Clustering Interval-valued Data Using an Overlapped Interval Divergence

Yongli Ren[1] **Yu-Hsn Liu**[2] **Jia Rong**[2] **Robert Dew**[2]

[1] School of Information Engineering, Zhengzhou University,
Zhengzhou 450052, China
Email: yonglitom@gmail.com

[2] School of Information Technology, Deakin University,
221 Burwood Highway, Vic 3125, Australia
Email: {yuhsnliu, jrong, rad}@deakin.edu.au

Abstract

As a common problem in data clustering applications, how to identify a suitable proximity measure between data instances is still an open problem. Especially when interval-valued data is becoming more and more popular, it is expected to have a suitable distance for intervals. Existing distance measures only consider the lower and upper bounds of intervals, but overlook the overlapped area between intervals. In this paper, we introduce a novel proximity measure for intervals, called *Overlapped Interval Divergence (OLID)*, which extends the existing distances by considering the relationship between intervals and their overlapped "area". Furthermore, the proposed *OLID* measure is also incorporated into different adaptive clustering frameworks. The experiment results show that the proposed *OLID* is more suitable for interval data than the Hausdorff distance and the city-block distance.

Keywords: Clustering, Distance, Similarity , Interval Valued Data.

1 Introduction

The importance of distance measures in machine learning and data mining is clear: a large number of learning problems, such as clustering and lazy learning, heavily rely on the similarity measurement over the data instance space. Accordingly, one of the main issues in these problems is the selection of a suitable metric for the concerned application domain. Most of existing distances have been designed for a relatively simple way: the data instance is described by a vector of random variables, each of which results in just one single value. However, in real life there are many situations where the use of interval-valued data is more suitable.

In general, interval-valued data come from two major sources:

1. many phenomena cannot be explained by using single-valued variables, and from their outset some data sets will include interval attributes. Many of natural language are expressed with intervals instead of single crisp values, e.g. "*I drink* 4-6 *cups of water a day.*" Similarly, in medical and engineering data, intervals also appear frequently, because of some tolerance in measuring real parameters. For example, age could be recorded as being in an interval, such as $[0, 10]$, $[30, 40]$ etc. In addition, it may not be possible to measure some characteristics accurately by a single value, e.g. the pulse rate at 70, but rather measures the variable as an $(x \pm \delta)$ value, namely (70 ± 1). The blood pressure may be recorded by its $[low, high]$ values, e.g. $[138, 180]$. These are all interval-valued attributes. A typical data set with interval-valued attributes may follow the lines of Table 1.

Table 1: Sample Interval Data Set

No.	Age	Weight	$\cdots$	Blood Pressure
1	$[12, 17]$	$[45, 50]$	$\cdots$	[90, 100]
2	$[25, 30]$	$[70, 80]$	$\cdots$	[138, 180]
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
21	$[20, 30]$	$[65, 70]$	$\cdots$	[110, 150]
22	$[10, 20]$	$[45, 70]$	$\cdots$	[90, 170]
23	$[30, 40]$	$[70, 75]$	$\cdots$	[70, 120]

2. As data sets increasingly suffer from the problem of scale, in terms of either the number of attributes or the number of instances. Researchers and practitioners from more diverse disciplines than ever before are attempting to use automated methods to analyze their data. It is often desirable to reduce the size of the data while maintaining their essential information as much as possible. One approach is to summarize large data sets in such a way that the resulting data set is of a manageable size. In this situation, interval data store variability better than standard single value data when real values describing the individual observations result in intervals in the description of the summarized data. Accordingly the summarized data could no longer be single values as in classical format, but instead be represented as intervals (Billard 2006).

The statistical treatment of interval-valued attributes has been considered in the context of *Symbolic Data Analysis* (SDA) (Diday 1988), which is a domain related to exploratory data analysis, multivariate analysis and pattern recognition. SDA aims to provide suitable methods for analyzing data set described through multi-valued attributes, including intervals, sets categories, or weight distributions. SDA has provided suitable tools for clustering interval-valued data: in 2004, Souza et al. proposed a clustering algorithm for interval data based on the *city-block* distance (de Souza & de A.T. de Carvalho 2004), and they applied the dynamic adaptive clus-

tering framework which incorporate the *city-block* distance to measure the distance between intervals. In 2006, De Carvalho et al. adopted a similar dynamic clustering framework but with the *Hausdorff* distance instead for intervals (de A.T. de Carvalho et al. 2006). Recently De Carvalho et al. further propose the single adaptive clustering framework, in which both the *city-block* and *Hausdorff* distances can be adopted (de A.T. De Carvalho & Lechevallier 2009). The single adaptive distances in (de A.T. De Carvalho & Lechevallier 2009) use the same adaptive parameters for all clusters, while it is different to the cluster adaptive distances in the early work (de A.T. de Carvalho et al. 2006), which use different adaptive parameters from cluster to cluster.

Most distances used for clustering interval data presented thus far have been designed for a relatively simple way: given two intervals, only the crisp values of their lower and upper bounds were considered, and the information about their overlapped area has been largely overlooked. However, in real life there are many situations where the ignorance of these overlapped area causes severe loss information, especially when both the distance between interval centers and the relative size of the overlapped area are concerned.

In this paper, we aim to fill the void by proposing a new distance for interval-valued data. By considering the intervals as a *hypercube* in a high dimensional space and take the overlapped area into consideration, the proposed *Overlapped Interval Divergence* is different from other interval distances which only consider their lower and upper bounds as single high dimensional points. As we will see from the later sections that by incorporating the proposed distances into both the single and the cluster adaptive clustering frameworks, we can get more accurate clustering results than existing distances.

