on most of the datasets.

REFERENCES

1. Allen, J. F, "Maintaining knowledge about temporal intervals", *Communications of ACM*, Vol. 26(11), pp 832-843, 1983
2. Roh J. W., Yi B. K, "Efficient indexing of interval time sequences", *Information Processing Letters*, Vol. 109(1), pp 1-12, 2008
3. De Souza R. M. C. R., De Carvalho F. D. A. T., "Clustering of interval data based on city–block distances", *Pattern Recognition Letters*, Vol. 25, pp 353-365, 2004
4. Brito P., "Modelling and Analysing Interval Data", In: Decker R., Lenz H.J. (eds) Advances in Data Analysis, *Studies in Classification, Data Analysis, and Knowledge Organization*, Springer, Berlin, Heidelberg: pp.197-208, 2004
5. Peng W, Li T,"Interval Data Clustering with Applications", *In Proceedings of 19 IEEE International Conference on Tools with Artificial Intelligence*, pp.355-362, 2006
6. Irpino A., Verde R., "Dynamic clustering of interval data using a Wasserstein-based distance", *Pattern Recognition Letters*, Vol. 29 (11), pp. 1648-1658, 2008
7. De Carvalho F. A. T., Brito P., Bock H.-H., "Dynamic clustering for interval data based on L_2 distance", *Computational Statistics*, Vol 21(2), pp 231–250, 2006
8. Moore R. E,"Methods and Applications of Interval Analysis", *SIAM*, 1979
9. Goshtasby A.A., "Similarity and dissimilarity measures", *Image Registration*, Springer, London, pp. 7-66, 2012
10. Ren Y., Liu Y.-H., Rong J., Dew R., "Clustering interval-valued data using an overlapped interval divergence", *In Proc. 8th Australasian Data Mining Conference (AusDM'09)* Melbourne, Australia,. pp. 35–42, 2009

AUTHORS PROFILE

Irani Hazarika, PhD is an Assistant Professor (Contractual) in Department of Computer Science, Gauhati University. Her area of research is Data Mining.

Anjana Kakoti Mahanta, Ph.D. is a Professor in Department of Computer Science, Gauhati University. She has more than30 years of teaching experience. Her current area of research is Algorithms and Data Mining. In 2007, under a bilateral exchange program of Indian National Science Academy (INSA) and Polish Academy of Sciences, Poland, she visited the University of Warsaw for three months and had done research work in collaboration with faculty members in the department of Computer Science of the University.