ISSN 1022 - 2502, E-ISSN 2408- 8528
Dhaka Univ. J. Sci. 72(1): 46-55, 2024 (January) DOI: https://doi.org/10.3329/dujs.v72i1.71184

Implementing Vertices Principal Component Analysis under Various Weighting Schemes for Interval Valued Observations with Applications to Data Mining

Md. Anwarul Islam Bhuiyan, Sohana Jahan and Mohammad Babul Hasan

Department Mathematics, University of Dhaka, Dhaka-1000, Bangladesh

(Received : 20 August 2023; Accepted : 13 December 2023)

Abstract

Data mining is the technique for deriving valuable data from a more extensive collection of raw data. It is the process of looking for irregularities, trends, and correlations in huge data sets in order to forecast results. Although a number of techniques have been developed to perform data mining on conventional data in the past years, there are huge scope to work with Interval Valued data (IVD). Working with IVD has been shown to be of significant importance when it comes to identifying the objective entity in a precise manner or representing incomplete knowledge on life situations. Unlike classical data where each object is represented by a point, in IVD the objects are represented by regions in Rp. In this paper, an extension of Principle Component Analysis (PCA) known as Vertices Principal Components method for interval-valued information has been explored. It additionally incorporated the relative contributions of the vertices depending on different choices of weighting schemes. A new idea for classification of the supervised IVD is proposed which is based on the idea of K-Nearest Neighbor (KNN) technique. The proposed approach is implemented on several benchmarking data sets. Numerical results suggest the proper choice of weighting schemes for each of the data set that will lead to better recognition rate.

Keywords: Data Mining, Interval Valued Data, Principal Component Analysis, Vertices Principal Component Analysis, K-Nearest Neighbor, Distance Matrix.

I. Introduction

Within the interest of knowledge, information plays an important role. Data[1,2] is made up of discrete values that describe amount, quality, fact, statistics, and other fundamental units of meaning. It can be expressed in words, details, observations, pictures, numbers, graphs, or symbols. Data is information that has been transformed into a format that is useful for transfer or processing in computing.

Generally, we use the classical data set to represent information or knowledge in which each data point is considered as single point. To represent the data set which is not possible to express by a particular point, Diday.[2] introduced the idea of symbolic data set to present such phenomenon. Symbolic dataset[1,2] consists of intervals, lists, histograms, distributions etc. Interval-valued data (IVD), a type of symbolic data, is given as an interval in which the observation object can occur frequently in the process of aggregating large databases into a form that is easy to manage. Medical Health Demographics, Iris data, Haemoglobin by gender age groups, Cholesterol by gender age groups, Blood pressure data, Mushroom data, temperature (in Celsius or Fahrenheit), mark grading etc. are the examples of IVD. In market research or in any other forms of medical, educational, social, economic or business research interval valued data[3,4] plays a pivotal role.

L. Billard et al.[2] extended the method of finding principal components in case of interval valued data. They proposed vertices method[2] on the basis of all vertices of hypercube and centre method using the centroid values. A. Douzal-Chouakria et al.[1] studied principal aspect evaluation for interval-valued data and added the concept of vertex contributions[1] to the underlying primary additives. By combining both the midpoints (or centers) and the radii (a measure of the interval width) of (IVD), P. D'Urso et al.[7] proposed an extension of convention or classical PCA which is Midpoint Radius Principal Component Analysis (MR-PCA)[7]. H. Wang et al.[8] proposed a new PCA method called Complete Information based Principal Component Analysis (CIPCA)[8] which defines the inner product of interval-valued variables and gives a proficient and powerful way for directing PCA for enormous scaled mathematical information, X. Qi et al.[9] introduced the Uniform Representative Framework (URF)[9] to better describe the structural information of the IVD. In addition, symmetric uncertainty (SU)[9] was applied to quantitatively measure the relationship between features and classes. t is quite common for real random variables to be seriously observed or so improbable that the results would need to be recorded as actual intervals containing specific data from an experiment. In some cases, due to specific confidentiality reasons, the specified value of a variable may be kept in an encrypted format. In such cases, researchers are interested to consider interval valued observations instead of classical data.

Our main contributions in this work are

- Here, selection process of appropriate weight scheme[15] is suggested depending on the relative contribution of the reconstructed datasets (expressed in terms of vertices) on principal components which will lead to better recognition rate of testing data.

*Author for correspondence. e-mail: sjahan.mat@du.ac.bd

- This paper works with supervised data. Vertices principal component analysis is applied on training data. Therefore, transformation matrix is obtained which is applied on testing data set to project the data in reduced dimensional space.
- Finally, a new idea for classification of the supervised IVD is proposed which is based on the idea of K-Nearest Neighbor (KNN) technique.
- The proposed approach is implemented on several benchmarking data sets. Numerical results suggest the proper choice of weighting schemes for each of the data set that will lead to better recognition rate

The rest of the paper is designed as follows.

In the next section, formulation process of the problem has discussed including classification. The following sections (III to IV) include a brief review of PCA, Vertices PCA, Variance-Covariance matrix and Calculation process of finding Principal component with contribution. In section V, the idea of Distance matrix for interval dissimilarities has discussed. In section VI, the determination process of appropriate weight scheme has discussed depending on relative contribution of vertices and idea Distance matrix for interval dissimilarities has applied as classification process. Finally, the conclusion of this work has drawn in section VII.