The rest of the paper is organized as follows. In section 2, the related work are presented. In section 3, we propose the *Overlapped Interval Divergence* with detailed analysis of its properties and the adaptive clustering algorithms employed in the work. In section 4, we present the experiment results that evaluate the proposed algorithms compared to single(cluster) adaptive *Hausdorff* distance and single(cluster) adaptive *city-block* distance under the synthetic data sets. Finally conclusions and future work are presented in section 5.

2 Dynamic Clustering for Interval-valued Data

Clustering, partitioning data into sensible groupings according to measured or perceived intrinsic characteristics or similarity, is one of the most fundamental unsupervised data mining tasks. It is useful for helping user to understand and interpret the general patterns in data when prior knowledge of the underlying distribution is missing. As the representation of data by means of intervals is becoming more and more frequent, researchers and practitioners from more diverse disciplines than ever before are attempting to extend existing methods for the comparison of interval data (Diday 1988).

2.1 Interval-Valued Data

According to symbolic data analysis (Diday 1988), an interval variable is a variable which takes the interval values such as $[a, b]$, where $a \leq b$ and $a, b \in \Re$. When $a = b$, this interval variable is becoming a normal single valued variable. Let D be a data set described by p interval variables. Each data instance $x_i \in D$ is represented as a vector of intervals: $x_i = (x_i^1, \cdots, x_i^p)$, where $x_i^j = [a_i^j, b_i^j]$.

A distance or proximity measure d is a non-negative function defined on each pair of interval-valued data instances, such that the closer the instances, the lower the value assumed by d. Two popular distance measures which have been widely used are the *city-block* distance (de Souza & de A.T. de Carvalho 2004) and the *Hausdorff* distance (de A.T. de Carvalho et al. 2006), which will be described later together with the dynamic clustering algorithms.

2.2 Adaptive Distances for Dynamic Clustering

Symbolic data analysis has provided clustering methods in which interval-valued data are considered. As the most influential symbolic data analysis method, the Dynamic Clustering Algorithm represents a group of unsupervised partition-based clustering algorithms. It can be proven that this group of algorithms generalizes several clustering algorithms including K-means and K-median algorithm.

The general Dynamic Clustering Algorithm looks for the partition of data set D into K clusters, and each cluster is represented by a single prototype vector of intervals, such that the sum of distance measures between each instance belonging to a cluster and the cluster's prototype is minimized.

Let $y_k = ([\alpha_k^1, \beta_k^1], \ldots, [\alpha_k^p, \beta_k^p])$ be the prototype for the k-th cluster P_k $(k = 1, \ldots, K)$. The Dynamic Clustering Algorithm is then trying to minimize the following criterion:

$$O = \sum_{k=1}^{K} \sum_{x_i \in P_k} d_k(x_i, y_k). \quad (1)$$

Popular distance measures for interval-valued data include the *Hausdorff* distance (de A.T. de Carvalho et al. 2006) and the *city-block* distance (de Souza & de A.T. de Carvalho 2004). When they are used with Dynamic Clustering Algorithms, they usually appear in one of two adaptive forms: the single adaptive distance uses the same parameter for all clusters; the cluster adaptive distance uses different parameters from cluster to cluster (de A.T. De Carvalho & Lechevallier 2009).

2.2.1 The Single Adaptive Distances

Literature (de A.T. De Carvalho & Lechevallier 2009) proposes the partitional clustering algorithm for interval-valued data by using a single adaptive *Hausdorff* distance:

$$d_k(x_i, y_k) = \sum_{j=1}^{p} \lambda^j (max[|a_i^j - \alpha_k^j|, |b_i^j - \beta_k^j|]), \quad (2)$$

in which $\lambda^j = (\lambda^1, \ldots, \lambda^p)$ is a weight vector for p interval variables.

Similarly, if using the *city-block* distance we can get:

$$d_k(x_i, y_k) = \sum_{j=1}^{p} \lambda^j [|a_i^j - \alpha_k^j| + |b_i^j - \beta_k^j|], \quad (3)$$

where the weight vector $\lambda^j = (\lambda^1, \ldots, \lambda^p)$ is also fixed for p interval variables.

De Carvalho et al. also propose an extended single adaptive *city-block* distance for interval-valued data clustering (de A.T. De Carvalho & Lechevallier 2009):

$$d_k(x_i, y_k) = \sum_{j=1}^{p} (\lambda_L^j |a_i^j - \alpha_k^j| + \lambda_U^j |b_i^j - \beta_k^j|), \quad (4)$$

in which there are two vectors of weight, one for the lower boundary$\lambda_L = (\lambda_L^1, \ldots, \lambda_L^p)$, and the other for the upper boundary $\lambda_U = (\lambda_U^1, \ldots, \lambda_U^p)$. These weight vectors are the same for each cluster.

2.2.2 The Cluster-Adaptive Distances

De Carvalho et al. introduce the dynamic clustering algorithm for interval data by adopting different adaptive *Hausdorff* distances for different clusters (de A.T. de Carvalho et al. 2006):

$$d_k(x_i, y_k) = \sum_{j=1}^{p} \lambda_k^j (max[|a_i^j - \alpha_k^j|, |b_i^j - \beta_k^j|]), \quad (5)$$

which is parameterized by K vectors $\lambda_k = (\lambda_k^1, \ldots, \lambda_k^p)$ $(k = 1, \ldots, K)$, one for each cluster.

Similarly, we can have the cluster adaptive *city-block* distance (de Souza & de A.T. de Carvalho 2004):

$$d_k(x_i, y_k) = \sum_{j=1}^{p} \lambda_k^j [|a_i^j - \alpha_k^j| + |b_i^j - \beta_k^j|], \quad (6)$$

which is also parameterized by K vectors $\lambda_k = (\lambda_k^1, \ldots, \lambda_k^p)$ $(k = 1, \ldots, K)$.

