IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 22, Issue 6, Ser. I (Nov. – Dec. 2020), PP 45-50
www.iosrjournals.org

Interval Data Clustering

Dr. Jyoti Prokash Goswami[1], Dr. Anjana Kakoti Mahanta[2]
[1](Associate Professor, Department, of Computer Applications, Assam Engineering College, India)
[2](Professor, Department of Computer Science, Gauhati University, India)

Abstract: *In this paper, a new framework for clustering interval data has been proposed. In this frame work, each cluster is represented by a representative which is a fuzzy set. First, we define the interval data and its cluster representative as the uniform fuzzy set representation. Next, we consider a data set with single attribute of interval data type for the sake of defining similarity and distance functions for interval data and then propose a way of handling this attribute in the process of clustering. It is assumed that the underlying domain of the attribute is a discrete, totally ordered bounded domain. Also, all intervals are assumed to be closed intervals. We suggest measures to calculate the distance/similarity between two interval data, an interval data and a cluster representative and two cluster representatives of interval data. Using these concepts, interval data clustering may also be implemented with any existing representative based or summary based clustering algorithm.*

Keywords: *cluster representative, interval data, membership value, spread of a cluster*

Date of Submission: 10-12-2020 Date of Acceptance: 25-12-2020

I. Introduction

There are many phenomena we have been facing with in reality that cannot be explained with the help of single-valued variables. Consequently, they have to be kept in a dataset with interval attributes. This type of data is a different one in the sense that, instead of representing a single value it may takes a range of continuous values for the variables. They are referred to as interval data also known as symbolic interval data [1]. Study of the interval valued data or simply interval data are most common and used in different fields of study. Some of the examples of this type are: daily weather temperature, weekly price variations of fish, record of blood pressure of a patient, etc. thus, a given data d is defined as a closed and bounded interval [1] and are often applied to represent quantity which may differ between an upper and a lower bounds. It comprises a range of continuous values for the variables. Most of the data in statistical analysis lie within some specified intervals. Interval data clustering is comparatively a new field of study in the field of data mining, particularly in case of clustering, and very limited work have been done so far. The most important point is how to define a suitable distance or similarity measure that may be efficiently implemented to determine the distance or similarity between two interval data. Two well known measures are Hausdorff and City Block distance measures which are calculated on the basis of only the upper and the lower bounds of the intervals. In another measure viz., OLID, a relationship has been established between the upper and the lower bounds and the overlapped areas of the two intervals and hence the distance between two interval data is defined. For any two intervals $I_1 = [x_1, y_1]$ and $I_2 = [x_2, y_2]$, the City Block and Hausdorff distance measures are defined as follows on the basis of their boundary values.

City Block distance: $(L_1) = I\, x_1 - x_2 I + I y_1 - y_2 I$

Hausdorff distance: $(d_H) = \max (I\, x_1 - x_2 I, I y_1 - y_2 I)$

In this paper, we have extended the idea of clustering categorical data of the paper [2] for interval data. The uniform fuzzy set representation of a categorical data point and a cluster representative, as proposed in [2], is also applied for interval data. On the basis of this, next we are trying to define similarity and distance functions for interval data. For this purpose we consider a categorical data set whose only attribute is of interval type or a data set with single attribute that takes values of interval data type and then suggest a way of handling this attribute in the clustering process. However, the same idea may be extended to a data set with more attributes of interval type. We assume the underlying domain of the attribute to be a discrete totally ordered bounded domain. Also we assume that all intervals are closed intervals, i.e., if $x \in [a, b]$ then $a \leq x \leq b$. Size of an interval [a, b] is defined as the cardinality of the set consisting of all elements in the interval [a, b]. For example, if the underlying domain is the set of integers, then **I [2, 8] I = 7.**

Here, we propose measures to determine the distance/similarity between two interval data , an interval data and a cluster representative and two cluster representatives of interval data. In this paper, some important concepts on interval data have been defined and a framework has been provided so that the genetic algorithm

based categorical data clustering algorithm proposed in [3] can be implemented for interval data. The remainder of this paper is organized as follows.

