Malaysian Journal of Mathematical Sciences 14(2): 249–271 (2020)

MALAYSIAN JOURNAL OF MATHEMATICAL SCIENCES

Journal homepage: http://einspem.upm.edu.my/journal

Modified Statistical Approach for Data Preprocessing to Improve Heterogeneous Distance Functions

Dalatu, P. I. *[1, 3] and Midi, H. [1, 2]

[1]*Department of Mathematics, Faculty of Science, Universiti Putra Malaysia, Malaysia*
[2]*Institute of Mathematical Research , Universiti Putra Malaysia, Malaysia*
[3]*Department of Mathematics, Faculty of Science, Adamaw State University, Mubi-Nigeria*

E-mail: dalatup@gmail.com
* *Corresponding author*

Received: 3 January 2019
Accepted: 27 March 2020

ABSTRACT

Clustering is one of the most important techniques used in data mining. The major aim of clustering is to partition a set of data objects into clusters such that data objects in the same cluster are more similar to each other than those in the other clusters. We proposed a modified statistical approach for data preprocessing to improve heterogeneous distance functions from Heterogeneous Euclidean-Overlap Metric (HEOM) by replacing the range function which serves as a local normalization by interquartile range function, and the approach is called Interquartile Range- Heterogeneous Euclidean-Overlap Metric (IQR-HEOM). The proposed approach is used to overcome the weakness of using range function as local normalization in HEOM. However, using range function and

dividing it by range allows outliers to have big effect on the contribution of the attributes. In addition, cohesion measures how closely related objects are in a cluster. While, silhouette measures how distinct or well-separated a cluster is from other clusters. To evaluate the performance of the proposed approach, simulation study and real life data sets were considered. Therefore, comparing the performance of the proposed approach and the existing methods, it is evidently clear that the suggested approach outperformed the existing methods, even with the contamination of the data, still the proposed approach had shown better performance.

Keywords: K-Means, simulation, interquartile, heterogeneous and clustering.

1. Introduction

In data mining, clustering is one of the most important techniques. The clustering can be used in many fields such privacy preserving, information retrieval and text analysis in Zhang et al. (2010). The main aim of clustering is to partition a set of data objects into clusters such that data objects in the same cluster are more similar to each other than those in the other clusters. Normally, data sets to be mined contain both numeric and categorical attributes.

Hence, most of the existing algorithms are limited to one of the two data types as the K-Means, K-Mode , Fuzzy K-mode. Cao et al. (2012) proposed a new dissimilarity measure for the K-mode algorithm, and presented a method to simultaneously find initial center and the number of clusters for the categorical data.

Up to the present time, there is some work for dealing with mixed data. It is important to note that many distance functions have been proposed to handle the issue. Moreover, lots of these functions work well for numeric attributes but do not appropriately handle nominal attributes. When all attributes are nominal, the simplest distance function is the Overlap Metric (OM), which simply counts the number of different attribute values of each pair of instances and is widely used by instance-based learning inAha (1992) and locally weighted learning. However, OM is a little rough to measure the distance between each pair of instances, because it fails to make use of additional information provided by nominal attribute values that can aid in generalization. In order to find reasonable distance function between instances with nominal attributes only, the Value Difference Metric (VDM) was proposed by Stanfill and Waltz (1986). In VDM, there were some drawbacks and limitations, which made them to propose a novel distance function: Frequency Difference Metric (FDM).

The FDM uses the joint frequencies of class label and attributes values, instead of the conditional probabilities, to compute the distance between two instances. Moreover, they stated that their proposed method is very simple, even much simpler than the basic VDM. That is, they used the Manhattan distance between the joint frequency vectors of these two instances as their distance.

In recent times (ChitraDevi et al. (2012)), lots of researchers in the literature had made use of Heterogeneous Euclidean-Overlap Metric (HEOM) in different ways and areas. In order to evaluate and cope with heterogeneous data, to contain both nominal and ordinal attributes, the two approaches are usually

taken as: (1) The data is transformed into one of the two data types that complies with the distance measure used, and (2) the two types of distance measures are combined and handle the data separately in accordance with the data type.

However, both approaches are problematic when it comes to the interpretation of results. Therefore, they proposed novel distance in their study called Neighbourhood Counting Metric (NCM), to overcome the problem. The proposed method is derived from a probability function and it can handle both nominal and ordinal attributes in a conceptually uniform way. The proposed method has the following advantages; the new NCM is conceptually simple, it is straight forward to implement. It also has the added property that it is independent of the underlying analytical or reasoning task (example classification). The measure is clear and unambiguous meaning is defined by the number neighbourhood of a query that include or cover a given data point.

In Peng et al. (2015), stated three different methods in existence for managing heterogeneous data. Firstly, is to convert nominal attributes to integer through coding and then consider them as numerical attributes. Its major problem is instability as the performance is easily affected by the use of a coding mechanism. Secondly, the method is to discrete numerical attributes, and then treat them as nominal attributes, classification and regression tree (CART), and other methods. Generally, discretization causes loss of information. Thirdly, the method is to learn a distance, such as the value difference metric (VDM), heterogeneous value difference metric (HVDM), heterogeneous euclidean-overlap metric (HEOM) and other methods (Stanfill and Waltz (1986)). This type of method can be combined based on distance (example K-Nearest Neighbour). The overlap is a simple and effective method to use. However, it only determines whether nominal attributes are equal to one another, and does not fully exploit classification.

The VDM was introduced by Stanfill and Waltz (1986) to provide an appropriate distance function for nominal attributes. A simplified version of the VDM (without the weighting schemes was proposed). The HVDM was proposed based on the following reasons; the Euclidean distance function is inappropriate for nominal attributes, VDM is inappropriate for continuous attributes, neither is sufficient enough on its own for use on heterogeneous application, that is one with both nominal and continuous attributes.

The Windowed Value Difference Metric (WVDM) samples the value of $P_{a,x,c}$ at each value x occurring in the training set for each a, instead of the midpoint of each range. However, the discretized ranges are not even used by

WVDM on continuous attributes, except to determine an appropriate window width, w_a, which is same as the range width in Discretized Value Difference Metric (DVDM) and Interpolated Value Difference Metric (IVDM).

In ChitraDevi et al. (2012), they presented various distance functions that can be used to perform cluster based outliers detection in wireless sensor networks. They made use of the efficiency of those resulting clustering to detect outliers by calculating the false alarm rate and false positive rate; using data collected from Intel Berkeley Laboratory. From their results obtained, they claimed that all the three distances (Euclidean, Manhattan, and HEOM) can be applied to perform clustering. However, they added that Euclidean and Manhattan requires normalization to avoid deviation from dynamic ranges, whereas HEOM requires no normalization as it performs local normalization using range. They concluded that based on the data used, HEOM provides greater accuracy even for slight data variation.

In this paper, we propose a new method to enhance HEOM to overcome the weakness of the existing method, whereby HEOM needs no normalization it executes local normalization using range function in ChitraDevi et al. (2012). The procedure used in HEOM, by dividing it with range allows outliers to have big effect on the contribution of the attributes. They further recommended using interquartile range which is more robust to range against outliers in data preprocessing. Therefore, we proposed a method by replacing range with interquartile range in HEOM (ChitraDevi et al. (2012)).

This paper is organized as follows: Section 2, presents materials and methods, comprises of conventional and proposed methods. Section 3, reviews and evaluates the simulation study. Section 4, gives results and discussion. Section 5, finally, gives some concluding remarks.

2. Materials and Methods

2.1 Conventional Methods

The two distance functions, Euclidean and Manhattan which handles only continuous input attributes and also applied to perform clustering as homogeneous distance functions. The HEOM; handles both continuous and nominal attributes with overlap metric for nominal attributes and normalize Euclidean distance for linear attributes.

2.1.1 Euclidean Distance Function

The Euclidean distance was first applied in clustering analysis. The Euclidean distance is calculated as in Lloyd (1982):

$$d(x_i, x_j) = \sqrt{\sum_{i=1}^{n}(x_i - x_j)^2} \tag{1}$$

This distance measure has the appealing property, the $d(x_i, x_j)$ can be interpreted as physical distance between p-dimensional points $x_i' = x_{i1}, x_{i2}, ..., x_{ip}$ and $x_j' = (x_{j1}, x_{j2}, ..., x_{jp})$ in Euclidean space.