Souza and De Carvalho also extended the cluster adaptive *city-block* distance by separately considering the lower and the upper bounds (de Souza & de A.T. de Carvalho 2004):

$$d_k(x_i, y_k) = \sum_{j=1}^{p} (\lambda_{kL}^j |a_i^j - \alpha_k^j| + \lambda_{kU}^j |b_i^j - \beta_k^j|), \quad (7)$$

where each cluster P_k is parameterized by two weight vectors: one for lower boundary $\lambda_{kL} = (\lambda_{kL}^1, \ldots, \lambda_{kL}^p)$, the other for upper boundary $\lambda_{kU} = (\lambda_{kU}^1, \ldots, \lambda_{kU}^p)$. Here, the weight vectors are also different from cluster to cluster.

2.2.3 The General Adaptive Clustering Algorithm

Once the strategy for adaptive distances is determined, the algorithm for clustering interval-valued data can be generated into a general process: it will randomly choose a partition of X into clusters $P = (P_1, \ldots, P_k)$, then iterate over the following steps.

- In the first step, determine K cluster prototypes $y = (y_1, \ldots, y_K)$ to represent each cluster.
- In the second step, fix the prototypes $y = (y_1, \ldots, y_K)$ and the partitions $P = (P_1, \ldots, P_k)$, and update the adaptive distances d_k so that the adequacy criterion O is minimized.
- In the third step, fix the prototypes and the adaptive distances, and determine the best partition $P = (P_1, \ldots, P_k)$ which minimizes the adequacy criterion O.

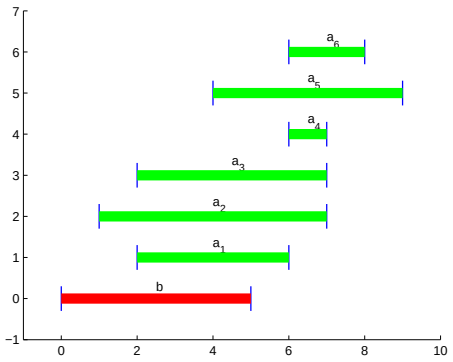


Figure 1: The interval $a_1, a_2, a_3, a_4, a_5, a_6$ and b.

3 Clustering with the Overlapped Interval Divergence

Most distances defined for interval-valued data have been designed for a relatively simple way: only the lower and upper bounds of the intervals are considered. However, in real life there are many situations where their overlapped area should also be considered (Li & Tong 2002, Li & Dai 2004, Jiang et al. 2005, Dai et al. 2004).

For example, considering seven intervals as shown in Fig. 1, the *city-block* and the *Hausdorff* distances from any interval $\{a_1, a_2, a_3, a_4, a_5, a_6\}$ to b are presented in Table 2. As we can see, intervals a_1, a_2, a_3 and a_5 are all overlapping with interval b, while intervals a_4 and a_6 have no overlapped area with b. According to the *city-block* distance, we have $d_c(a_1, b) = d_c(a_2, b) = 3$ and $d_c(a_4, b) = d_c(a_5, b) = 8$. It is evident that the *city-block* distance can not distinguish a_1 from a_2, or distinguish a_4 from a_5. Similarly, if following the *Hausdorff* distance, we will have $d_H(a_1, b) = d_H(a_2, b) = d_H(a_3, b) = 2$ and $d_H(a_4, b) = d_H(a_6, b) = 6$, which means the *Hausdorff* can not distinguish among a_1, a_2 and a_3, or between a_4 and a_6.

This is contradict to the intuition that a_1, a_2, a_3 are different from each other, especially when the relative size of the overlapped area is a concern. In this section, we are addressing this problem by proposing a new proximity measure for intervals.

3.1 The Overlapped Interval Divergence (OLID)

Any interval-valued data generalizes a single-valued data because it represents a range of values, and have "area" in nature. In addition, there will be an overlapped area between any two intervals, even though the overlapped area might be empty. For intervals, two factors are related to the proximity between two intervals: one is the distance between their centers; another one is the relative size of their overlapped area. By considering the above two factors together, we propose an *Overlapped Interval Divergence (OLID)* for intervals.

Definition 1 *Given two intervals* $a = [a_1, a_2]$ *and* $b = [b_1, b_2]$, *let* $c_a = \frac{a_1+a_2}{2}, r_a = \frac{a_2-a_1}{2}$ *and* $c_b = \frac{b_1+b_2}{2}, r_b = \frac{b_2-b_1}{2}$. *Then the* Overlapped Interval Divergence (OLID) *from interval* a *to* b *is defined as:*

$$div(a, b) = l(a, b) \cdot (1 - \frac{OA(a, b)}{2r_a + 1}). \quad (8)$$

Table 2: Distances to $b = [0, 5]$

Distances	$a_1 = [2, 6]$	$a_2 = [1, 7]$	$a_3 = [2, 7]$	$a_4 = [6, 7]$	$a_5 = [4, 9]$	$a_6 = [6, 8]$
city-block	3	3	4	8	8	9
Hausdorff	2	2	2	6	4	6
OLID	0.4	0.8571	1	3	3.3333	4

where $OA(a, b)$ is the Overlapped Area between a and b, and $l(a, b)$ is a distance originated from Hausdorff *distance by considering all points inside the intervals:*

$$l(a,b) = max_{a' \in [a_1,a_2]}\{min_{b' \in [b_1,b_2]}\{u(a^{'}, b^{'})\}\}, \quad (9)$$

in which $u(a^{'}, b^{'})$ is the Euclidean distance between $a^{'}$ and $b^{'}$.

For an interval $a = [a_1, a_2]$, the relationship between a and any other interval $b = [b_1, b_2]$ could be divided into the following six types as shown in Fig. 2:

Falling Inside This kind of relationship occurs when interval a is completely falling inside of interval b, as shown in Fig. 2(a). In this situation we have the *OLID* $div(a, b) = 0$.

Covering This relationship happens when interval b is completely falling into interval a, as shown in Fig. 2(b). In this situation we have the *OLID* from a to b as $(|c_a - c_b| + r_a - r_b)(1 - \frac{2r_b}{2r_a+1})$.