Problem Formulation

Given a set of interval valued data where the data points belong to two or more classes. The problem is to reduce the dimension of data set to select important features and use these features to identify the class or label of unknown or new data.

Suppose the original Interval Valued Dataset consists of m observations $\mu_i = (\mu_{i1}, \mu_{i2}, \dots, \mu_{ip}), i = 1,2,,,,,m$ with p variables $X = (X_1, X_2, , , , X_p)$. Each observation $\mu_{ij=}[a_{ij}, b_{ij}], i = 1,2, \dots, m, j = 1,2, \dots, p$ is a non-trivial interval valued data that is $a_{ij} \leq b_{ij}$. Thus, the data set has the form

$$\begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{pmatrix} = \begin{pmatrix} [a_{11}, b_{11}] & [a_{12}, b_{12}] & \cdots & [a_{1p}, b_{1p}] \\ [a_{21}, b_{21}] & [a_{22}, b_{22}] & \cdots & [a_{2p}, b_{2p}] \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ [a_{m1}, b_{m1}] & [a_{m2}, b_{m2}] & \cdots & [a_{mp}, b_{mp}] \end{pmatrix}.$$

The basic steps that are taken in this work to deal with this interval valued data is illustrated in Fig. 1.

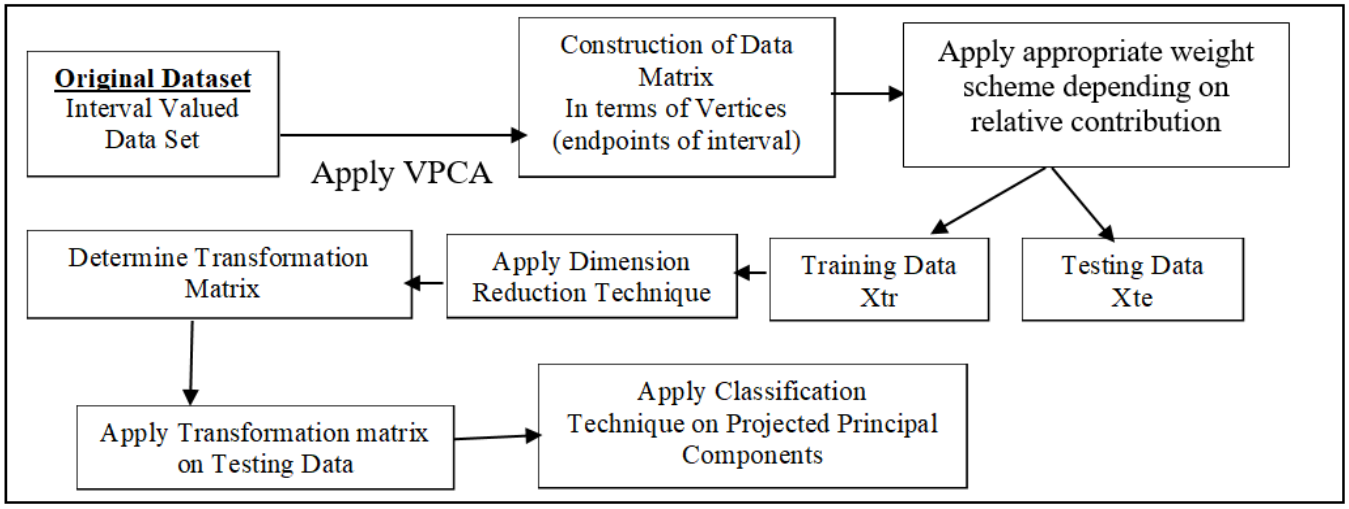

Fig. 1. Basic structure for classification for IVD by dimension reduction using VPCA

Principal Component Analysis (PCA)

Dimension reduction is a very important pre-processing step for classification of data. Many data set are usually reduced to 2 or 3-dimension for visualization purpose. Also, data with less dimension are simpler to investigate. PCA[11,12], is one kind of dimensionality-reduction strategy that's frequently utilizes to reduce the dimensionality of data yet preserves important information that highlight the similitudes and contrasts.

PCA reduces the dimension of the data sets by computing new variables Principal Components (PCs) that are constructed as linear combinations or mixtures of the initial variables. The first principal component has maximum variance (among all linear combinations) and accounts for as much variation in the data as possible.

Principal Component Analysis[12] can be spitted into the following steps:

1. Standardize the initial variables.

2. Computation of covariance matrix.
3. Finding the eigenvectors and eigenvalues of the covariance matrix and choose the

   dominant eigenvalue to calculate the Principal Components.
4. Represent the original data set in terms of Principal Component's axes.

Vertices Principal Component Analysis (VPCA)

VPCA[2] was first introduced by L. Billard et al. in 2008. They proposed this method on the basis of all vertices of hypercube. In this section we will briefly discuss the process of applying VPCA on IVD.