2.1.2 Manhattan Distance Function

The Manhattan distance was first used in clustering analysis. It sums the difference between their components. Manhattan distance is computed as follows in Lloyd (1982):

$$d_{man}(x_i, x_j) = \sum_{i=1}^{n} \mid x_i - x_j \mid . \tag{2}$$

where $i, j = 1, 2, \ldots, n$.

2.1.3 Heterogeneous Euclidean-Overlap Metric (HEOM)

The Heterogeneous Euclidean-Overlap Metric; handles both continuous and nominal attributes with overlap metric for nominal attributes and normalize Euclidean distance for linear attributes ChitraDevi et al. (2012). In order to tackle the issues of applications with both continuous and nominal attributes is to apply a heterogeneous distance function that uses dissimilar attributes distance functions on diverse categories of attributes, the unique technique that has been used is overlap metric for combined nominal attributes and normalized Euclidean distance for linear attributes. Therefore, da (the function defines the distance between two values x_i and x_j of an attribute a as denoted in Equation 5) yields a value that normally in the interval $0, \ldots, 1$ (range $[0; 1]$), whether the attributes is nominal or continuous. Heterogeneous Euclidean-Overlap Metric (HEOM) (maybe heterogeneous) is computed as in ChitraDevi et al. (2012),

this function describes the distance between two values x_i and x_j of a given attribute a as:

$$d_a(x_i, x_j) = \begin{cases} 1 & \text{if } x_i \text{ or } x_j \text{ is unknown, else} \\ overlap(x_i, x_j) & \text{if } a \text{ is nominal, else} \\ rn - diff_a(x_i, x_j) & \text{if } a \text{ is continuous} \end{cases}$$

Unknown attribute values are controlled by returning an attribute distance 1 (i.e., maximal distance) if either of the attribute values is unknown. The function overlap (each data points is mapped to a small set of features to different clusters) and the range-normalized difference $rn - diff$ are defined respectively as:

$$overlap(x_i, x_j) = \begin{cases} 0, & \text{if} \quad x_i = x_j \\ 1, & otherwise \end{cases}$$

$$rn - diff_a(x_i, x_j) = \frac{|x_i - x_j|}{range_a} \tag{3}$$

The value $range_a$ is used to normalize the attribute, and is given as:

$$range_a = max_a - min_a \tag{4}$$

where max_a and min_a are the maximum and minimum values, respectively, observed in the training set for attribute a. The data for the above definition yields a value which is in the range $0, \ldots, 1$, whether the attribute is nominal or linear. Therefore, the general distance between two input vectors x_i and x_j is given by the Heterogeneous Euclidean-Overlap Metric function (HEOM) as in ChitraDevi et al. (2012):

$$HEOM(x_i, x_j) = \sqrt{\sum_{i=1}^{n} d_a(x_i - x_j)^2} \tag{5}$$

The distance function eliminates the special effects of the random ordering of nominal values, but its excessively naive method to handling nominal attributes fails to make use of added evidence provided by nominal attribute values that assist in simplification.

2.2 IQR-Heterogeneous Euclidean-Overlap Metric (HEOM)

In this section we will discuss the proposed heterogeneous distance functions. The proposed method is based on the HEOM (refer to ChitraDevi et al. (2012)). Therefore, ChitraDevi et al. (2012) claimed that by using range function as a local normalization in HEOM, and comparing the method to two conventional distance functions of Euclidean and Manhattan. They concluded that HEOM provides greater accuracy in performance. They based their argument by using cluster outliers detection in wireless sensor networks. The HEOM method was criticized far back, that the distance function removes the effect of the arbitrary ordering of nominal values. This means that this type of approach is too simple in handling nominal attributes which fails to make use of additional information given in nominal attribute values that can assist in generalization. Recently, the proposed outliers detection model (ODM), is built by using the K-Means clustering algorithm. They applied interquartile range (IQR) as data preprocessing instead of range values. They further added that the procedure used in HEOM, by dividing it with range allows outliers to have intense influence on the contribution of the attributes. They believed that, normal objects lie between the lower and upper extremes. They recommended using interquartile range which is more robust to range against outliers in data preprocessing.

For example, if variables happens to have values in the range of $0, \ldots, 10$, in almost every case but with abnormal (and probable error) value of 50. Hence, dividing all the values by the range would nearly give the results in all less than 0.2. Generally, a silhouette measure of less than 0.20 shows a poor quality result, a measure between 0.20 and 0.50 indicates a fair result, while values of more than 0.50 shows good results.

Therefore, to increase the accuracy and break down points of the proposed method, the interquartile range (IQR) with a break down point of 25%, and has little resistance to outliers due to its focus on the center of the distribution (Rousseeuw and Hubert (2011)) are used. Using IQR that is less sensitive to outliers to range values. However, the weakness of HEOM according to ChitraDevi et al. (2012), is going to be dealt with by this proposed method. The proposed method is summarized as follows:

The function *interquartilerange* is computed by using the ideas from ChitraDevi et al. (2012):

$$iqrn - diff_a(x_i, x_j) = \frac{\mid x_i - x_j \mid}{iqr_a} \tag{6}$$

where

$$iqr_a = Q3_a - Q1_a \tag{7}$$

IQR-Heterogeneous Euclidean-Overlap Metric ($IQR - HEOM$) is defined as,

$$IQR - HEOM(x_i, x_j) = \sqrt{\sum_{i=1}^{n} d_a(x_i - x_j)^2}, \tag{8}$$

where, d_a is being calculated for $IQR - HEOM$ based on conditions in Equations 6 and 7.

2.3 K-Means Clustering Algorithm

The K-means clustering algorithm consist of four steps, which are iterated until convergence Mohamad and Usman (2013). The iteration will stop when the clusters produced are stable, which means there are no more movement of objects crossing any group. The K-Means algorithms are enlisted by Lloyd (1982) are as follows.

The K-Means clustering algorithm is broadly used in data mining to group data with similar features together. Assumed n data points, the algorithm distributes them into k groups in three stages: (1) evaluate the distances between data points with each of k clusters and assign the data to the nearest cluster, (2) calculate the center of each cluster, (3) update the clusters repeatedly, until the k clusters change no more or stabilized. The aim of the algorithm is to minimize the cost function. The cost function (Khan (2012)),

$$J = \sum_{i=1}^{n} \sum_{j=1}^{k} \parallel x_i - c_j \parallel^2 \tag{9}$$

where, $\parallel x_i - c_j \parallel^2$ is an arbitrary distance measure between a data point x_i and the cluster center c_j is assigned to the distance of the n data points from their individual centers.

The algorithm consists of the following steps (Khan (2012)):

1. Initialize the centers at random;
2. Assign data points to their respective clusters having the nearest mean;
3. Compute new centers as means of the clusters assigned in step 2;
4. Repeat steps 2 and 3 until no change is made in the centers.

2.4 Two Internal Validity Measures

2.4.1 Silhouette Coefficients

The Silhouette coefficients contrast the average distance to elements in the same with the average distance to elements in other clusters. Based on ChitraDevi et al. (2012), the Silhouette is computed as follows:

$$s(i) = \frac{(b(i) - a(i))}{max\{a(i), b(i)\}} \tag{10}$$

where, i represents any object in the data set,
$a(i)$ is the average distance (dissimilarity) of i to other objects in the same cluster A, and
$b(i)$ is the minimum (lowest) (average distance of i to all objects in the neighboring clusters B). The dissimilarity is computed using distance measures.

The value of Silhouette coefficient usually varies from -1 to $+1$, and the cluster arrangement is extremely efficient when the value is nearer to $+1$.

2.4.2 Cohesion values

The Cohesion is computed as in ChitraDevi et al. (2012):

$$Cohesion(c_i) = SSE = \sum_{i=1}^{n} \sum_{j=1}^{k} (x_i - m_j)^2, \tag{11}$$

where, m_j represent each means for the clusters and $j = 1, \ldots, k$ (total number of clusters with their various means are being computed). Cohesion is also called as Sum of Squared Errors (SSE) or Within Sum of Squares.