Left Overlapping This corresponds to the situation where interval b overlaps with a on the left side of a, as shown in Fig. 2(c). We have $a_1 \in [b_1, b_2]$ and $a_2 \notin [b_1, b_2]$, then $div(a, b) = (|c_a - c_b| + r_a - r_b)(1 - \frac{r_a+r_b-|c_a-c_b|}{2r_a+1})$.

Right Overlapping This corresponds to the situation where interval b overlaps with a on the right side of a, as shown in Fig. 2(d). We have $a_1 \notin [b_1, b_2]$ and $a_2 \in [b_1, b_2]$, then $div(a, b) = (|c_a - c_b| + r_a - r_b)(1 - \frac{r_a+r_b-|c_a-c_b|}{2r_a+1})$.

Left Neighboring This happens when interval b is not overlapping with a, and b is on the left side of a, as shown in Fig. 2(e).

Right Neighboring This happens when interval b is not overlapping with a, and b is on the right side of a, as shown in Fig. 2(f).

By considering all the types of situations, we can get the *Overlapped Interval Divergence* function as follows:

$$\begin{aligned} & div(a,b) \\ & = l(a,b) \cdot (1 - \tfrac{OA(a,b)}{2r_a+1}) \\ & = \begin{cases} 0 & i \\ (|c_a - c_b| + r_a - r_b)(1 - \frac{2r_b}{2r_a+1}) & ii \\ |c_a - c_b| & iii \\ (|c_a - c_b| + r_a - r_b)(1 - \frac{r_a+r_b-|c_a-c_b|}{2r_a+1}) & iv \\ (|c_a - c_b| + r_a - r_b)(1 + \frac{|c_a-c_b|-(r_a+r_b)}{2r_a+1}) & v \end{cases} \end{aligned} \quad (10)$$

in which, i denotes when $|c_a - c_b| \leq r_b - r_a$; ii denotes when $|c_a - c_b| \leq r_a - r_b$; iii denotes when $r_a = r_b = 0$; iv denotes when $|r_a - r_b| < |c_a - c_b| < r_a + r_b$; v denotes when $|c_a - c_b| \geq r_a + r_b$.

It is interesting to note that when both intervals degrade into single values, the *OLID* divergence becomes the regular L_1 distance.

3.2 The Dynamic Clustering Algorithms based on Adaptive OLID

Based on the general framework of the adaptive clustering algorithm introduced in Section 2.2, the clustering algorithms based on single or cluster adaptive *OLID* are developed to discover the best partition of the original data sets into K clusters, which holds the minimum adequacy criterion O_{min}.

$$O_{single} = \sum_{k=1}^{K} \sum_{i \in P_k} d_{single}(x_i, y_k), \quad (11)$$

in which

$$d_{single}(x_i, y_k) = \sum_j^p [\lambda^j (max\{div(a_i^j, \alpha_k^j), div(\alpha_k^j, a_i^j)\})] \quad (12)$$

is the *single adaptive OLID* measuring the dissimilarity between an object $x_i (i = 1, \cdots, n)$ and a cluster prototype $y_k (k = 1, \cdots, K)$, which is the median of $x \in P_k$ and multiplied by a weight vector $\lambda^j (j = 1, \cdots, p)$. Here, since *OLID* is asymmetric, we use the *max* function to make it symmetric.

In each iteration, the weight vector $\lambda^j (j = 1, \cdots, p)$ is calculated according to the following expression:

$$\lambda^j = \frac{\{\prod_{l=1}^p (\sum_{k=1}^K [\sum_{i \in P_k} (max\{div(a_i^l, \alpha_k^l), div(\alpha_k^l, a_i^l)\})])\}^{\frac{1}{p}}}{\sum_{k=1}^K [\sum_{i \in P_k} (max\{div(a_i^j, \alpha_k^j), div(\alpha_k^j, a_i^j)\})]}, \quad (13)$$

which satisfies $\lambda^j > 0$ and $\prod_{j=1}^p \lambda^j = 1$.

The Pseudo-code of the algorithms are presented in Alg. 1, which is both for single and cluster adaptive *OLID* algorithms.

For the dynamic clustering algorithm based on cluster adaptive distances share the same algorithm schema with the one based on single adaptive distances, but using a specific adequacy criterion $O_{cluster}$ since the adaptive distances are different from cluster to cluster:

$$O_{cluster} = \sum_{k=1}^{K} \sum_{i \in P_k} d_{cluster}(x_i, y_k), \quad (14)$$

in which the dissimilarity between the object x_i and the corresponding cluster prototype y_k can be calculated by the *cluster adaptive OLID* equation as follows:

$$d_{cluster}(x_i, y_k) = \sum_j^p [\lambda_k^j (max\{div(a_i^j, \alpha_k^j), div(\alpha_k^j, a_i^j)\})]. \quad (15)$$