To apply VPCA the first step is to construct data in term of vertices as discussed below

Transformation Process of Interval Valued Dataset to Classical Dataset

Suppose the number of nontrivial intervals in μ_i is s_i, then the number of vertices corresponding to μ_i is 2^{s_i}. Thus, the total number of vertices in dataset $(\mu_1, \mu_2, \dots, \mu_m)$ is

$$n = \sum_{i=1}^{m} n_i = \sum_{i=1}^{m} 2^{s_i} . \quad (1)$$

The data matrix for the observation μ_i can be written as

$$X_{\mu_i} = \begin{pmatrix} x_{11}^i & \cdots & x_{1p}^i \\ \vdots & \cdots & \vdots \\ x_{k1}^i & \cdots & x_{kp}^i \\ \vdots & \cdots & \vdots \\ x_{n_i 1}^i & \cdots & x_{n_i p}^i \end{pmatrix},$$

where $x_k^i = \left(x_{k1}^i, x_{k2}^i, \dots, x_{kp}^i\right)$ is the coordinate of the vertex $k = 1,2, \dots n_i$ related to the hypercube H_i representing the observation $\mu_i, i = 1,2, \dots m$.

Then the complete data matrix X in terms of vertices is the following $n \times p$ matrix

$$X = \left(x_{ij}\right)_{1 \le i \le n, 1 \le j \le p} = \begin{pmatrix} X_{\mu_1} \\ \vdots \\ X_{\mu_m} \end{pmatrix}$$

$$= \begin{pmatrix} \begin{pmatrix} x_{11}^1 & \cdots & x_{1p}^1 \\ \vdots & \cdots & \vdots \\ x_{n_1 1}^1 & \cdots & x_{n_1 p}^1 \end{pmatrix} \\ \vdots \\ \begin{pmatrix} x_{11}^m & \cdots & x_{1p}^m \\ \vdots & \cdots & \vdots \\ x_{n_m p}^m & \cdots & x_{n_m p}^m \end{pmatrix} \end{pmatrix}, \quad (2)$$

where $x_{ij} \in \mathbb{R}^p,\ i = 1,2, \dots, n;$ where $n = \sum_{i=1}^{m} n_i$.

Variance-Covariance matrix

After generating the matrix X, the Variance – Covariance matrix is calculated by

$$V = X^T D X , \quad (3)$$

where X is defined as (2) and D is the $n \times n$ diagonal matrix containing the weight functions

$$D = diag\left(W_1^1, W_2^1, \dots, W_{n_i}^1, \dots, W_1^m, \dots, W_{n_m}^m\right), \quad (4)$$

where $\left(W_1^i, W_2^i, \dots, W_{n_i}^i\right)$ denotes the weight corresponding to the observation μ_i. In the next section, different weigh scheme that can be applied to this variance -covariance matrix is discussed.

Weights

In determining relative contribution of vertices on principal component, weights imposed on vertices plays an important role. L. Billard et al. suggested different weighting schemes in their work: i) Equal weight for each observation, ii) weights based on internal variations of hypercubes, iii) weights inversely proportional to the volume of hypercubes. Each of these weight schemes are analysed by applying them of several datasets which are discussed in numerical part. The following section includes a brief description of these weight scheme

Weight Scheme 1 (Equal weight for each observation)

Let the weight function associated with the vertex k (of μ_i) be $W_k^i, k = 1,2, \dots, n_i, i = 1,2, \dots, m.$ Here each observation μ_i has n_i vertices and have a weight factor W_i. Therefore,

$$W_i = \sum_{k=1}^{n_i} W_k^i, \sum_{i=1}^{m} W_i = 1.$$

If we consider equal weight for each observation then

$$W_i = \frac{1}{m}, i = 1,2,,, m \quad (5)$$

and therefore $W_k^i = \frac{1}{m * n_i}, k = 1,2, \dots, n_i,\ i = 1,2, \dots, m.$

Weight Scheme 2 (W_i based on internal variations of hypercubes)

Suppose V_i is the volume of hypercube H_i associated with the observation μ_i given by

$$V_i = \prod_{a_{ij} \ne b_{ij}} \left(b_{ij} - a_{ij}\right). \quad (6)$$

Weight based on internal variations of hypercubes $H_i,\ i = 1,2, \dots, m$ can be written as

$$W_i = \frac{V_i}{\sum_{i=1}^{m} V_i}. \quad (7)$$

Weight Scheme 3 (W_i inversely proportional to the volume of hypercubes)

In this case, the weight functions are considered to be inversely proportional to the volume of hypercube i.e.

$$W_i = \frac{1-\frac{V_i}{\sum_{i=1}^{m} V_i}}{\sum_{i=1}^{m}\left[1-\frac{V_i}{\sum_{i=1}^{m} V_i}\right]}. \quad (8)$$

Vertices Principal Components and Relative Contribution of Vertices

Here we first determine the eigenvectors and eigenvalues of the weighted variance–covariance matrix *V* and choose the dominant eigenvalue to calculate the Principal Components.