Furthermore, it is important to mention that our experiment are going to be evaluated in two different ways, (i) normalized data, and (ii) NonNormalized data; the following are carried out to compare the performance of the proposed method and the existing methods. Firstly, we perform the K-Mean clustering algorithm based on conventional and proposed methods. Secondly, the silhouette coefficients and cohesion values are compared under each distance functions to evaluate the performance of proposed method, ChitraDevi et al. (2012):

$$Cohesion(c_i) = SSE = \sum_{i=1}^{n} \sum_{j=1}^{k} (x_i - m_j)^2, \tag{12}$$

where, m_j represent each means for the clusters and $j = 1, \ldots, k$ (total number of clusters with their various means are being computed). Cohesion is also called as Sum of Squared Errors (SSE) or Within Sum of Squares.

Furthermore, it is important to mention that our experiment are going to be evaluated in two different ways, (i) normalized data, and (ii) NonNormalized data; the following are carried out to compare the performance of the proposed method and the existing methods. Firstly, we perform the K-Mean clustering algorithm based on conventional and proposed methods. Secondly, the silhouette coefficients and cohesion values are compared under each distance functions to evaluate the performance of proposed method.

3. Simulation Study

In this section, Monte Carlo simulation study is presented to compare the performance of some existing methods such as Euclidean distance, Manhattan distance and specifically the Heterogeneous Euclidean-Overlap Metric (HEOM) (ChitraDevi et al. (2012)), with our proposed method Interquartile Range-Heterogeneous Euclidean-Overlap Metric ($IQR - HEOM$).

The simulation is conducted on three examples comprising two and four variables each in this study. The example is popular and often being used by many researchers who study the stability of clusters by applying cohesion

values and silhouette coefficients (see, used by de Amorim and Hennig (2015)).

Following ChitraDevi et al. (2012); two variables (x_1, x_2) and four variables (x_1, x_2, x_3, x_4) are generated with sample size $(n = 50, 100, 160)$ each, measures are calculated based on 1000 replications, and then using Equations 1, 2, 5, 8, 9, 10, and 11 on the data for evaluation. These experiments involve different width clusters as; 0.01, 0.02, 0.03, 0.05, 0.07, 0.09, 0.12, 0.15, 0.18, 0.21, 0.24, 0.27, and 0.30. It consists of two sections, Non-normalized data and Normalized data.

The performance of our proposed method is calculated based on the parameters for cohesion values as maximum > 1 (minimum < 1) and silhouette coefficients as maximum $= 1$ (minimum $= 0$).

The first experiment:
Each of the explanatory variables (x_1, x_2) and (x_1, x_2, x_3, x_4) are generated from uniform distribution with parameters $[-10, 10]$, and $n = 50$ sample size each. The variables are estimated using cohesion values and silhouette coefficients parameters.

The second experiment:
Each of the explanatory variables (x_1, x_2) and (x_1, x_2, x_3, x_4) are generated from uniform distribution with parameters $[-10, 10]$, and $n = 100$ sample size each. The variables are estimated using cohesion values and silhouette coefficients parameters.

The third experiment:
Each of the explanatory variables (x_1, x_2) and (x_1, x_2, x_3, x_4) are generated from uniform distribution with parameters $[-10, 10]$, and $n = 160$ sample size each. The variables are estimated using cohesion values and silhouette coefficients parameters.

After data is put in place as transformed and also untransformed data respectively; the Euclidean distance, Manhattan distance, $HEOM$, and $IQR-HEOM$ are applied to the transformed and untransformed data. For each of the experimental runs, there are 1000 replications. Then, the cohesion values and silhouette coefficients computed under each distance functions.

4. Results and Discussion

Table 1, 2, 3, 4, 5, and 6 present the average values of 1000 replications of the cohesion values and silhouette coefficients. Tables 1, 2, 3, 4, 5, and 6 show the average cohesion values and silhouette coefficients for various width clusters. This results, signify that the non normalized simulated data under all the methods had fair performance, unlike the normalized data. This indicates that the performance of the proposed method is even fairly good by attaining up to 0.27 width cluster of $n = 160$ sample size having four variables in non normalized data and had performed excellent well under normalized data by exhausting all the width clusters successful. Therefore, it is evidently clear that the proposed method performance is much better based on the evaluation carried on the simulated data compared to the existing methods.

Table 7 presents three different sample size ($n = 50, 100, 160$), with two (x_1, x_2) and four (x_1, x_2, x_3, x_4) attribute variables each, when the data is contaminated at 5% and 10%. Following the same simulation study as preceding section, each of the variable is generated from uniform distribution with range $[-10, 10]$ and contaminated data is generated from uniform distribution with range $[15, 16]$. The Euclidean distance, Manhattan distance, $HEOM$ and $IQR - HEOM$ were then applied to the data.

The average cohesion values, silhouette coefficients, and computational timing tests based on the 1000 simulation runs are presented in Table 7. From the table it is clearly seen that, the proposed method has made tremendous achievement, despite that the data is contaminated. However, the existing methods compared to the proposed method had performed fairly well in clean data.

Table 1: Average Cohesion and Silhouette for various Width Clusters, (n = 50 (x_1, x_2))

Normalized(Classic)				Normalized(NADS)		
		Parameters			Parameter	
Width	Methods Dist.	Coh. Max> 1(Min< 1)	Silh. Max=1(Min=0)	Methods Dist.	Coh. Max> 1(Min< 1)	Silh. Max=1(Min=0)
0.01	Euc.	2.59E-04	1	Euc.	2.19E-05	0.99164
0.01	Manh.	3.62E-05	1	Manh.	2.19E-05	1
0.01	HEOM	3.81E-04	1	HEOM	2.19E-05	1
0.01	**IQR-HEOM**	**3.81E-06**	**1**	**IQR-HEOM**	**2.69E-06**	**1**
0.02	Euc.	5.38E-04	1	Euc.	7.09E-06	0.98591
0.02	Manh.	5.38E-04	1	Manh.	7.09E-06	0.98591
0.02	HEOM	0.00031	0.98618	HEOM	7.09E-06	0.98591
0.02	**IQR-HEOM**	**7.39E-05**	**1**	**IQR-HEOM**	**8.53E-07**	**1**
0.03	Euc.	6.38E-04	1	Euc.	7.09E-06	0.98591
0.03	Manh.	6.38E-04	1	Manh.	7.09E-06	0.98581
0.03	HEOM	0.00031	0.98618	HEOM	7.09E-06	0.98591
0.03	**IQR-HEOM**	**2.57E-07**	**1**	**IQR-HEOM**	**4.95E-08**	**1**
0.05	Euc.	0.00038	1	Euc.	0.00025	0.91372
0.05	Manh.	0.00371	1	Manh.	0.00015	0.93960
0.05	HEOM	0.00374	0.96947	HEOM	8.32E-05	0.96932
0.05	**IQR-HEOM**	**5.71E-05**	**1**	**IQR-HEOM**	**2.59E-08**	**1**
0.07	Euc.	9.71E-05	0.99190	Euc.	0.00097	1
0.07	Manh.	9.71E-06	1	Manh.	0.00025	0.91372
0.07	HEOM	0.01417	0.00403	HEOM	0.00030	0.90335
0.07	**IQR-HEOM**	**7.39E-06**	**1**	**IQR-HEOM**	**7.09E-06**	**1**
0.09	Euc.	0.29676	0.60749	Euc.	0.25566	0.52493
0.09	Manh.	0.12175	0.74468	Manh.	0.14686	0.39603
0.09	HEOM	7.14679	0.38833	HEOM	0.14686	0.39603
0.09	**IQR-HEOM**	**0.15280**	**1**	**IQR-HEOM**	**0.00237**	**1**
0.12	Euc.	0.00031	0.98618	Euc.	0.00612	0.60818
0.12	Manh.	0.00031	0.98618	Manh.	0.00252	0.74517
0.12	HEOM	0.10103	0.75346	HEOM	0.00211	0.75289
0.12	**IQR-HEOM**	**7.78E-04**	**1**	**IQR-HEOM**	**4.82E-07**	**1**
0.15	Euc.	0.00031	0.98618	Euc.	0.00625	0.54005
0.15	Manh.	0.00031	0.98618	Manh.	0.00470	0.60847
0.15	HEOM	0.20166	0.64425	HEOM	0.00417	0.64449
0.15	**IQR-HEOM**	**3.61E-04**	**1**	**IQR-HEOM**	**2.89E-08**	**1**
0.18	Euc.	0.00031	0.98618	Euc.	0.00151	0.45341
0.18	Manh.	0.00031	0.98618	Manh.	0.00815	0.57351
0.18	HEOM	0.03674	0.58727	HEOM	0.00778	0.58465
0.18	**IQR-HEOM**	**0.013108**	**1**	**IQR-HEOM**	**7.65E-08**	**1**
0.21	Euc.	0.00031	0.98618	Euc.	0.01775	0.0.52011
0.21	SManh.	0.00031	0.98618	Manh.	0.00790	0.57854
0.21	HEOM	0.37184	0.58581	HEOM	0.00764	0.58442
0.21	**IQR-HEOM**	**0.018562**	**0.95468**	**IQR-HEOM**	**3.59E-07**	**1**
0.24	Euc.	0.00251	0.97805	Euc.	0.02187	0.48116
0.24	Manh.	0.00031	0.98618	Manh.	0.00144	0.50832
0.24	HEOM	0.46667	0.55411	HEOM	0.00960	0.55483
0.24	**IQR-HEOM**	**0.02487**	**0.93618**	**IQR-HEOM**	**7.09E-08**	**1**
0.27	Euc.	0.00374	0.96947	Euc.	0.03030	0.44641
0.27	Manh.	0.00031	0.98618	Manh.	0.01823	0.50613
0.27	HEOM	0.59173	0.51784	HEOM	0.01233	0.51657
0.27	**IQR-HEOM**	**0.05011**	**0.97690**	**IQR-HEOM**	**7.09E-08**	**1**
0.30	Euc.	0.00515	0.95496	Euc.	0.03269	0.45973
0.30	Manh.	0.00374	0.96947	Manh.	0.02377	0.43992
0.30	HEOM	1.06160	0.48058	HEOM	0.01906	0.48078
0.30	**IQR-HEOM**	**0.11682**	**0.98124**	**IQR-HEOM**	**7.09E-08**	**1**