In the next section, a brief review on the interval data clustering has been made. In section 3 and its sub-section we visualize an interval as a set of discrete unit intervals. 'Spread' of a cluster, the concept of 'membership value', and set superimposition, three terms that are being used by us are defined next. Then in section-4, we describe the way of constructing a cluster representative. Next, we propose methods for measuring distances between two intervals and between a cluster representative and an interval considering both the aspects of overlapping and disjoint case separately. We also define a measure to find distance between two cluster representatives in section-7. A common generalized formula for measuring distance is proposed in section- 8 and finally the chapter is concluded in section-9.

II. Review of related works

The traditional data analysis methods of interval data have taken into consideration only the representative of the intervals viz. center or mean ignoring the detail structure of intervals. A modified method of dissimilarity measure of interval data for the purpose of handling the problem of clustering is proposed in [4] where various approaches of distance / dissimilarity measure between interval data are presented, showing relations among them. Also a comprehensive experimental study based on the proposed measure in clustering the interval data is made on various data sets and shows how their approach results better performance over the traditional ones in terms of producing more meaningful and explanatory result. A two stage-approach of clustering interval data is proposed there observing the fact of existence of natural two-level hierarchical representation viz., the representative of interval data and the fine representation that also shows its structure information. In the first stage, they consider the representative level / coarse level representation of interval data for obtaining a rough partition of the data and in the next stage, final clustering is made by refining the results of the first stage using the fine representation. Experimentally it has been proved that this two stage approach maintains clustering quality and reduces the computation cost. A novel fuzzy clustering method for interval valued data is proposed in [5] with an adaptable variable selection and found to be useful for three reasons. First, in the classification of interval data with high dimension and low sample-size, results poor quality because of the presence of noise occurring from irrelevant and redundant variables (dimensions), an adaptable variable selection is to be done to reduce or summarize variables. Second, fuzzy clustering results clusters with uncertain boundaries which is well adjusted with the uncertainty situations of classification to data and thereby give more robust result when compared to hard clustering, for the noise of data. Third, by using this adaptable representation of interval data, the original data can be transformed into a more manageable data for the purpose of avoiding curse of dimensionality. A novel proximity measure for intervals called Overlapped Interval Divergence (OLID) is proposed in [6] which considers the relationship between intervals and their overlapped areas and thus extends the existing distance measures where only the upper and the lower bounds of intervals take part. Genetic algorithm has been used in [7] to propose an automatic interval data clustering algorithm where the number of clusters is suitably determined applying the overlapped distance between intervals. The Davies and Bouldin index are modified there for the sake of optimizing clustering results and improving the operators of the original genetic algorithm viz., crossover, mutation and selection. The algorithm has been tested on data sets with different characteristics and proved to be more advantageous in comparison to the existing ones.

III. Important Concepts

The concepts of cluster representative and similarity measure that were used for categorical data in [2] and [3] have been extended to the case of interval data. In addition to them, few more concepts relevant to this paper are described in this section.

3.1 Visualization of intervals as set.

Although it is known that every interval is a set, we reemphasis this here since our aim is to extend the definition of a cluster representative that was for categorical data in [2] and [3] to interval data. We consider an interval as a set of discrete unit intervals. This is possible since we have assumed that our underlying domain is discrete. For example, if the domain is I, the set of integers, then the interval [4, 7] is considered as set {[4],[5],[6],[7]}. This is done in order to define cluster representatives using the concept that we have used for categorical data.

3.2 Spread of a Cluster.

If a cluster is formed with the intervals $I_1 = [a_1, b_1]$, $I_2 = [a_2, b_2]$,……, $I_n = [a_n, b_n]$ then we define the spread of the cluster as another interval [a, b], where $a = \min \{ a_1, a_2, \ldots\ldots, a_n\}$ and $b = \max \{ b_1, b_2, \ldots\ldots b_n\}$. For a cluster consisting of a single interval data spread is equal to the interval itself.