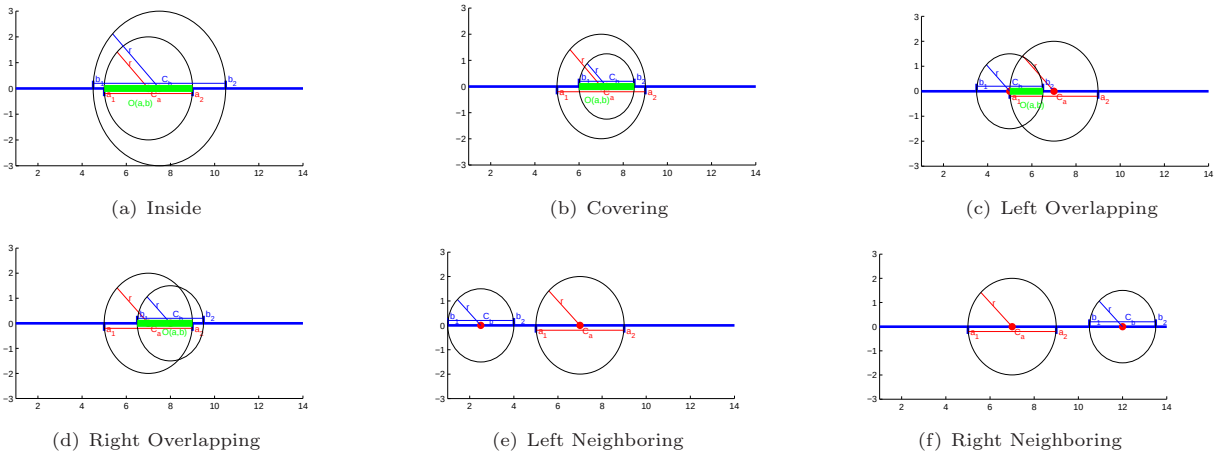
where the weight vector is

$$\lambda_k^j = \frac{\{\prod_{l=1}^p \sum_{i \in P_k} (max\{div(a_i^l, \alpha_k^l), div(\alpha_k^l, a_i^l)\})\}^{\frac{1}{p}}}{\sum_{i \in P_k} (max\{div(a_i^j, \alpha_k^j), div(\alpha_k^j, a_i^j)\})}, \quad (16)$$

which satisfies $\lambda_k^j > 0$ and $\prod_{j=1}^p \lambda_k^j = 1$.

4 Experiment and Discussion

To evaluate the performance of our *OLID* measurement, we investigate it within the framework of dynamic clustering algorithms on synthetic data sets.



(a) Inside (b) Covering (c) Left Overlapping

(d) Right Overlapping (e) Left Neighboring (f) Right Neighboring

Figure 2: Different Relationships Between Two Intervals

Algorithm 1 The adaptive OLID algorithm

Input:
Data Set X.
The number of clusters K.
Output:
A partition P of X into K clusters.
Algorithm Process:

1: **Initialization:**
2: $P \leftarrow$ random partition of input data X into K clusters;
3: **Iterative Research:**
4: $Flag \leftarrow False$;
5: while not $Flag$
6: $Flag \leftarrow TRUE, change \leftarrow 0$;
7: Calculate the prototypes $y_k (k = 1, \cdots, K)$;
8: Calculate the weight vector $\lambda^j (j = 1, \ldots, p)$:

$$\lambda^j = \frac{\{\prod_{l=1}^{p}(\sum_{k=1}^{K}[\sum_{i\in P_k}(max\{div(a_i^l,\alpha_k^l),div(\alpha_k^l,a_i^l)\})]\}^{\frac{1}{p}}}{\sum_{k=1}^{K}[\sum_{i\in P_k}(max\{div(a_i^j,\alpha_k^j),div(\alpha_k^j,a_i^j)\})]}$$

or

$$\lambda_k^j = \frac{\{\prod_{l=1}^{p}\sum_{i\in P_k}(max\{div(a_i^l,\alpha_k^l),div(\alpha_k^l,a_i^l)\})\}^{\frac{1}{p}}}{\sum_{i\in P_k}(max\{div(a_i^j,\alpha_k^j),div(\alpha_k^j,a_i^j)\})},$$

9: for each element $x_i \in X$
10: $k \leftarrow$ the label of the cluster which x_i belongs to;
11: $k_{new} = argmin_k(d_{single}(x_i, y_k))$ or $k_{new} = argmin_k(d_{cluster}(x_i, y_k))$;
12: if $k \neq k_{new}$
13: Assign x_i to $P_{k_{new}}$;
14: $change = change + 1$;
15: end if
16: end for
17: if $change = 0$, then $Flag \leftarrow TRUE$;
18: end while

The results generated based on the *Hausdorff* and the *city-block* distances are also included for comparison purpose. This section starts with an introduction of the experimental data sets, then we describe the Corrected Rand Index (CR Index), which is widely used in the similar studies to evaluate the performance of the clusterings; finally, the experiments results of our *OLID* measurement are shown together with a discussion based on the performance comparison with the other popular measurements.

We compare our measurement with several popular distances, in the context of one/two weight single/cluster adaptive clustering algorithms. As a random initialization step in the dynamic clustering framework, we run each algorithm on each data set for 100 times, and use the average CR over these 100 running for comparison.

4.1 Data Sets

In the experiments, three synthetic interval data sets are employed, which are designed to be well-separated, not-so-well-separated and over-lapping respectively.

4.1.1 Synthetic Data Sets

Following a similar strategy in (de Souza & de A.T. de Carvalho 2004, de A.T. de Carvalho et al. 2006, de A.T. De Carvalho & Lechevallier 2009), three types of the synthetic data sets are generated according to a bivariate normal distribution in a two-dimensional real number space, $\Re^2$. The first one represents a well-separated data set. The second one represents a not-so-well-separated data set, in which the class covariance matrices of the bivariate distribution are unequal; while for the third data set, the class covariance matrices are nearly the same. The parameters listed in Table 3 are set up to generate these three data sets respectively.

Each data set contains 450 data instances. A priori classification is done for evaluation convenience, by which four labels are set up to group the data instances into four classes with different sizes: *Class 1* and *Class 2* have the same size of 150, *Class 3* contains 50 data instances, and *Class 4* takes 100 ones.

As shown in Fig. 3(a), Fig. 3(c) and Fig. 3(e), each data instance (a_i, b_i) in well-separated, not-so-well-separated and over-lapping data sets is a seed of a vector of intervals:

$$([a_i - \gamma_1, a_i + \gamma_1], [b_i - \gamma_2, b_i + \gamma_2]),$$

where γ_1 and γ_2 are randomly picked up from intervals of $[1, 5]$, $[1, 10]$, $[1, 15]$ and $[1, 20]$. The data sets can be also represented by interval values as shown in Fig. 3(b), Fig. 3(d) and Fig. 3(f).