The v^{th} symbolic vertices principal components of weighted variance–covariance matrix *V* given by equation (3) for the observation μ_i is given by

$$\left.\begin{array}{c} Y_{iv}^{*} = \left[y_{iv}^{a}, y_{iv}^{b}\right], v = 1,2,\dots,p\,, \\ y_{iv}^{a} = min\{y_{kv}^{i}\}\,, \\ y_{iv}^{b} = max\{y_{kv}^{i}\} \\ \text{where } y_{kv}^{i} = PC_{kv}^{i}\,, i = 1,2,\dots,m, \\ k = 1,2,\dots,n_i \text{ and } v = 1,2,\dots,p. \end{array}\right\} \quad (9)$$

Then principal components related to v^{th} eigenvector and associated with $x_{ij} \in \mathbb{R}^p$ is ,

$$PC_v = \sum_{j=1}^{p} e_{vj}\left(x_{ij} - \bar{X}_J\right), \quad (10)$$

where $e_v = \left(e_{v1}, e_{v2}, , , , , e_{vp}\right), v = 1,2, , , p$ and the weighted sample mean is

$$\bar{X}_j = \sum_{i=1}^{m}\sum_{k=1}^{n_i} w_k^i x_{kj}^i.$$

Relative contribution of Vertices on Principal Components

Finally, the relative contribution of vertices on principal components PC_v is determined by

$$Ctr\left(x_k^i, PCv\right) = \frac{\left(y_{vk}^i\right)^2}{\left[d(x_k^i, G)\right]^2}, \quad (11)$$

where *G* is the centroid of the data set.

Next, the required principal components (PC_v) are determine which will be used to project the data set into lower dimensional space.

In the next section, distance matrix for interval valued data is discussed which will be used for classification of testing data.

Distance Matrix for Interval Valued Dataset

To develop distance matrix[10] for interval valued dataset, the ranges of dissimilarities[3] must be represented by ranges of distances. For this the hypercubes are considered as the objects and the upper and lower bound of the distance interval are approximated.

Let the rows of the matrix *X* of order $n \times p$ contains p intervals each of which represents the coordinates of the centers (denoted by x_{is}) of edges of the hypercubes and the corresponding radius or spread , where *n* is the number of objects and *p* is the dimensionality.

The coordinates of center x_{is} are defined by

$$x_{is} = \frac{y_{is}^a + y_{is}^b}{2}, s = 1,2,\dots,p \quad (12)$$

and the distance from the center of hypercube i along the axis *s* called the spread, is denoted by $r_{is} \geq 0$ and is defined by

$$r_{is} = \frac{\left|y_{is}^a - y_{is}^b\right|}{2}, s = 1,2,\dots,p. \quad (13)$$

Then the maximum Euclidean distance between rectangles *i* and *j* is given by:

$$d_{ij}^{(U)}(X,R) = \left(\sum_{s=1}^{p}\left[\left|x_{is} - x_{js}\right| + \left(r_{is} + r_{js}\right)\right]^2\right)^{\frac{1}{2}} \quad (14)$$

and the minimum Euclidean distance by

$$d_{ij}^{(L)}(X,R) = \left(\sum_{s=1}^{p} \max\left[\left|x_{is} - x_{js}\right| - \left(r_{is} + r_{js}\right)\right]^2\right)^{\frac{1}{2}} \quad (15)$$

Thus, the distance matrix Δ containing the entries as intervals of distances is given by

$$\Delta = \left(\left[d_{ij}^{(L)}, d_{ij}^{(U)}\right]\right)_{i,j=1,2,,,,m} = \begin{pmatrix} \left[d_{11}^{(L)}, d_{11}^{(U)}\right] & \left[d_{12}^{(L)}, d_{12}^{(U)}\right] & \cdots & \left[d_{1m}^{(L)}, d_{1m}^{(U)}\right] \\ \left[d_{21}^{(L)}, d_{21}^{(U)}\right] & \left[d_{22}^{(L)}, d_{22}^{(U)}\right] & \cdots & \left[d_{2m}^{(L)}, d_{2m}^{(U)}\right] \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ \left[d_{m1}^{(L)}, d_{m1}^{(U)}\right] & \left[d_{m2}^{(L)}, d_{m2}^{(U)}\right] & \cdots & \left[d_{mm}^{(L)}, d_{mm}^{(U)}\right] \end{pmatrix}. \quad (16)$$

Eq.14 and 15 and therefore the distance matrix Δ will be used for the classification task in the next section.

II. Experiment and Results

Dataset description

We have implemented the idea of Vertices Principal Component Analysis on two different interval valued datasets (Facial dataset and Blood Pressure dataset).

Face Dataset

The importance of face recognition[5] problem has increased rapidly specially in the context of security. However, for different portraits of the same person sometimes there may be slight variations in measurements due to several reasons. In that case, to develop a model for face recognition,

interval valued measurement will lead to better performance. In this work, a VPCA is applied on face dataset[3] which is collected from Leroy et al. (1996). In this dataset, each observation contains 6 (six) random variables each of which is an interval and represents, the distance between two points of some specific part of face image is measured based on the number of pixels on that image. For example, the variable X1 stands for the distance spanned by the eyes (AD in Fig. 02) that is the distance between the outer corner of two eyes , X2 defines the distance between the inner corner of two eyes (BC), X3 stands for the distance between outer corner of right eye and upper middle lip (AH), X4 denotes corresponding length for left eye (DH), X5 stands for the distance between the outside of the mouth on right side and upper middle lip (EH) and X6 indicates the corresponding length to the left side of the mouth (GH).