Table 2: Average Cohesion and Silhouette for various Width Clusters, (n = 50, (x_1, x_2, x_3, x_4))

Normalized(Classic)				Normalized(NADS)		
		Parameters			Parameter	
Width	Methods Dist.	Coh. Max> 1(Min< 1)	Silh. Max=1(Min=0)	Methods Dist.	Coh. Max> 1(Min< 1)	Silh. Max=1(Min=0)
0.01	Euc.	1.59E-04	1	Euc.	1.19E-05	0.98164
0.01	Manh.	2.62E-05	1	Manh.	1.19E-05	1
0.01	HEOM	2.81E-04	1	HEOM	1.19E-05	1
0.01	**IQR-HEOM**	**2.81E-06**	**1**	**IQR-HEOM**	**1.69E-06**	**1**
0.02	Euc.	4.38E-04	1	Euc.	6.09E-06	0.98591
0.02	Manh.	1.38E-04	1	Manh.	6.09E-06	0.98482
0.02	HEOM	0.00032	0.97722	HEOM	6.09E-06	0.98428
0.02	**IQR-HEOM**	**6.39E-05**	**1**	**IQR-HEOM**	**7.53E-07**	**1**
0.03	Euc.	5.38E-04	1	Euc.	6.09E-06	0.98482
0.03	Manh.	5.38E-04	1	Manh.	7.09E-06	0.98482
0.03	HEOM	0.00031	0.98527	HEOM	6.09E-06	0.98482
0.03	**IQR-HEOM**	**1.57E-07**	**1**	**IQR-HEOM**	**3.95E-08**	**1**
0.05	Euc.	0.00038	1	Euc.	0.00025	0.91483
0.05	Manh.	0.00371	1	Manh.	0.00015	0.93851
0.05	HEOM	0.00374	0.96836	HEOM	7.32E-05	0.96821
0.05	**IQR-HEOM**	**4.71E-05**	**1**	**IQR-HEOM**	**1.59E-08**	**1**
0.07	Euc.	7.71E-05	0.99392	Euc.	0.00097	1
0.07	Manh.	8.71E-06	1	Manh.	0.00025	0.91483
0.07	HEOM	0.01417	0.00403	HEOM	0.00030	0.90446
0.07	**IQR-HEOM**	**6.39E-06**	**1**	**IQR-HEOM**	**6.09E-06**	**1**
0.09	Euc.	0.29676	0.60749	Euc.	0.25566	0.52493
0.09	Manh.	0.12175	0.74468	Manh.	0.14686	0.39603
0.09	HEOM	6.14679	0.38833	HEOM	0.14686	0.39603
0.09	**IQR-HEOM**	**0.15280**	**1**	**IQR-HEOM**	**0.00237**	**1**
0.12	Euc.	0.00031	0.98729	Euc.	0.00612	0.60818
0.12	Manh.	0.00031	0.98729	Manh.	0.00252	0.74629
0.12	HEOM	0.10103	0.75457	HEOM	0.00211	0.75392
0.12	**IQR-HEOM**	**6.78E-04**	**1**	**IQR-HEOM**	**3.82E-07**	**1**
0.15	Euc.	0.00031	0.98728	Euc.	0.00625	0.54215
0.15	Manh.	0.00031	0.98729	Manh.	0.00470	0.60956
0.15	HEOM	0.21772	0.64536	HEOM	0.00417	0.64558
0.15	**IQR-HEOM**	**2.61E-04**	**1**	**IQR-HEOM**	**1.89E-08**	**1**
0.18	Euc.	0.00031	0.98727	Euc.	0.00151	0.45452
0.18	Manh.	0.00031	0.98729	Manh.	0.00815	0.57462
0.18	HEOM	0.03674	0.58838	HEOM	0.00778	0.58576
0.18	**IQR-HEOM**	**0.01310**	**1**	**IQR-HEOM**	**6.65E-08**	**1**
0.21	Euc.	0.00031	0.98729	Euc.	0.01775	0.52122
0.21	SManh.	0.00031	0.98729	Manh.	0.00790	0.57743
0.21	HEOM	0.37184	0.58692	HEOM	0.00764	0.58553
0.21	**IQR-HEOM**	**0.01856**	**0.95579**	**IQR-HEOM**	**2.59E-07**	**1**
0.24	Euc.	0.00251	0.97916	Euc.	0.02187	0.48227
0.24	Manh.	0.00031	0.98729	Manh.	0.00144	0.50943
0.24	HEOM	0.46778	0.55522	HEOM	0.00960	0.55594
0.24	**IQR-HEOM**	**0.02487**	**0.93729**	**IQR-HEOM**	**6.09E-08**	**1**
0.27	Euc.	0.00374	0.96836	Euc.	0.03030	0.44752
0.27	Manh.	0.00031	0.98729	Manh.	0.01823	0.50724
0.27	HEOM	0.59294	0.51673	HEOM	0.01233	0.51768
0.27	**IQR-HEOM**	**0.05011**	**0.97581**	**IQR-HEOM**	**6.09E-08**	**1**
0.30	Euc.	0.00515	0.95385	Euc.	0.03269	0.45862
0.30	Manh.	0.00374	0.96958	Manh.	0.02377	0.43883
0.30	HEOM	1.06163	0.48169	HEOM	0.01906	0.48189
0.30	**IQR-HEOM**	**0.11682**	**0.98235**	**IQR-HEOM**	**6.09E-08**	**1**

Table 3: Average Cohesion and Silhouette for various Width Clusters, (n = 100, (x_1, x_2))