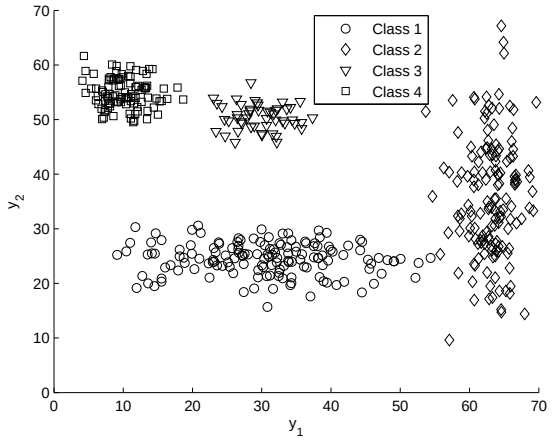
4.2 Clustering Validation

The Corrected Rand (CR) index, which was introduced in (Hubert & Arabie 1985), is one of the most popular clustering validation indexes (de Souza & de A.T. de Carvalho 2004, de A.T. de Carvalho et al. 2006, de A.T. De Carvalho & Lechevallier 2009).

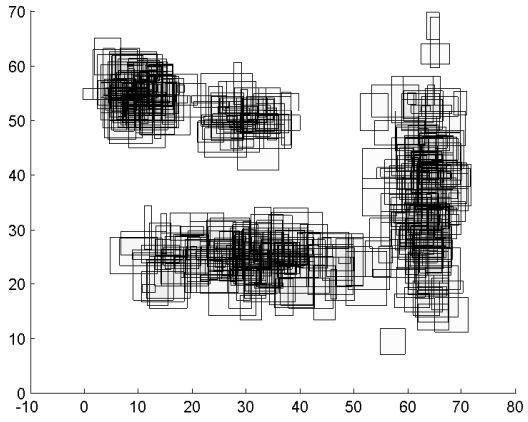
Given two partitions of the same data set, $U = \{u_1, u_2, \ldots, u_R\}$ and $V = \{v_1, v_2, \ldots, v_C\}$, which have

Table 3: Parameters in the three synthetic seed data sets

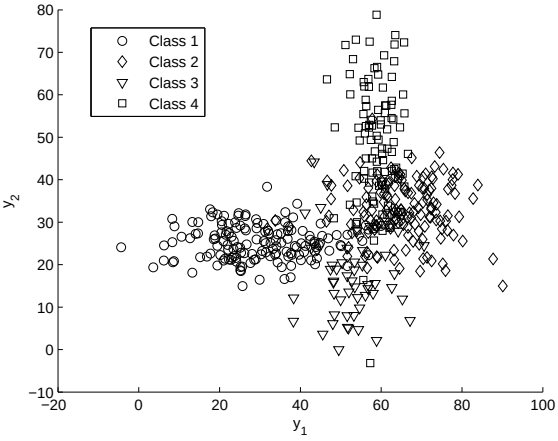
	Well-separated				**Not so Well-separated**				**Over-lapping**			
	μ_1	μ_2	σ_1	σ_2	μ_1	μ_2	σ_1	σ_2	μ_1	μ_2	σ_1	σ_2
Class 1	31	25	10	3	30	25	12	4	30	25	10	3
Class 2	63	34	3	12	64	33	9	7	64	32	9	4
Class 3	30	50	3	3	52	17	7	9	52	17	10	4
Class 4	10	55	3	3	59	50	4	14	59	39	9	3



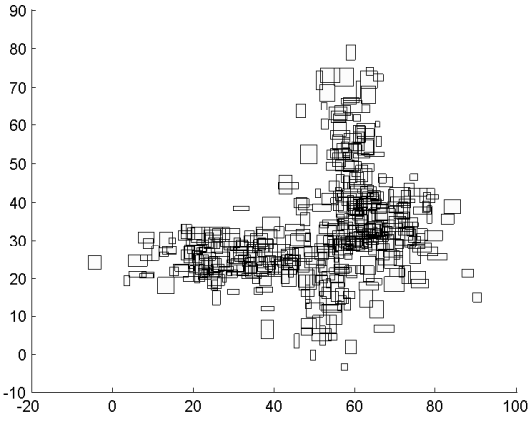
(a) Well-Separated Seeds



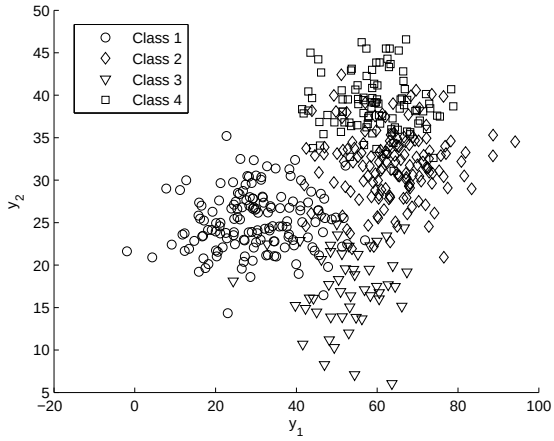
(b) Well-Separated Intervals



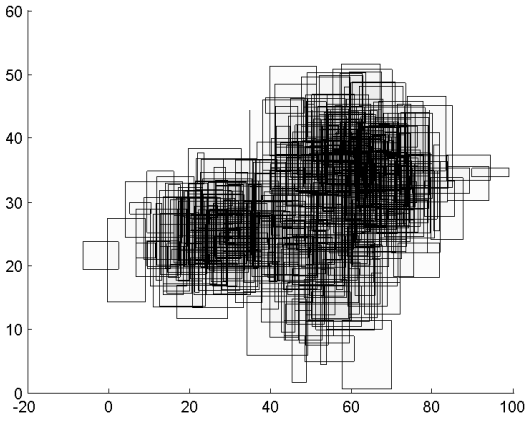
(c) Not so Well-Separated Seeds



(d) Not so Well-Separated Intervals



(e) Over-lapping Seeds



(f) Over-lapping Intervals

Figure 3: Three Synthetic Data Sets ($\gamma_1, \gamma_2 \in [1, 10]$)