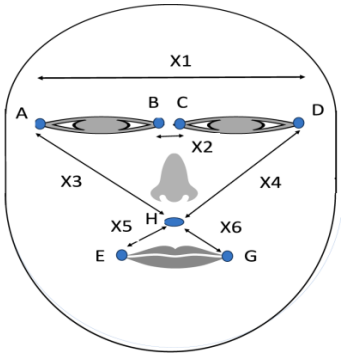

Fig. 2. Faces[3,4]: Description of 6 random variables

Due to the different conditions of alignment, illumination, pose and occlusion, the lengths will vary for different image of the same person. That's why three sequences of images are considered for one person. Here this dataset included 9 (nine) men with 3 (three) sequences of images which gives a total of 27 observations. The complete dataset will be 27×6 dimensional Interval valued dataset.

Table 1. First three rows of 27×6 dimensional interval valued face dataset

Person	X1 = AD	X2 = BC	X3 = AH	X4 = DH	X5 = EH	X6 = GH
FRA1	[155, 157]	[58, 61.01]	[101.45, 103.28]	[105, 107.3]	[61.4, 65.73]	[64.2, 67.8]
FRA2	[154, 160.01]	[57, 64]	[101.98, 105.55]	[104.35, 107.3]	[60.88, 63.03]	[62.94, 66.47]
FRA3	[154.01, 161]	[57, 63]	[99.36, 105.65]	[101.04, 109.04]	[60.95, 65.6]	[60.42, 66.4]

Blood Pressure Dataset

Blood pressure data has been obtained from the data table jointly formed by Lynn Billard and Edwin Diday in 2007. This dataset is available in the webpage of Department of statistics, Franklin College of Arts and Sciences, University of Georgia. It consists of 15 data points each with 3 attributes. The three attributes each of which are given in interval valued form, describe the Pulse Rate, Systolic Pressure, and Diastolic Pressure of a person.

Table 2. Blood Pressure dataset

Person	Pulse Rate	Systolic Pressure	Diastolic Pressure
1	[44, 68]	[90, 110]	[50, 70]
2	[60, 72]	[90, 130]	[70, 90]
3	[56, 90]	[140, 180]	[90, 100]
4	[70, 112]	[110, 142]	[80, 108]
5	[54, 72]	[90, 100]	[50, 70]
6	[70, 100]	[134, 142]	[80, 110]
7	[72,100]	[130, 160]	[76, 90]
8	[76, 98]	[110, 190]	[70, 110]
9	[86, 96]	[138, 180]	[90, 110]
10	[86, 100]	[110, 150]	[78, 100]
11	[53, 55]	[160, 190]	[205, 219]
12	[50, 55]	[180, 200]	[110, 125]
13	[73, 81]	[125, 138]	[78, 99]
14	[60, 75]	[175, 194]	[90, 100]
15	[42, 52]	[105, 115]	[70, 82]

Numerical Results

The procedure of applying vertices principal component analysis and classification for both datasets is given bellow:

- Reconstruct the data set in terms of vertices.
- Chose an appropriate Weighting Scheme depends on relative contribution of vertices on principal components.
- Split the constructed data set into training set those are labelled and testing whose classes will be identified.
- Determine distance between a test image to each of the training images to classify the testing image.

In Face dataset, there are 27 observations and 6 variables. So according to (1) the total number of vertices will be $27 \times 2^6 = 1728$ that means we will get 1728×6 dimensional dataset in terms of vertices. The comparison of Average Relative Contributions (ARC) of vertices on Principal Components (PC) (Using Eq. 11 and 12) for three different choices is given in the following Fig. 03.

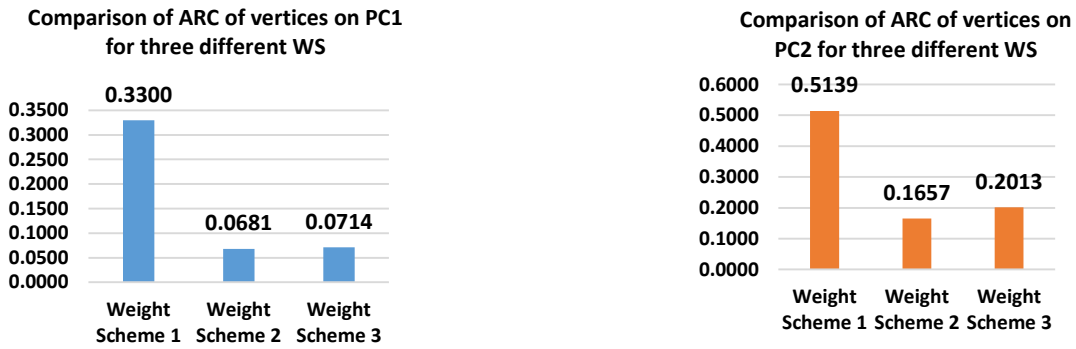

Fig. 3. Average Relative Contributions (ARC) of vertices on Principal Components (PCs) for three different weighting schemes for face dataset

From Fig. 3, we observed that the average relative contribution of vertices on principal components is the highest for the choice of weight scheme 1 that means equal weight, So calculation of the next portion will be carried using equal weight to each observation. Split the constructed dataset into training set which contains the information of the first two images of each person and testing set which contains the information of the third image of each person. Calculate first two principal components of the training set. Then applying the idea of Supervised PCA (eigenvectors obtained from the training set), calculate the first two principal components of the testing set.