3.3 Membership Value.

The membership value, in case of an interval data, is defined for each discrete unit interval as well as for its spread as its contribution in forming the cluster. So we may consider a cluster either having only a single interval or more than one interval data for the purpose of defining membership value. Membership value calculation is important from the cluster representative construction point of view.

In case of cluster representative formed with more than one interval either of overlapped or disjoint or of both types, the membership value of each discrete unit interval within the spread of the representatives may be any number m, ($1 \leq m \leq 0$). A '1' as the membership value for a discrete unit interval say a_i implies that a_i is present in all the intervals which form the cluster.

Example 1 : (Membership Value & Spread)

Let, **f** and **g** are the two interval data which form a cluster.

Now,

Case1. They are Disjoint Say, f = [2, 5] and g = [8, 12]

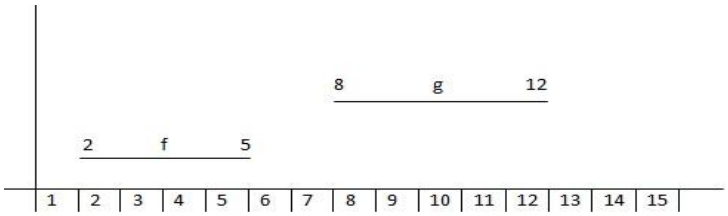

Figure no 1 : Disjoint data

Membership value of the interval [2, 5] = 1/2; [6, 7] is 0; and that of [8, 12] = 1/2;
Spread of the cluster is [2, 12].

Case2. They are Overlapped: Say, f= [2, 7] and g = [5, 12]

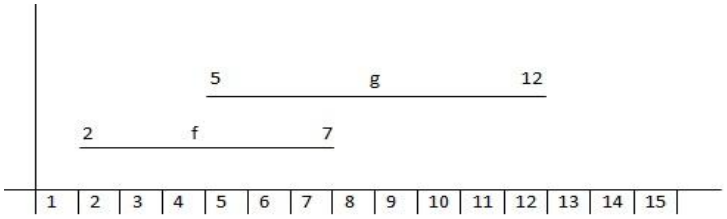

Figure no 2 : Overlapped data

Membership value of the interval [2, 4] = 1/2; [5, 7] is 1; and that of [8, 12] = 1/2;
Spread of the cluster is [2, 12].

3.4 Set superimposition.

The concept of set superimposition proposed in [9] is also used here for the purpose of constructing a cluster representative. We can define the term with an example.

Let us consider the same interval data or sets as shown in Fig. 2. So f ∩ g i.e. [2,7] ∩ [5,12] = [5,7] $\neq \phi$.

Now, using the symbol "S" for superimposition as in [9], f (S) g will be consisting of the superimposed intervals $[2,4]^{1/2}$, $[5,7]^{1}$ and $[8\text{-}12]^{1/2}$ Fig. 3, where the membership of [5,7] is 1 due to double representation.

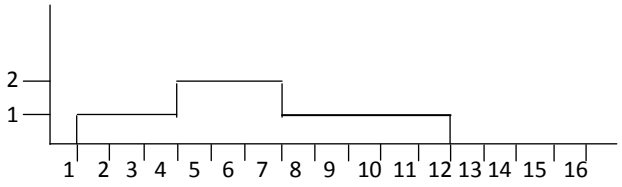

Figure no 3: Set superimposition

IV. Construction of a Cluster Representative

The cluster representative has been formed with the idea as was done in [2] and [3] in categorical data clustering. At any point of time, a cluster may have data intervals which are overlapped and / or disjoint in nature and the cluster is represented by its representative. So, a cluster representative is formed with the

membership values of the discrete / consecutive intervals obtained by superimposing the intervals [8],[9] present in the clusters.

For example, considering a cluster C_1 consisting of the following seven interval data-
(1). 4 — 8, (2). 6 — 10, (3). 5 — 9, (4). 12 — 15, (5). 10 — 14, (6). 16 — 19 and (7). 22 — 25
the cluster representative for C_1 as shown in the fig. 4, is defined as:
CR_1={ [4,4]: $\frac{1}{7}$; [5,5]: $\frac{2}{7}$; [6,8] : $\frac{3}{7}$; [9,10]: $\frac{2}{7}$; [11]: 0; [12, 14] : $\frac{2}{7}$; [15,19]: $\frac{1}{7}$; [20,21] : 0; [22,25]: $\frac{1}{7}$; }

Figure no 4: Cluster Representative for Cluster C_1 obtained by superimposing the interval data.