Normalized(Classic)				Normalized(NADS)		
		Parameters			Parameter	
Width	Methods Dist.	Coh. Max> 1(Min< 1)	Silh. Max=1(Min=0)	Methods Dist.	Coh. Max> 1(Min< 1)	Silh. Max=1(Min=0)
0.01	Euc.	4.59E-04	1	Euc.	4.19E-05	0.99386
0.01	Manh.	1.62E-05	1	Manh.	4.19E-05	1
0.01	HEOM	1.81E-04	1	HEOM	4.19E-05	1
0.01	**IQR-HEOM**	**1.81E-06**	**1**	**IQR-HEOM**	**4.69E-06**	**1**
0.02	Euc.	3.38E-04	1	Euc.	5.09E-06	0.98773
0.02	Manh.	3.38E-04	1	Manh.	5.09E-06	0.98772
0.02	HEOM	0.00031	0.98436	HEOM	5.09E-06	0.98773
0.02	**IQR-HEOM**	**5.39E-05**	**1**	**IQR-HEOM**	**6.53E-07**	**1**
0.03	Euc.	4.38E-04	1	Euc.	5.09E-06	0.98772
0.03	Manh.	4.38E-04	1	Manh.	5.09E-06	0.98772
0.03	HEOM	0.00031	0.98772	HEOM	5.09E-06	0.98772
0.03	**IQR-HEOM**	**4.57E-07**	**1**	**IQR-HEOM**	**2.95E-08**	**1**
0.05	Euc.	0.00038	1	Euc.	0.00025	0.93594
0.05	Manh.	0.00371	1	Manh.	0.00015	0.93742
0.05	HEOM	0.00374	0.96769	HEOM	6.32E-05	0.96754
0.05	**IQR-HEOM**	**3.71E-05**	**1**	**IQR-HEOM**	**4.59E-08**	**1**
0.07	Euc.	7.71E-05	0.99372	Euc.	0.00097	1
0.07	Manh.	7.71E-06	1	Manh.	0.00025	0.93594
0.07	HEOM	0.01417	0.00403	HEOM	0.00030	0.92557
0.07	**IQR-HEOM**	**5.39E-06**	**1**	**IQR-HEOM**	**5.09E-06**	**1**
0.09	Euc.	0.29898	0.62967	Euc.	0.25788	0.52675
0.09	Manh.	0.12397	0.74686	Manh.	0.14868	0.39825
0.09	HEOM	7.14897	0.38655	HEOM	0.14864	0.39825
0.09	**IQR-HEOM**	**0.15462**	**1**	**IQR-HEOM**	**0.00237**	**1**
0.12	Euc.	0.00031	0.98836	Euc.	0.00612	0.60636
0.12	Manh.	0.00031	0.98826	Manh.	0.00252	0.74739
0.12	HEOM	0.10103	0.75568	HEOM	0.00211	0.75467
0.12	**IQR-HEOM**	**5.78E-04**	**1**	**IQR-HEOM**	**4.82E-07**	**1**
0.15	Euc.	0.00031	0.98836	Euc.	0.00625	0.54227
0.15	Manh.	0.00031	0.98436	Manh.	0.00470	0.60625
0.15	HEOM	0.20344	0.64203	HEOM	0.00417	0.64227
0.15	**IQR-HEOM**	**1.61E-04**	**1**	**IQR-HEOM**	**4.89E-08**	**1**
0.18	Euc.	0.00031	0.98436	Euc.	0.00151	0.45123
0.18	Manh.	0.00031	0.98436	Manh.	0.00815	0.57133
0.18	HEOM	0.03674	0.58705	HEOM	0.00778	0.58243
0.18	**IQR-HEOM**	**0.01310**	**1**	**IQR-HEOM**	**5.65E-08**	**1**
0.21	Euc.	0.00031	0.98416	Euc.	0.01775	0.0.52221
0.21	SManh.	0.00031	0.98436	Manh.	0.00790	0.57632
0.21	HEOM	0.37362	0.58363	HEOM	0.00764	0.58220
0.21	**IQR-HEOM**	**0.01856**	**1**	**IQR-HEOM**	**2.59E-07**	**1**
0.24	Euc.	0.00251	0.97623	Euc.	0.02187	0.48224
0.24	Manh.	0.00031	0.98436	Manh.	0.00144	0.508610
0.24	HEOM	0.42225	0.5323	HEOM	0.00960	0.5326
0.24	**IQR-HEOM**	**0.02487**	**1**	**IQR-HEOM**	**5.09E-08**	**1**
0.27	Euc.	0.00374	0.94723	Euc.	0.03030	0.44423
0.27	Manh.	0.00031	0.96436	Manh.	0.01823	0.50431
0.27	HEOM	0.59251	0.51562	HEOM	0.01233	0.51435
0.27	**IQR-HEOM**	**0.05011**	**0.97872**	**IQR-HEOM**	**5.09E-08**	**1**
0.30	Euc.	0.00515	0.95274	Euc.	0.03269	0.45752
0.30	Manh.	0.00374	0.96725	Manh.	0.02377	0.43770
0.30	HEOM	1.06160	0.48236	HEOM	0.01906	0.48256
0.30	**IQR-HEOM**	**0.116825**	**0.98246**	**IQR-HEOM**	**5.09E-08**	**1**

Table 4: Average Cohesion and Silhouette for various Width Clusters, (n = 100, (x_1, x_2, x_3, x_4))

Normalized(Classic)				Normalized(NADS)		
		Parameters			Parameter	
Width	Methods Dist.	Coh. Max> 1(Min< 1)	Silh. Max=1(Min=0)	Methods Dist.	Coh. Max> 1(Min< 1)	Silh. Max=1(Min=0)
0.01	Euc.	5.59E-04	1	Euc.	5.19E-05	0.96431
0.01	Manh.	6.62E-05	1	Manh.	5.19E-05	1
0.01	HEOM	6.81E-04	1	HEOM	5.19E-05	1
0.01	**IQR-HEOM**	**6.81E-06**	**1**	**IQR-HEOM**	**5.69E-06**	**1**
0.02	Euc.	8.38E-04	1	Euc.	4.09E-06	0.98273
0.02	Manh.	8.38E-04	1	Manh.	4.09E-06	0.98273
0.02	HEOM	0.00031	0.98345	HEOM	4.09E-06	0.98374
0.02	**IQR-HEOM**	**4.39E-05**	**1**	**IQR-HEOM**	**5.53E-07**	**1**
0.03	Euc.	3.38E-04	1	Euc.	4.09E-06	0.98274
0.03	Manh.	3.38E-04	1	Manh.	4.09E-06	0.98254
0.03	HEOM	0.00031	0.98345	HEOM	4.09E-06	0.98274
0.03	**IQR-HEOM**	**5.57E-07**	**1**	**IQR-HEOM**	**4.95E-08**	**1**
0.05	Euc.	0.00038	1	Euc.	0.00025	0.91705
0.05	Manh.	0.00371	1	Manh.	0.00015	0.93733
0.05	HEOM	0.00374	0.96714	HEOM	5.32E-05	0.96705
0.05	**IQR-HEOM**	**8.71E-05**	**1**	**IQR-HEOM**	**2.59E-08**	**1**
0.07	Euc.	6.71E-05	0.99473	Euc.	0.00097	1
0.07	Manh.	6.71E-06	1	Manh.	0.00025	0.91045
0.07	HEOM	0.01417	0.00403	HEOM	0.00030	0.90002
0.07	**IQR-HEOM**	**4.39E-06**	**1**	**IQR-HEOM**	**4.09E-06**	**1**
0.09	Euc.	0.29343	0.60316	Euc.	0.25566	0.52160
0.09	Manh.	0.12142	0.74135	Manh.	0.14353	0.39330
0.09	HEOM	4.14346	0.38500	HEOM	0.14353	0.39334
0.09	**IQR-HEOM**	**0.15280**	**1**	**IQR-HEOM**	**0.00237**	**1**
0.12	Euc.	0.00031	0.98345	Euc.	0.00612	0.60545
0.12	Manh.	0.00031	0.98345	Manh.	0.00252	0.74244
0.12	HEOM	0.10103	0.75013	HEOM	0.00211	0.75556
0.12	**IQR-HEOM**	**4.78E-04**	**1**	**IQR-HEOM**	**1.82E-07**	**1**
0.15	Euc.	0.00031	0.98345	Euc.	0.00625	0.54332
0.15	Manh.	0.00031	0.98345	Manh.	0.00470	0.60514
0.15	HEOM	0.20433	0.64152	HEOM	0.00417	0.64116
0.15	**IQR-HEOM**	**6.61E-04**	**1**	**IQR-HEOM**	**5.89E-08**	**1**
0.18	Euc.	0.00031	0.98345	Euc.	0.00151	0.45014
0.18	Manh.	0.00031	0.98345	Manh.	0.00815	0.57024
0.18	HEOM	0.03674	0.58454	HEOM	0.00778	0.58132
0.18	**IQR-HEOM**	**0.01310**	**1**	**IQR-HEOM**	**4.65E-08**	**1**
0.21	Euc.	0.00031	0.98345	Euc.	0.01775	0.0.52344
0.21	SManh.	0.00031	0.98345	Manh.	0.00790	0.57521
0.21	HEOM	0.37451	0.58254	HEOM	0.00764	0.58115
0.21	**IQR-HEOM**	**0.01856**	**1**	**IQR-HEOM**	**6.59E-07**	**1**
0.24	Euc.	0.00251	0.97432	Euc.	0.02187	0.48443
0.24	Manh.	0.00031	0.98345	Manh.	0.00144	0.50505
0.24	HEOM	0.46224	0.55144	HEOM	0.00960	0.55150
0.24	**IQR-HEOM**	**0.02487**	**1**	**IQR-HEOM**	**4.09E-08**	**1**
0.27	Euc.	0.00374	0.96614	Euc.	0.03030	0.44314
0.27	Manh.	0.00031	0.98345	Manh.	0.01823	0.50340
0.27	HEOM	0.59440	0.51451	HEOM	0.01233	0.51324
0.27	**IQR-HEOM**	**0.05011**	**0.97363**	**IQR-HEOM**	**4.09E-08**	**1**
0.30	Euc.	0.00515	0.95163	Euc.	0.03269	0.45640
0.30	Manh.	0.00374	0.96614	Manh.	0.02377	0.43665
0.30	HEOM	1.06160	0.48325	HEOM	0.01906	0.48345
0.30	**IQR-HEOM**	**0.11682**	**0.98451**	**IQR-HEOM**	**4.09E-08**	**1**