R and C clusters respectively, the CR index can be estimated by the following equation:

$$CR = \frac{\sum_{i=1}^{R}\sum_{j=1}^{C}\binom{n_{ij}}{2} - \binom{n}{2}^{-1}\sum_{i=1}^{R}\binom{n_{i.}}{2}\sum_{j=1}^{C}\binom{n_{.j}}{2}}{\frac{1}{2}[\sum_{i=1}^{R}\binom{n_{i.}}{2} + \sum_{j=1}^{C}\binom{n_{.j}}{2}] - \binom{n}{2}^{-1}\sum_{i=1}^{R}\binom{n_{i.}}{2}\sum_{j=1}^{C}\binom{n_{.j}}{2}} \quad (17)$$

in which $\binom{n}{2} = \frac{n(n-1)}{2}$, n_{ij} represents the number of instances that are in clusters u_i and v_i; $n_{i.}$ indicates the number of instances in cluster u_i; $n_{.j}$ indicates the number of instances in cluster v_j; and n is the

total number of instances in the data set.

The value of CR index for a certain clustering algorithm falls into the range of $[-1, 1]$. A CR index value of 1 indicates two clustering results are exactly the same, whereas the value 0 or below indicates that the cluster agreement found by chance (Milligan 1996). When comparing the clustering result with the true clustering partition, the higher the CR index value is, the better the result is.

4.3 Results and Discussion

Each clustering distance is incorporated with the single and cluster adaptive clustering framework, then run 100 times before the averaged CR is calculated.

4.3.1 Results for Synthetic Data Sets

Table 4 presents the values of the average and standard deviation (in parenthesis) of the CR index for the well-separated data set. It is evident that the proposed *OLID* distance in the cluster framework can get the best results on all the testing data sets. This is consistent with De Carvalho's findings (de A.T. De Carvalho & Lechevallier 2009), which discovered that the cluster adaptive clustering framework performs well on the well-separated data sets.

Table 5 presents the values of the average and standard deviation (in parenthesis) of the CR index for the not-so-well-separated data set. This is also consistent with De Carvalho's findings (de A.T. De Carvalho & Lechevallier 2009), which discovered that the cluster adaptive clustering framework performs well on the not-so-well-separated data sets. It is interesting to note that the proposed *OLID* distance in this framework can always lead to the best results on all testing data sets.

Table 6 presents the values of the average and standard deviation(in parenthesis) of the CR index for the over-lapping data set. This is also consistent with De Carvalho's findings (de A.T. De Carvalho & Lechevallier 2009) that the single adaptive clustering framework performs well on the over-lapping data sets. It is as expected that the proposed *OLID* distance outperforms all the other distances in the single adaptive clustering framework on all data sets.

4.3.2 Paired t-test Results

The two-tailed, paired t-test with 95% confidence level has been used to evaluate *OLID* with *city-block* and Hausdorff distance under single and cluster frameworks. The results are presented in Table 7. From the table, we can see that in the cluster adaptive clustering framework, the proposed *OLID* measure significantly improves the existing *city-block* and *Hausdorff* distances. In the single adaptive clustering framework, the *OLID* measure performs significantly better than the *Hausdorff* distance, and it is also significantly better than the two weight *city-block* distance, though the difference between it and the one weight *city-block* distance is not significant.

5 Conclusion

The choice of a distance measurement is essential for the success of many machine learning and data mining tasks, such as clustering and lazy learning. The trend of data representation as the interval-valued data calls for more sophisticated methods to evaluate the distance or similarity between interval-valued data instances.

The work is motivated by the fact that, most existing distance measures for interval-valued data only considered the lower and upper bounds, and overlooked the relative size of their overlapped area. In this paper, we introduce a new distance measurement based on the *Hausdorff* distance and the relative size of the overlapped area. We show its properties, and use it into different dynamic clustering frameworks: the single adaptive *OLID* algorithm and the cluster adaptive *OLID* algorithm. Our experiment results indicate the significant improvement of the proposed *OLID* measure over existing distances. In addition, our results further confirm that the single adaptive clustering framework is suitable for the overlapping data sets, while the cluster adaptive clustering framework is suitable for the well-separated data sets (de A.T. De Carvalho & Lechevallier 2009).

References

Billard, L. (2006), Symbolic data analysis: What is it?, *in* 'Proceedings of 17th Symposium on Computational Statistics (COMPSTAT'06)', Physica-Verlag HD, Rome, Italy, pp. 261–269.

Dai, H., Li, G. & Zhou, Z.-H. (2004), Ensembling mml causal discovery, *in* 'Proceedings of The Eighth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'04)', pp. 260–271.

de A.T. de Carvalho, F., de Souza, R. M., Chavent, M. & Lechevallier, Y. (2006), 'Adaptive hausdorff distances and dynamic clustering of symbolic interval data', *Pattern Recognitionn Letters* **27**(3), 167–179.

de A.T. De Carvalho, F. & Lechevallier, Y. (2009), 'Partitional clustering algorithms for symbolic interval data based on single adaptive distances', *Pattern Recognition* **42**(7), 1223–1236.

de Souza, R. M. & de A.T. de Carvalho, F. (2004), 'Clustering of interval data based on city-block distances', *Pattern Recognitionn Letters* **25**(3), 353–365.

Diday, E. (1988), *Classification methods of data analysis*, Elesvier, North Holland, Amsterdam, chapter The symbolic approach in clustering, related methods of data analysis, pp. 673–684.