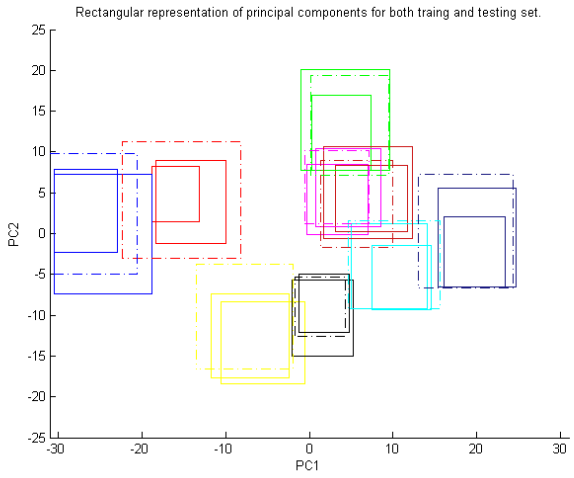

Fig. 4. Rectangular representation[3,4] of principal components for training and testing sets for face datasets where each rectangle is drawn considering minimum and maximum value of PCs. Solid lined rectangle represents the training set and the dotted lined rectangle represents the testing set.

From Fig. 4, it is clear that same-coloured solid lined rectangles and the dotted lined rectangle stay almost together which means testing data match with training data.

Table 3. Distance Matrix of order 18 × 9 of face dataset

[13.42, 15.48]	[13.43, 33.45]	…	…	…	…
[15.70, 17.63]	[10.35, 34.57]	…	…	…	…
…	…	…	….	[7.23, 16.87]	[11.95, 17.17]
…	…	….	…	[9.10, 27.97]	[15.19, 18.00]

Each element of Table 3 represents the distance interval where the lower limit (using Eq. 14) has obtained by considering the minimum distance between each testing image with all training images. Similarly, for the upper limit (using Eq. 15) the maximum distance has considered.

Table 4 represents the difference matrix where each element is computed by taking the difference between the upper limit and lower limit of each interval of distance matrix which is given in Table 3.

Table 4. Difference matrix of order 18 × 9 of face dataset

Training Image	**Distances**								
	FRA3	HUS3	INC3	ISA3	JPL3	KHA3	LOT3	PHI3	ROM3
FRA1	2.06	20.01	15.33	20.75	12.79	24.72	17.83	14.04	18.77
FRA2	**1.93**	24.22	15.65	23.59	15.55	26.03	22.09	14.91	20.85
HUS1	26.69	2.58	25.05	23.55	10.32	27.69	17.23	14.85	28.05
HUS2	31.47	**1.78**	29.80	25.75	12.95	31.74	20.51	16.00	33.41
INC1	15.26	22.38	**0.64**	21.34	16.25	28.01	19.00	15.90	19.49
INC2	18.03	29.00	6.59	23.83	22.09	30.75	22.86	21.42	21.83
ISA1	26.52	23.22	23.01	3.37	21.84	20.91	14.64	20.00	18.18
ISA2	27.97	24.36	24.53	**1.09**	22.09	24.00	17.07	16.46	17.37
JPL1	17.63	14.93	19.62	20.95	2.35	28.34	15.84	3.72	19.36
JPL2	20.03	11.42	21.24	20.69	**1.74**	29.46	18.07	4.37	21.36
KHA1	31.01	27.31	32.51	27.65	24.03	7.56	15.40	27.23	29.60
KHA2	29.53	27.98	32.02	26.13	24.80	**4.26**	14.97	27.46	28.34
LOT1	31.39	22.38	27.09	20.63	17.68	11.00	2.49	21.04	26.63
LOT2	28.70	20.35	24.48	17.57	15.67	13.22	**1.19**	19.07	23.70
PHI1	20.16	15.78	21.18	14.45	7.48	29.07	19.29	**3.36**	18.89
PHI2	21.85	14.71	23.38	16.38	5.75	32.67	21.79	3.95	20.68
ROM1	24.21	26.33	21.33	13.77	20.01	22.99	17.20	19.64	5.22
ROM2	25.18	29.50	22.49	16.89	20.80	26.60	20.98	18.87	**2.82**

Table 5. Classification of testing images for face dataset

Testing Image	**Minimum Differences**	**Match with Training Image**
FRA3	1.93	FRA2
HUS3	1.78	HUS2
INC3	0.64	INC1
ISA3	1.09	ISA2
JPL3	1.74	JPL2
KHA3	4.26	KHA2
LOT3	1.19	LOT2
PHI3	3.36	PHI1
ROM3	2.82	ROM2

Table 5 indicates that all the images of the testing set are recognized based on the minimum differences of distance matrix that means recognition rate is 100 percent.

For blood pressure dataset, at first need to classify the whole data set into three parts based on the following Table 6.