Here, each fraction indicates the membership values of the corresponding subintervals. The union of the subintervals is the spread of the cluster. A cluster representative is actually a partition of the spread of the cluster where membership values are associated with the individual members of the partition. It is also true that the membership value of a partition lies between 0 and 1, both inclusive.

V. Distance between two Intervals

Here, we are trying to define the distance between two interval data considering both the disjoint and overlapped cases separately with examples.
Let f = [f_l , f_u] and g = [g_l , g_u] are two interval data, where f_l and g_l are the lower ends and f_u , g_u are the upper ends . In defining the distance between two such interval data, we consider an interval as a set of discrete unit intervals as already stated in section 3.1.

5.1. Disjoint Intervals (NOL)
If the two intervals **f** and **g** are disjoint, then distance between them is measured as the modulus of difference between their mean, i.e.
$d_{NOL}(f, g) = |\text{ mean}(f) - \text{mean}(g)|$, where, mean of an interval data x = [x_l , x_u] is calculated as : $(x_l + x_u)/2$
Alternatively, it can also be calculated as: $[x_l + x_{l+1} + \ldots\ldots + x_u]/ [[x_u - x_l]+1]$

Example 2: Distance between two disjoint intervals.
Let, f = [2, 6] and g = [8, 12], as shown in the figure 4

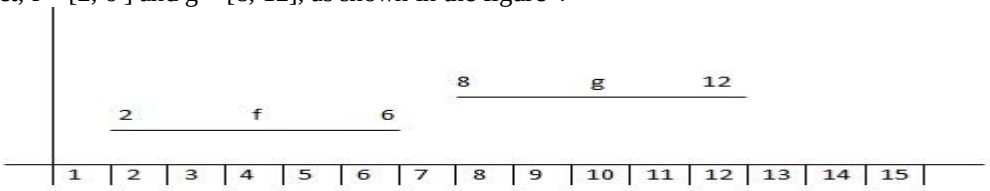

Figure no 5 : Two disjoint Intervals

Now, $d(f, g) = |[(2+6)/2 - (8+12)/2]| = 6$
Alternatively, $|[(2+3+\ldots..+6)] / [(6 - 2+1] - [(8+9+\ldots..+12)| / [(12-8)+1]$
$= |4 - 10|; = 6$

5.2. Overlapping intervals (OL)
In case of overlapping intervals we define the similarity between them using Jaccard coefficient defined as:

$$\text{Sim}(f, g) = \frac{|\,intersection(f,g)\,|}{|\,union(f,g)\,|}$$

Here, intersection and union of f and g are defined by considering them as sets of discrete intervals as described in 3.1.
For example, in case of the interval f= [f_1, f_n], the set of discrete intervals is [f_1, f_2, …… f_{n-1},f_n], if $f_1 \leq f_2 \leq \ldots\ldots \leq f_n$.
Then distance is defined as: $d_{OL}(f, g) = 1 - \text{sim}(f, g)$; Here, $0 \leq d \leq 1$.

If the intervals are same, then their similarity value is 1, i.e., $d_{OL}(f, g) = 0$

Example 3: Distance between two overlapping intervals.

Considering the integer domain, let f = [2 , 8] and g= [4 , 12], as shown in the fig. 6.