Table 5: Average Cohesion and Silhouette for various Width Clusters, (n = 160, (x_1, x_2))

Normalized(Classic)				Normalized(NADS)		
		Parameters			Parameter	
Width	Methods Dist.	Coh. Max> 1(Min< 1)	Silh. Max=1(Min=0)	Methods Dist.	Coh. Max> 1(Min< 1)	Silh. Max=1(Min=0)
0.01	Euc.	6.59E-04	1	Euc.	6.19E-05	0.99528
0.01	Manh.	7.62E-05	1	Manh.	6.19E-05	1
0.01	HEOM	7.81E-04	1	HEOM	2.19E-05	1
0.01	**IQR-HEOM**	**7.81E-06**	**1**	**IQR-HEOM**	**6.69E-06**	**1**
0.02	Euc.	9.38E-04	1	Euc.	3.09E-06	0.98155
0.02	Manh.	9.38E-04	1	Manh.	3.09E-06	0.98264
0.02	HEOM	0.00031	0.98244	HEOM	3.09E-06	0.98155
0.02	**IQR-HEOM**	**3.39E-05**	**1**	**IQR-HEOM**	**5.53E-07**	**1**
0.03	Euc.	2.38E-04	1	Euc.	3.09E-06	0.98155
0.03	Manh.	2.38E-04	1	Manh.	3.09E-06	0.98145
0.03	HEOM	0.00031	0.98345	HEOM	3.09E-06	0.98154
0.03	**IQR-HEOM**	**6.57E-07**	**1**	**IQR-HEOM**	**8.95E-08**	**1**
0.05	Euc.	0.00038	1	Euc.	0.00025	0.91745
0.05	Manh.	0.00371	1	Manh.	0.00015	0.93524
0.05	HEOM	0.00374	0.96513	HEOM	4.32E-05	0.96573
0.05	**IQR-HEOM**	**1.71E-05**	**1**	**IQR-HEOM**	**6.59E-08**	**1**
0.07	Euc.	5.71E-05	0.99555	Euc.	0.00097	1
0.07	Manh.	5.71E-06	1	Manh.	0.00025	0.91735
0.07	HEOM	0.01417	0.00403	HEOM	0.00030	0.90331
0.07	**IQR-HEOM**	**3.39E-06**	**1**	**IQR-HEOM**	**3.09E-06**	**1**
0.09	Euc.	0.29676	0.60315	Euc.	0.25122	0.52157
0.09	Manh.	0.12432	0.74124	Manh.	0.14242	0.39247
0.09	HEOM	7.14235	0.38477	HEOM	0.14242	0.39247
0.09	**IQR-HEOM**	**0.15644**	**1**	**IQR-HEOM**	**0.00237**	**1**
0.12	Euc.	0.00031	0.98645	Euc.	0.00612	0.60454
0.12	Manh.	0.00031	0.98254	Manh.	0.00252	0.74153
0.12	HEOM	0.10103	0.75713	HEOM	0.00211	0.75645
0.12	**IQR-HEOM**	**3.78E-04**	**1**	**IQR-HEOM**	**8.82E-07**	**1**
0.15	Euc.	0.00031	0.98254	Euc.	0.00625	0.54441
0.15	Manh.	0.00031	0.98254	Manh.	0.00470	0.60403
0.15	HEOM	0.20522	0.64061	HEOM	0.00417	0.64005
0.15	**IQR-HEOM**	**7.61E-04**	**1**	**IQR-HEOM**	**6.89E-08**	**1**
0.18	Euc.	0.00031	0.98254	Euc.	0.00151	0.45704
0.18	Manh.	0.00031	0.98254	Manh.	0.00815	0.57315
0.18	HEOM	0.03674	0.58363	HEOM	0.00778	0.58021
0.18	**IQR-HEOM**	**0.01310**	**1**	**IQR-HEOM**	**3.65E-08**	**1**
0.21	Euc.	0.00031	0.98245	Euc.	0.01775	0.0.52455
0.21	SManh.	0.00031	0.98251	Manh.	0.00790	0.57410
0.21	HEOM	0.37540	0.58145	HEOM	0.00764	0.58006
0.21	**IQR-HEOM**	**0.01856**	**1**	**IQR-HEOM**	**7.59E-07**	**1**
0.24	Euc.	0.00251	0.97441	Euc.	0.02187	0.48152
0.24	Manh.	0.00031	0.98254	Manh.	0.00144	0.50476
0.24	HEOM	0.46223	0.55055	HEOM	0.00960	0.55057
0.24	**IQR-HEOM**	**0.02487**	**1**	**IQR-HEOM**	**3.09E-08**	**1**
0.27	Euc.	0.00374	0.96503	Euc.	0.03030	0.44205
0.27	Manh.	0.00031	0.98254	Manh.	0.01823	0.50257
0.27	HEOM	0.59137	0.51340	HEOM	0.01233	0.516571
0.27	**IQR-HEOM**	**0.05011**	**1**	**IQR-HEOM**	**3.09E-08**	**1**
0.30	Euc.	0.00515	0.95052	Euc.	0.03269	0.45537
0.30	Manh.	0.00374	0.96503	Manh.	0.02377	0.43556
0.30	HEOM	1.06160	0.48414	HEOM	0.01906	0.48434
0.30	**IQR-HEOM**	**0.11682**	**1**	**IQR-HEOM**	**3.09E-08**	**1**

Table 6: Average Cohesion and Silhouette for various Width Clusters, (n = 160, (x_1, x_2, x_3, x_4))