Table 6. Blood Pressure (BP) Levels

BP at Normal Ranges	Systolic: <120 mm Hg Diastolic: <80 mm Hg
BP at Risk Level	Systolic: From 120 to 139 mm Hg Diastolic: From 80 to 89 mm Hg
High Blood Pressure (BP)	Systolic: ≥140 mm Hg Diastolic: ≥90 mm Hg

Table 7. Different Classes of Patients based on Blood Pressure Labels

Blood Pressure Levels	Persons	Label
Normal	1,5,15	1
At Risk (prehypertension)	2,13	2
High Blood Pressure (hypertension)	3,4,6,7,8,9,10,11,12,14	3

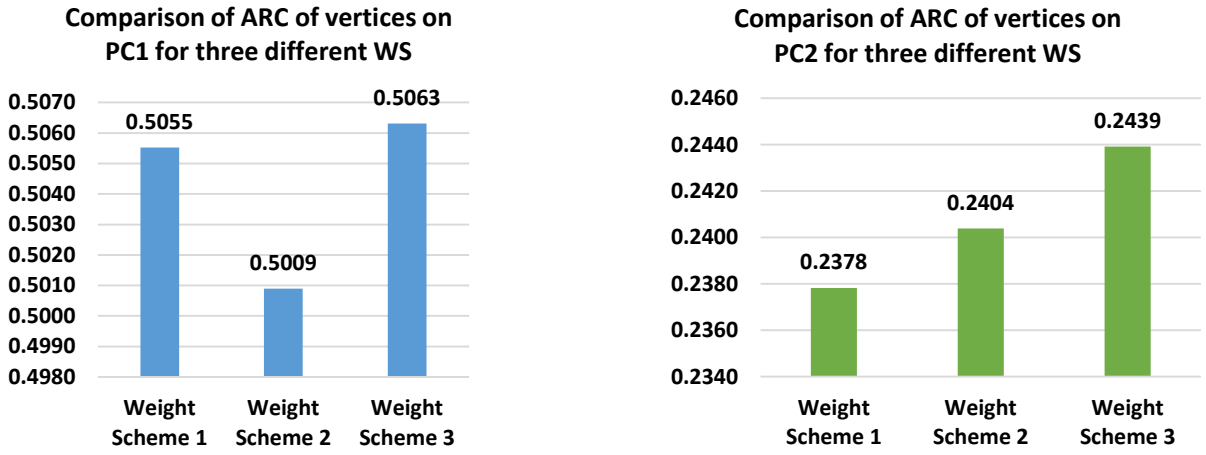

Fig. 5. Average Relative Contributions (ARC) of vertices on Principal Components (PCs) for three different weighting schemes for Blood Pressure dataset

According to the above Fig. 05, for blood pressure dataset the ARC of vertices on Principal Components is the highest for the choice of Weight Scheme 3. So, calculation of the next portion will be carried using Weights those are inversely proportional to the volume of the hypercube.

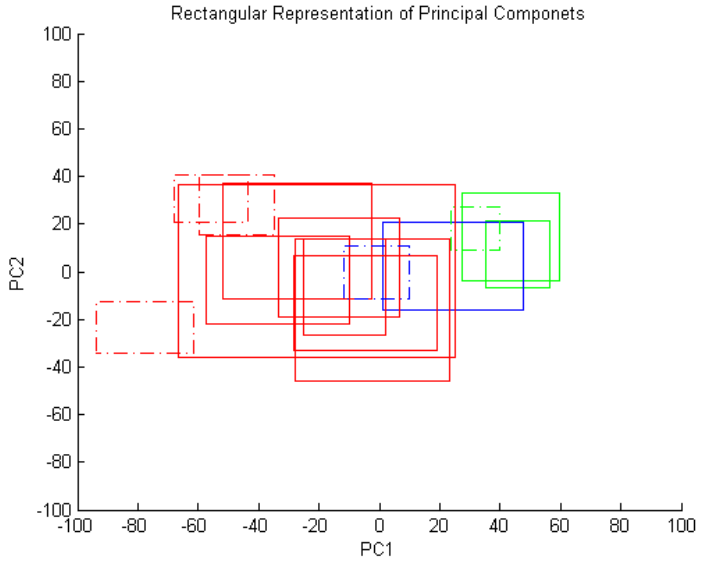

Fig. 6. Rectangular representation[3,4] of principal components for training and testing sets of Blood pressure dataset where each rectangle is drawn considering minimum and maximum value of PCs. Solid lined rectangle represents the training set and the dotted lined rectangle represents the testing set.

From Fig. 6, it is clear that same-coloured solid lined rectangles and the dotted lined rectangle stay almost together which means testing data match with training data. Split the constructed dataset into training set and testing test according to Table 8.

Table 8. Training and Testing sets for Blood Pressure dataset

Data Sets	Persons
Training Set	Normal (2), At Risk (1), High (7)
Testing Set	Normal (1), At Risk (1), High (3)

Table 9 gives the recognition rate of classification of testing image of blood pressure data based in different values of distance interval of the distance matrix. The number of misclassified testing data is one (out of five) for the upper limit of distance interval that is the recognition rate is 80%. But in terms of lower limit recognition rate is 60% and 40% for the choice of minimum differences.