Now, sim (f, g) = $\frac{|(4,5,6,7,8|}{|(2,3,4,5,6,7,8,9,10,11,12)|}$;

$= \frac{5}{11} = 0.455$ and d(f, g) = 1 – 0.455 ; = 0.545

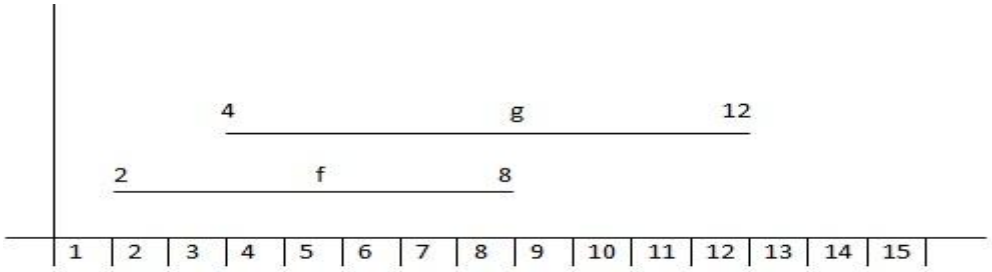

Figure no. 6 : Two Overlapping Intervals

VI. Distance between an interval data and a Cluster representative

To define the distance between an interval data and a cluster representative we first calculate the weighted mean of the cluster representative. As we have seen in the previous section that the data belong to a cluster which form the representative are distributed not in a uniform manner throughout the spread of the representative. In other words, membership values of each discrete interval may be different. For this reason, it would not be appropriate to define the distance between an interval data and a cluster representative simply considering the mean of the representative. Weighted mean of the cluster representative is calculated by adding the product of the numbers of data (weights) belong to each unit/discrete interval (or that share/overlap each discrete interval) and their corresponding interval unit number (ordered) and then dividing it by the sum of the number of data of the each discrete/contiguous interval within the spread. Thus, the weighted mean (W.mean) for the cluster representative of CR_1 would be:

W.mean = (4*1+5*2+(6+7+8)*3+(9+10)*2+11*1+(12+13+14)*2+(15+16+17
+18+19)*1+ (22+23+24+25)*1 / (1+2+9+4+1+6+5+4)
= 11.97

Now we are in a position to define the distance measures for the following cases that are very much essential for the purpose of implementing our proposed clustering algorithm.

Case1. The interval data is disjoint from the spread of the representative. (NOL)

The distance between a cluster representative 'CR_1' and a data 'f', whose interval/spread is different from the spread of the representative i.e. (spread of the $CR_1 \cap$ spread of f) = ϕ, is calculated by obtaining the difference between the mean of the data and the weighted mean of the representative.

Thus, $d_{NOL}(f,CR_1) = | \text{mean}(f) – \text{W.mean}(CR_1) |$

Case2 . The interval data is overlapping with the spread of the representative. (OL)

In case of determining distance between an interval data 'f' and a Cluster representative CR_1, where (spread of the $CR_1 \cap$ spread of f) ≠ϕ, we first calculate the similarity between them as has been done in case of two overlapping data, as follows:

$$\text{sim}(f, CR_1) = \frac{|(f \cap CR1)|}{|(f \cup CR1)|};$$

Where, union, intersection and cardinality of intervals are straight forward extension of the corresponding definitions in fuzzy set theory. It is possible to bring the fuzzy set operations into picture because the partitions of the spread of a cluster representative have membership values between 0 and 1, both inclusive as stated in section 3.3. In case of an interval this membership value is 1.

Then distance is defined as: $d_{OL}(f, CR_1) = 1- \text{sim}(f, CR_1)$;

VII. Distance between two Cluster representatives

In this section, we define a measure to find distance between two cluster representatives taking into consideration of their both overlapped and disjoint spreads.

Let CR_1 and CR_2 are the two cluster representatives, and then distance between them is defined again separately depending on whether they are overlapping or disjoint.