Normalized(Classic)				Normalized(NADS)		
		Parameters			Parameter	
Width	Methods Dist.	Coh. Max> 1(Min< 1)	Silh. Max=1(Min=0)	Methods Dist.	Coh. Max> 1(Min< 1)	Silh. Max=1(Min=0)
0.01	Euc.	5.59E-04	1	Euc.	3.19E-05	0.99431
0.01	Manh.	7.62E-05	1	Manh.	5.19E-05	1
0.01	HEOM	4.81E-04	1	HEOM	5.19E-05	1
0.01	**IQR-HEOM**	**4.81E-06**	**1**	**IQR-HEOM**	**5.69E-06**	**1**
0.02	Euc.	4.38E-04	1	Euc.	3.09E-06	0.98372
0.02	Manh.	4.38E-04	1	Manh.	5.09E-06	0.98364
0.02	HEOM	0.00031	0.98345	HEOM	2.09E-06	0.98363
0.02	**IQR-HEOM**	**5.39E-05**	**1**	**IQR-HEOM**	**4.53E-07**	**1**
0.03	Euc.	4.38E-04	1	Euc.	4.09E-06	0.98653
0.03	Manh.	4.38E-04	1	Manh.	5.09E-06	0.98354
0.03	HEOM	0.00031	0.98436	HEOM	5.09E-06	0.98373
0.03	**IQR-HEOM**	**5.57E-07**	**1**	**IQR-HEOM**	**6.95E-08**	**1**
0.05	Euc.	0.00038	1	Euc.	0.00025	0.91594
0.05	Manh.	0.00371	1	Manh.	0.00015	0.93742
0.05	HEOM	0.00374	0.96725	HEOM	6.32E-05	0.96714
0.05	**IQR-HEOM**	**3.71E-05**	**1**	**IQR-HEOM**	**4.59E-08**	**1**
0.07	Euc.	7.71E-05	0.99372	Euc.	0.00097	1
0.07	Manh.	7.71E-06	1	Manh.	0.00025	0.91150
0.07	HEOM	0.01417	0.00403	HEOM	0.00030	0.90112
0.07	**IQR-HEOM**	**4.39E-06**	**1**	**IQR-HEOM**	**4.09E-06**	**1**
0.09	Euc.	0.29343	0.60316	Euc.	0.25122	0.52182
0.09	Manh.	0.12441	0.74246	Manh.	0.14353	0.39330
0.09	HEOM	8.14349	0.38611	HEOM	0.14353	0.39334
0.09	**IQR-HEOM**	**0.132802**	**1**	**IQR-HEOM**	**0.00235**	**1**
0.12	Euc.	0.00031	0.98344	Euc.	0.00612	0.60545
0.12	Manh.	0.00031	0.98253	Manh.	0.00252	0.74243
0.12	HEOM	0.10103	0.75124	HEOM	0.00211	0.75067
0.12	**IQR-HEOM**	**5.78E-04**	**1**	**IQR-HEOM**	**2.82E-07**	**1**
0.15	Euc.	0.00031	0.98346	Euc.	0.00625	0.54447
0.15	Manh.	0.00031	0.98345	Manh.	0.00470	0.60516
0.15	HEOM	0.20433	0.64153	HEOM	0.00417	0.641167
0.15	**IQR-HEOM**	**4.61E-04**	**1**	**IQR-HEOM**	**4.89E-08**	**1**
0.18	Euc.	0.00031	0.98436	Euc.	0.00151	0.453563
0.18	Manh.	0.00031	0.98346	Manh.	0.00815	0.57234
0.18	HEOM	0.03674	0.58505	HEOM	0.00778	0.584243
0.18	**IQR-HEOM**	**0.01311**	**1**	**IQR-HEOM**	**4.65E-08**	**1**
0.21	Euc.	0.00031	0.98345	Euc.	0.01775	0.0.52344
0.21	SManh.	0.00031	0.98254	Manh.	0.00790	0.57526
0.21	HEOM	0.37540	0.58145	HEOM	0.00764	0.58116
0.21	**IQR-HEOM**	**0.01856**	**1**	**IQR-HEOM**	**8.59E-07**	**1**
0.24	Euc.	0.00251	0.97350	Euc.	0.02187	0.48661
0.24	Manh.	0.00031	0.98345	Manh.	0.00144	0.50565
0.24	HEOM	0.46223	0.55055	HEOM	0.00960	0.55438
0.24	**IQR-HEOM**	**0.02487**	**1**	**IQR-HEOM**	**6.09E-08**	**1**
0.27	Euc.	0.00374	0.96725	Euc.	0.03030	0.44423
0.27	Manh.	0.00031	0.98345	Manh.	0.01823	0.50340
0.27	HEOM	0.59620	0.51239	HEOM	0.01233	0.51102
0.27	**IQR-HEOM**	**0.05022**	**1**	**IQR-HEOM**	**6.09E-08**	**1**
0.30	Euc.	0.00515	0.95274	Euc.	0.03269	0.45751
0.30	Manh.	0.00374	0.96614	Manh.	0.02377	0.43665
0.30	HEOM	1.061603	0.48414	HEOM	0.01906	0.48434
0.30	**IQR-HEOM**	**0.11165**	**0.99235**	**IQR-HEOM**	**6.09E-08**	**1**

Table 7: Average Cohesion, Silhouette values and Computing Time, (n = 50, 100, 160)

n	Cont.	Method	x_1, x_2		x_1, x_2, x_3, x_4	
			Av.Coh.(Silh.)	C.Time	Av.Coh.(Silh.)	C.Time
50	5%	Euc.	0.0007(0.9920)	52	0.0002(0.9872)	57
		Manh.	0.0003(0.9941)	54	0.0001(0.9883)	56
		HEOM	0.0004(0.9835)	57	0.0001(0.9547)	60
		IQR-HEOM	**1.95E-08(1)**	**47**	**2.36E-07(1)**	**50**
	10%	Euc.	0.0003(0.9372)	62	0.0001(0.8900)	65
		Manh.	0.0005(0.9409)	61	0.0003(0.8875)	62
		HEOM	0.0001(0.8955)	64	0.0004(0.8146)	66
		IQR-HEOM	**4.29E-07(1)**	**54**	**5.33E-07(1)**	**56**
100	5%	Euc.	0.0016(0.6911)	64	0.0027(0.6657)	66
		Manh.	0.0023(0.6892)	63	0.0049(0.6660)	65
		HEOM	0.0032(0.6370)	66	0.5943(0.5780)	68
		IQR-HEOM	**3.92E-06(1)**	**59**	**1.33E-06(1)**	**61**
	10%	Euc.	0.0395(0.4753)	71	0.0317(0.4753)	73
		Manh.	0.0537(0.4778)	69	0.0517(0.4011)	71
		HEOM	0.0327(0.3690)	72	0.0149(0.3084)	74
		IQR-HEOM	**2.69E-05(1)**	**64**	**5.33E-05(1)**	**66**
160	5%	Euc.	0.7146(0.2414)	70	0.8317(0.2340)	72
		Manh.	0.6946(0.2710)	69	0.7938(0.2554)	71
		HEOM	0.9872(0.1915)	72	1.0325(0.1961)	74
		IQR-HEOM	**0.0001(0.9821)**	**64**	**0.0004(0.9504)**	**67**
	10%	Euc.	0.9546(0.1213)	73	0.9934(0.1172)	75
		Manh.	0.9033(0.1424)	70	0.9726(0.1253)	74
		HEOM	1.2437(0.0175)	73	2.1513(0.0017)	77
		IQR-HEOM	**0.0037(0.9673)**	**66**	**0.0022(0.9452)**	**68**

4.1 Real Data Applications

In this section, the Iris, Hayes-Roth, and Tae datasets are considered to verify the performance of our proposed methods:

Iris dataset: The iris dataset was applied by many researchers. The dataset contains 3 classes of 150 instances each, where each class refers to a type of iris plant. It comprises the following attributes information: (1) Sepal length in cm, (2) Sepal width in cm, (3) Petal length in cm, and (4) Petal width in cm. The classes are listed as follows: (1) iris Setosa, (2) iris Verisiclor, and (3) iris Virginica.

Hayes-Roth dataset: The Hayes-Roth dataset was used by many researchers such as Uddin et al. (2017), and Ryu and Eick (2005). The dataset contains 3 classes of 160 instances each, with 4 attributes namely: (1) hobby, (2) age, (3) educational, and (4) marital status.

Tae (Teaching Assistant Evaluation) dataset: The Tae dataset was used by many researchers. The dataset contains 3 classes of 151 instances each, with 5 attributes namely: (1) native, (2) instructor, (3) course, (4) semester, and (5) size.