Table 9. Classification of testing data of Blood Pressure Dataset based on Distance Matrix

	Match with Training Patient's Label		
Testing Patients (Label)	Using Upper Limit of Distance Interval	Using Lower Limit of Distance Interval	Using Difference of Distance Interval
15(1)	1	3	1
13(2)	3	1	3
11(3)	3	3	1
12(3)	3	3	1
14(3)	3	3	3
Recognition Rate	**4 out of 5 (80%)**	**3 out of 5 (60%)**	**02 out of 5 (60%)**

III. Conclusion

For the time being, more and more research is devoted to Interval Valued Observations than classical datasets. Different dimension reduction techniques are being explored and modified by researchers, which have traditionally been used for classical data. First part of this research includes the implementation of the vertices principal component analysis on different interval valued observations. To reduce the complexity, some authors have been suggested to work with midpoint[2,17,18] of each interval, thus converting the interval valued data into a classical data. But use of midpoints of the intervals ignores internal variations present in the data. Methods involving a range variable could not always distinguish between differing observations with similar range values Thus VPCA proved superior most especially with respect to computational complexity and optimum covering envelopes. In this work, different weight schemes depending on the vertices relative contribution on principal components are analysed to choose appropriate weights that will lead to better classification. Second part of this research involves distance matrix for interval valued data to classify testing data points. It is observed that, the recognition rate varies with different information of distance interval (like length of distance interval, upper limit, lower limit). For face data, length of distance interval gave 100% accuracy in recognizing testing images, whereas for blood pressure data, upper limit of distance interval showed 80% accuracy. It should be noted that, classifier for IVD is yet to develop. Idea of classifiers for classical data such as Support Vector Machine (SVM), K-NN can be extended to use them for interval Valued Observations to get better identification.

Acknowledgement

This research work is supported by the Bangladesh University Grant Commission under the UGC PhD fellowship, 2021.

References

1. Douzal-Chouakria1 A., L. Billard and E. Diday, 2011. Principal Component Analysis for Interval Valued Observations, *Statistical Analysis and Data Mining*, **4,** 228-246, hal- 00659996.
2. Billard L., A. Douzal-Chouakria1 and E. Diday, 2008. Symbolic principal component for interval-valued Observations, hal-00361053.
3. Denoeux T. and M. Masson, 2000. Multidimensional scaling of interval-valued dissimilarity data, *Pattern Recognition Letters*, **21(1)**, 83–92.
4. Ana B. Ramos-Guajardo and P. Grzegorzewski, 2016. Distance-based linear discriminant analysis for interval valued data, *Information Sciences*, **372**, 591–607.
5. Jahan, S. 2018. On Dimension Reduction Using Supervised Distance Preserving Projection for Face Recognition, *Universal Journal of Applied Mathematics*, **6(3),** 94-105, doi: 10.13189/ujam.2018. 060303.
6. Jahan, S. 2020. Supervised Distance Preserving Projection using Alternating Direction Method of Multipliers, *Journal of Industrial and Management Optimization, AIMS*, **16(4)**,1783-1799.
7. DâUrso P. and P. Giordani, 2004. A least squares approach to principal component analysis for interval valued data, *Chemometrics and Intelligent Laboratory Systems*, **70**, 179-192.
8. Wang H., R. Guan and J. Wu, 2012. CIPCA: Complete-Information-based Principal Component Analysis for interval-valued data, *Neurocomputing*, **86,** 158-169.
9. Qi, X. H. Guo, Z. Arthem and W.Wang, 2020. An Interval-Valued Data Classification Method Based on the Unified Representation Frame, *IEEE Access*, **8**, 17002-17012.
10. Groenen, P. J. F. S. Winsberg, O. RodrÃguez and E. Diday, 2006. I-Scal: Multidimensional scaling of interval dissimilarities, *Computational Statistics Data Analysis*, **51**, 360-378.

11. Mishra, S. U. Sarkar, S. Taraphder, S. Datta, D. Swain and R. Saikhom, 2017. Multivariate Statistical Data Analysis-Principal Component Analysis (PCA), *International Journal of Livestock Research*, **7(5)**, 60-78.

12. Halko, A. N. P. Martinsson, Y. Shkolnisky and M. Tygert, 2011. An Algorithm for the Principal Component Analysis of large Data sets, *SIAM Journal on Scientific Computing*, **33(5)**, 2580–2594.

13. Theodoridis S. and K. Koutroumbas, 2009. Pattern Recognition. *Elsevier Inc*.

14. Theodoridis S. and K. Koutroumbas, 2010. An Introduction to Pattern Recognition, A MATLAB approach, *Elsevier Inc.*

15. Imani M. and G. A. Montazer, 2019. A survey of emotion recognition methods with emphasis on E-Learning environments, *Journal of Network and Computer Applications*, **147**, 102423.

16. Rodriguez O., Edwin Diday, and Suzanne Winsberg, 2000, Generalization of the Principal Components Analysis to Histogram Data, in Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery, Lyon, France, 13–16 September 2000.

17. Garro J. A., O. R. Rodríguez, 2019, Optimized Dimensionality Reduction Methods for Interval-Valued Variables and Their Application to Facial Recognition. *Entropy*, **21**(10):1016.,