Case 1. Spread of the two representatives are disjoint from each other. (NOL)

In this case, distance between them is calculated as the modulus of difference between their weighted mean

as given below:

$$d_{NOL}(CR_1, CR_2) = | W.mean(CR_1) - W.mean(CR_2) |$$

Case 2. Spread of the two representatives are Overlapping with each other. (OL)

When the spreads of the two cluster representatives are overlapping, we first calculate their similarities using the Jaccard coefficient as was used for two overlapped intervals. To obtain the distance we then subtract similarity value from 1 . The Jaccard coefficient in this case is defined as:

$$sim(CR_1, CR_2) = \frac{|(CR1 \cap CR2)|}{|(CR1 \cup CR2)|};$$

Thus distance is defined as: $d_{OL}(CR_1, CR_2) = 1 - sim(CR_1, CR_2)$;

VIII. Common formula (Generalized) for Measuring Distance

After defining the distance measures separately for each case, now we are in a position to propose a Common Generalized formula to measure distance between two interval data or a data and a representative, considering both the cases of disjoint and overlapped interval.

Let $I_1 = [a, b]$ and $I_2 = [c, d]$ are two interval data, now we define a factor NOVERLAP (I_1, I_2), as follows.

NOVERLAP $(I_1, I_2) = 0$; if either $a \leq c \leq b \leq d$ or $c \leq a \leq d \leq b$, i.e., the intervals are not disjoint.

$= 1$; otherwise.

When distance between a cluster representative and an interval data is to be calculated then the spread of the cluster is taken, which is an interval.

Now, $d(I_1, I_2) = d_{OL}(I_1, I_2) + NOVERLAP(I_1, I_2) * d_{NOL}(I_1, I_2)$

It is to be noted that $d_{OL}(I_1, I_2) = 0$, if I_1 and I_2 are disjoint.

IX. Conclusion

In this paper, we have mentioned the importance of interval data in different field of study. Interval data clustering is relatively a new field of study. Distance measures commonly used are based only on the upper and lower bounds of the intervals except the measures in [6],[7]. We have considered here the importance of overlapping portion along with the two bounds in defining a common distance measure applicable for both overlapped and disjoint data. In doing so, we have first defined some important concepts relevant to our study and on the basis of these we propose methods for measuring distances between two intervals, between a cluster representative and an interval, and between two cluster representatives considering both the aspects of overlapping and disjoint case separately. Thus, the same ideas applied to categorical data clustering in [2], have applied in creating a frame work for implementing our GA based clustering algorithm proposed in [3] for interval data.

References

[1]. G. Cabanes, Y. Bennani, and R. Destenay, A Hardy. A new topological clustering algorithm for interval data, Pattern Recognition, 46 (2013) ELSEVIER. pp. 3030-3039.

[2]. J. P .Goswami, and A. K. Mahanta . Categorical Data Clustering Based on an Alternative Data Representation Technique, International Journal of Computer Applications, (0975-8887), Vol.-72, No. 5, pp. 7-12, May 2013.

[3]. J. P. Goswami and A. K. Mahanta . A Genetic Algorithm Based Ensemble Approach for categorical Data Clustering: 12th IEEE India International Conference INDICON 2015.

[4]. Wei Peng and Tao Li. Interval Data Clustering with Applications. ICTAI 2006, 18th IEEE International Conference pp. 355-362.

[5]. Mika Sato – Ilic. Symbolic Clustering with Interval Valued Data. Procedia Computer Science 6 (2011) pp, 358-363

[6]. Y. Ren, Y. Liu, J. Rong, R. Dew, Clustering Interval -Valued Data Using an Overlapped Interval Divergence. Proceeding of the 8th Australasian Data Mining Conference (AusDM'09), pp. 35-42, 2009.

[7]. T. V. Van, D. Phamtoan, L.H. Tuan, T.N.Trang, An Automatic Clustering for Interval Data using the Genetic Algorithm. Springer, April 2020 Annals of Operation Research.

[8]. H. K. Baruah. Set Superimposition and its application to the Theory of Fuzzy Sets, Journal of Assam Science Society, Vol. 10 No. 1 and 2, (1999), pp. 25-31.

[9]. A .K. Mahanta, F. A. Mazarbhuiya, H. K. Baruah. Finding Calendar- based Periodic Patterns. Pattern Recognition Letters 29 (2008) 1274-1284, ELSEVIER.

Dr. Jyoti Prokash Goswami. "Interval Data Clustering." *IOSR Journal of Computer Engineering (IOSR-JCE)*, 22(6), 2020, pp. 45-50.