The performances of our methods are compared to other methods, are evaluated based on the average external validity measures and computational time.

Table 8: Average Silh. coefficients and Coh. values under each Dist. Functions for Iris, Hayes-Roth and Tae Datasets

Iris Dataset								
Methods	Euclidean		Manhattan		HEOM		**IQR-HEOM**	
Parameters	Silh.	Coh.	Silh.	Coh.	Silh.	Coh.	**Silh.**	**Coh.**
	Max =1	Max> 1	Max =1	Max> 1	Max=1	Max> 1	**Max=1**	**Max>1**
Width	(Min=0)	(Min<1)	(Min=0)	(Min<1)	(Min=0)	(Min<1)	**(Min=0)**	**(Min<1)**
0.01	1	3.26E-06	1	3.49E-06	1	3.58E-05	1	6.50E-07
0.02	1	2.71E-05	1	1.57E-05	1	2.87E-04	**1**	**4.94E-06**
0.05	0.997125	1.03E-05	0.997125	1.03E-05	0.800502	0.006038	**1**	**3.97E-05**
0.10	0.981305	0.000164	0.981305	0.000164	0.383485	0.069038	**1**	**1.83E-04**
0.15	0.845241	0.003832	0.946758	0.000829	0.187431	0.159342	**1**	**3.49E-04**
0.20	0.721084	0.010818	0.898028	0.002183	0.252529	0.213464	**1**	**7.73E-04**
Hayes-Roth Dataset								
Methods	Euclidean		Manhattan		HEOM		**IQR-HEOM**	
Parameters	Silh.	Coh.	Silh.	Coh.	Silh.	Coh.	**Silh.**	**Coh.**
	Max=1	Max> 1	Max=1	Max> 1	Max=1	Max> 1	**Max=1**	**Max>1**
Width	(Min=0)	(Min<1)	(Min=0)	(Min<1)	(Min=0)	(Min<1)	**(Min=0)**	**(Min<1)**
0.01	0.996474	8.83E-07	0.996474	8.83E-07	1	9.81E-08	**1**	**8.66E-10**
0.02	0.997278	9.86E-06	0.997278	9.86E-06	1	6.82E-07	**1**	**7.79E-09**
0.05	0.981316	0.000104	0.981316	0.000104	0.996474	8.83E-07	**1**	**8.52E-08**
0.10	0.970324	0.000470	0.970324	0.000470	0.991109	2.69E-05	**1**	**8.85E-07**
0.15	0.961897	0.001516	0.961897	0.001516	0.983836	7.18E-05	**0.997278**	**8.83E-07**
0.20	0.956271	0.001977	0.956271	0.001977	0.976156	0.000168	**0.996474**	**9.86E-06**
Tae Dataset								
Methods	Euclidean		Manhattan		HEOM		**IQR-HEOM**	
Parameters	Silh.	Coh.	Silh.	Coh.	Silh.	Coh.	**Silh.**	**Coh.**
	Max=1	Max> 1	Max=1	Max> 1	Max=1	Max> 1	**Max=1**	**Max>1**
Width	(Min=0)	(Min<1)	(Min=0)	(Min<1)	(Min=0)	(Min<1)	**(Min=0)**	**(Min<1)**
0.01	1	2.87E-09	1	3.67E-08	1	3.59E-04	**1**	**3.64E-08**
0.02	1	1.77E-05	1	3.83E-08	0.966560	0.055188	**1**	**2.93E-07**
0.05	1	3.82E-05	1	2.35E-07	0.896195	0.743929	**1**	**2.58E-07**
0.10	1	3.58E-04	1	1.58E-06	0.736611	3.252349	**1**	**1.69E-06**
0.15	1	1.63E-04	1	1.45E-05	0.662927	6.279123	**1**	**1.28E-05**
0.20	0.956782	1.67E-04	0.983415	3.78E-05	0.452501	22.42122	**1**	**3.17E-04**

Table 8 presents the performance of silhouette coefficients and cohesion values on distance functions. It can be seen that the $IQR-HEOM$ has recorded the highest performance under all the three datasets used. However, the proposed method is better in performance compared to the existing traditional methods.

5. Concluding Remarks

In this paper, we proposed a method to overcome the weakness of using range function as a local normalization in $HEOM$. The procedure applied

in $HEOM$, by dividing it with range allows outliers to have big effect on the contribution of the attributes. They further recommended using interquartile range which is more robust to range against outliers in data preprocessing. The new method is called $IQR - HEOM$. The proposed method is based on the use of IQR function to increase the accuracy and breakdown point (refer to ChitraDevi et al. (2012)). In our proposed method we make use of IQR function, which has breakdown point of 25%, and has little resistance to outliers due to its focus on the center of the distribution in Rousseeuw and Hubert (2011).

To investigate the performance of our proposed method, simulation study and real life datasets are considered. The results indicate that $HEOM$ method has least performance. This may be due to the fact that it uses range function as a local normalization approach; which has 0% breakdown point and is not resistance to outliers.

We also present three different sample size ($n = 50, 100, 160$), with two and four attribute variables each, contaminate at 5% and 10% each and are generated to evaluate the average cohesion values, silhouette coefficients and computing time (minutes). However, despite the contamination of the data, still the proposed method had shown good performance.

The proposed method has good performance, evidently by achieving the maximum point of 1 in the silhouette coefficients and attaining almost the minimum of less than 1 in the cohesion values. Therefore, from the results, it can be concluded that the $IQR-HEOM$ method is better, which shown good performance in the simulation and real life data sets compared to the existing classical methods.

References

Aha, D. W. (1992). Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms. *International Journal of Man-Machine Studies*, 36(2):267–287.

Cao, F., Liang, J., Li, D., Bai, L., and Dang, C. (2012). A dissimilarity measure for the k-modes clustering algorithm. *Knowledge-Based Systems*, 26:120–127. doi:10.1016/j.knosys.2011.07.011.

ChitraDevi, N., Palanisamy, V., Baskaran, K., and Prabeela, S. (2012). A novel distance for clustering to support mixed data attributes and promote data reliability and network lifetime in large

scale wireless sensor networks. *Procedia Engineering*, 30(2012):669–677. https://doi.org/10.1016/j.proeng.2012.01.913.

de Amorim, R. C. and Hennig, C. (2015). Recovering the number of clusters in data sets with noise features using feature rescaling factors. *Information Sciences*, 324:126–145. https://doi.org/10.1016/j.ins.2015.06.039.

Khan, F. (2012). An initial seed selection algorithm for k-means clustering of georeferenced data to improve replicability of cluster assignments for mapping application. *Applied Soft Computing*, 12(11):3698–3700. https://doi.org/10.1016/j.asoc.2012.07.021.

Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137.

Mohamad, I. and Usman, D. (2013). Standardization and its effects on k-means clustering algorithm. *Research Journal of Applied Sciences, Engineering and Technology*, 6:3299–3303. http://dx.doi.org/10.19026/rjaset.6.3638.

Peng, S., Hu, Q., Chen, Y., and Dang, J. (2015). Improved support vector machine algorithm for heterogeneous data. *Pattern Recognition*, 48(6):2072–2083. https://doi.org/10.1016/j.patcog.2014.12.015.

Rousseeuw, P. J. and Hubert, M. (2011). Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):73–79.

Ryu, T. W. and Eick, C. F. (2005). A database clustering methodology and tool. *Information Sciences*, 171(1-3):29–59. https://doi.org/10.1016/j.ins.2004.03.016.

Stanfill, C. and Waltz, D. (1986). Toward memory-based reasoning. *Communications of the ACM*, 29(12):1213–1228. https://doi.org/10.1145/7902.7906.

Uddin, Z., Ahsanuddin, M., and Khan, D. A. (2017). Teaching physics using microsoft excel. *Physics Education*, 52(5):053001.

Zhang, W., Yoshida, T., Tang, X., and Wang, Q. (2010). Text clustering using frequent itemsets. *Knowledge-Based Systems*, 23(5):379–388. https://doi.org/10.1016/j.knosys.2010.01.011.