Hubert, L. & Arabie, P. (1985), 'Comparing partitions', *Journal of Clasification* **2**(1), 193–218.

Jiang, Y., Ling, J., Li, G., Dai, H. & Zhou, Z.-H. (2005), Dependency bagging, *in* 'Proceedings of The Tenth International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing (RSFDGrC 2005)', pp. 491–500.

Li, G. & Dai, H. (2004), 'What will affect software reuse: A causal model analysis', *International Journal of Software Engineering and Knowledge Engineering* **14**(3), 351–364.

Li, G. & Tong, F. (2002), 'Unsupervised discretization algorithm based on mixture probabilistic model', *Jisuanji Xuebao/Chinese Journal of Computers* **25**(2), 158–164.

Milligan, G. (1996), *Clustering and Classification*, World Scientific, Singapore, chapter Clustering validation: results and implications for applied analysis, pp. 341–375.

Table 4: Well-Separated Data Set: comparison of the distances

	OLID		**city-block**				**Hausdorff**	
Range of γ_i	One weight		One weight		Two weight		One weight	
$(i = 1, 2)$	Single	Cluster	Single	Cluster	Single	Cluster	Single	Cluster
$\gamma_i \in [1, 5]$	**0.7464**	**0.8102**	0.7393	0.8005	0.7378	0.7875	0.7228	0.7859
CR Index	(0.0935)	(0.1038)	(0.0879)	(0.0954)	(0.0926)	(0.0967)	(0.0726)	(0.1071)
$\gamma_i \in [1, 10]$	0.761	**0.8003**	**0.773**	0.796	0.7556	0.7869	0.7489	0.7932
CR Index	(0.1108)	(0.0895)	(0.1197)	(0.0821)	(0.1096)	(0.0768)	(0.1158)	(0.0998)
$\gamma_i \in [1, 15]$	**0.7777**	**0.8122**	0.7455	0.7715	0.7558	0.7805	0.77	0.7932
CR Index	(0.098)	(0.1037)	(0.1088)	(0.0863)	(0.1157)	(0.0939)	(0.1349)	(0.0962)
$\gamma_i \in [1, 20]$	**0.7352**	**0.8038**	0.6994	0.7442	0.6982	0.755	0.7173	0.7345
CR Index	(0.1007)	(0.0993)	(0.0743)	(0.0699)	(0.0734)	(0.0872)	(0.0909)	(0.0801)

Table 5: Not So Well-Separated Data Set: comparison of the distances

	OLID		**city-block**				**Hausdorff**	
Range of γ_i	One weight		One weight		Two weight		One weight	
$(i = 1, 2)$	Single	Cluster	Single	Cluster	Single	Cluster	Single	Cluster
$\gamma_i \in [1, 5]$	**0.5235**	**0.5555**	0.4967	0.5275	0.4905	0.5315	0.4691	0.5303
CR Index	(0.0470)	(0.0589)	(0.0183)	(0.0808)	(0.0354)	(0.0768)	(0.0520)	(0.0834)
$\gamma_i \in [1, 10]$	0.5139	**0.5682**	0.5262	0.5561	**0.5305**	0.5526	0.4788	0.5332
CR Index	(0.0642)	(0.0611)	(0.0386)	(0.0392)	(0.0364)	(0.0435)	(0.0829)	(0.0937)
$\gamma_i \in [1, 15]$	**0.4329**	**0.4869**	0.4309	0.4683	0.4327	0.4691	0.4132	0.4198
CR Index	(0.0091)	(0.0347)	(0.0172)	(0.0118)	(0.0071)	(0.0017)	(0.0181)	(0.0428)
$\gamma_i \in [1, 20]$	0.4606	**0.5225**	**0.4883**	0.4967	0.4792	0.4946	0.3568	0.3769
CR Index	(0.0412)	(0.0487)	(0.0335)	(0.0843)	(0.0232)	(0.0725)	(0.0270)	(0.0444)

Table 6: Over-Lapping Data Set: comparison of the distances

	OLID		**city-block**				**Hausdorff**	
Range of γ_i	One weight		One weight		Two weight		One weight	
$(i = 1, 2)$	Single	Cluster	Single	Cluster	Single	Cluster	Single	Cluster
$\gamma_i \in [1, 5]$	**0.6079**	**0.6058**	0.5964	0.5679	0.5855	0.5953	0.5777	0.5619
CR Index	(0.114)	(0.1241)	(0.0999)	(0.0977)	(0.0993)	(0.1014)	(0.1137)	(0.105)
$\gamma_i \in [1, 10]$	**0.5958**	**0.5826**	0.5672	0.5451	0.5689	0.5521	0.5598	0.5355
CR Index	(0.0536)	(0.0765)	(0.0627)	(0.081)	(0.066)	(0.0809)	(0.0292)	(0.0548)
$\gamma_i \in [1, 15]$	**0.5211**	**0.5350**	0.5168	0.5040	0.5203	0.5163	0.4816	0.4720
CR Index	(0.0211)	(0.0722)	(0.0394)	(0.0759)	(0.0479)	(0.0847)	(0.0445)	(0.0566)
$\gamma_i \in [1, 20]$	**0.6416**	0.5546	0.5755	**0.5548**	0.5691	0.5395	0.5335	0.5090
CR Index	(0.0845)	(0.0947)	(0.0654)	(0.0799)	(0.0654)	(0.0911)	(0.0604)	(0.0684)

Table 7: Paired t-test Results

Algorithms	**Distance**	p-value	**t**
Single OLID vs. Single city-block	One weight	0.0961	1.8198
	Two weight	0.0492	2.2103
Single OLID vs. Single Hausdorff	one weight	0.0013	4.2584
Cluster OLID vs. Cluster city-block	One weight	0.0004	5.0638
	Two weight	0.0000	7.5389
Cluster OLID vs. Cluster Hausdorff	one weight	0.0006	4.738