Pertanika J. Sci. & Technol. 26 (4): 1823 - 1836 (2018)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Statistical Estimators as an Alternative to Standard Deviation in Weighted Euclidean Distance Cluster Analysis

Paul Inuwa Dalatu[1,2]* and Habshah Midi[1,3]

[1]Department of Mathematics, Faculty of Science, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia
[2]Department of Mathematics, Faculty of Science, Adamawa State University, Mubi, PMB 25 Mubi Adamawa State, Nigeria
[3]Institute of Mathematical Research, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia

ABSTRACT

Clustering is basically one of the major sources of primary data mining tools. It makes researchers understand the natural grouping of attributes in datasets. Clustering is an unsupervised classification method with the major aim of partitioning, where objects in the same cluster are similar, and objects which belong to different clusters vary significantly, with respect to their attributes. However, the classical Standardized Euclidean distance, which uses standard deviation to down weight maximum points of the *ith* features on the distance clusters, has been criticized by many scholars that the method produces outliers, lack robustness, and has 0% breakdown points. It also has low efficiency in normal distribution. Therefore, to remedy the problem, we suggest two statistical estimators which have 50% breakdown points namely the *Sn* and *Qn* estimators, with 58% and 82% efficiency, respectively. The proposed methods evidently outperformed the existing methods in down weighting the maximum points of the *ith* features in distance-based clustering analysis.

Keywords: Clustering, estimators, K-Means, simulation, weighted

INTRODUCTION

Clustering analysis is an unsupervised learning. It is widely known as unsupervised learning algorithm because it does not involve any statistical assumption to data (Cao et al., 2009). Velmurugan and Santhanam (2011) stated that data modeling places clustering in a historic viewpoint embedded in mathematics, statistics, and

ARTICLE INFO

Article history:
Received: 10 November 2017
Accepted: 23 March 2018
Published: 24 October 2018

E-mail addresses:
dalatup@gmail.com (Paul Inuwa Dalatu)
habshah@upm.edu.my (Habshah Midi)
* Corresponding author

ISSN: 0128-7680
e-ISSN: 2231-8526

numerical analysis. The major aim of clustering is to disintegrate a dataset into dissimilar subsets called clusters or groups, whereby, data in a particular subset have the same membership or characteristics while different subset presenting dissimilar membership from data in distinct subset.

Generally, the current clustering algorithms obtainable in the literature is aimed to offer hard clusters based on K-Means algorithm. The K-Means particularly practices Euclidean distance to measure the alteration between a data object and its cluster centroid. These distances are commonly calculated from raw data and not from normalized data. Whereas, using Euclidean distances, the distance between any two objects is not affected by the addition of new objects to the analysis. The clustering outcomes can be significantly affected by differences in scale among the dimension in which the distances are calculated through. Data normalization is the linear transformation of data to a specific range (Visalakshi &Thangavel, 2009).

Usually, in computing the Euclidean distance function, all features add the same to the function value. Subsequently, different features are usually calculated with different metrics or different magnitudes, and these must be normalized (or standardized) before using the distance function (Munz et al., 2007; Xu & Tian, 2015).

Therefore, one of the issues of K-Means weakness is that it may not always yield global optimum outcomes (Reddy et al., 2012), which necessitates normalizing different metrics when using Euclidean distance function in cluster analysis. However, Xu and Tian (2015) addressed the reported issue and proposed a Standardized Weighted Euclidean Distance:

$$d\left(x_i, x_j\right) = \sqrt{\sum_{i=1}^{n}\left(\frac{x_i - x_j}{s_i}\right)^2}\ , \qquad [1]$$

where s_i (s_i standard deviation of dataset) is an empirical normalization and weighing factor of the *ith* feature. It is observe that the bigger s_i , the smaller is the effect of the *ith* feature on the distance function.

In recent times, some researchers have identified the limitations and drawbacks of the standardized normalization (see Mohamad & Usman, 2013, Jayalakshmi & Santhakumaran, 2011, Jain et al., 2005). Gnanadesikan et al. (1995) studied and conducted experiments on the performance of nine methods on eight most important simulated and real data. Their outcomes revealed and demonstrated weakness of weighting based on the standard deviation. Furthermore, Milligan and Cooper (1988) presented simulation studies on standardization issue. They experimented eight standardization approaches, and the classical Z-Score (i.e. standard deviation) normalization was found to be less effective in various circumstances. However, Sarstedt and Mooi (2014) recommended that in

most cases, normalization by range performed better compared to standard deviation. Furthermore, Matthews (1979) argued that, by down weighting the whole sample using method like Standardized Weighted Euclidean based on variability may probably eliminate the significant between-cluster consistency. Hence, this motivated us to propose *Qn* and *Sn* estimators to replace standard deviation in Standardized Euclidean distance called *Qn* and *Sn* Weighted Euclidean distance, respectively.

This paper is structured as follows: Section 2 provides the materials and methods. Section 3 presents the results and discussion. Section 4 proffers the conclusions of the study.

MATERIALS AND METHODS

Conventional Distance Functions

According to Giancarlo et al. (2010), distance functions are vital components of classification and clustering techniques. Therefore, in comparing performance of the proposed distance function, the K-Means clustering algorithm is also executed using various traditional distance functions, such as the Euclidean and the weighted Euclidean distance.

Euclidean Distance

The most popular distance measure for numerical data is possibly the Euclidean distance, also well-known as L_2 norm, as defined in (Shirkorshidi et al., 2015):

$$d\left(x_i, x_j\right) = \sqrt{\sum_{i=1}^{n}\left(x_i - x_j\right)^2} \qquad [2]$$

This distance measure has the appealing property in which the $d\left(x_i, x_j\right)$ can be interpreted as the 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.

Standardized Euclidean

The Standardized Euclidean (or sometimes called Weighted Euclidean) was first proposed by Orloci (1967) based on the fact that Euclidean distance had some demerit of absoluteness on the method. This standardization eliminates the limiting effect of all attribute variables in samples on the maximum likely distance. Therefore, this function will assist in giving equal weight to different values in the set and the distance will become scale invariant. Recently, it was criticized by Gerstenberger and Vogel (2015) that as far as standard deviation was applied to down weight some maximum points, it was prone to outliers and lack robustness. The Standardized Euclidean is computed as in Xu and Tian (2015), Equation [1] revisited.

Proposed Weighted Euclidean Distance Functions

In this section we discuss the two proposed Weighted Euclidean distance functions. The two proposed functions are based on the Weighted Euclidean or sometimes called Standardized Euclidean of Xu and Tian (2015). Xu and Tian (2015) claimed that the larger s_i (denotes the standard deviation of the dataset), the smaller was the influence of the *ith* feature on the distance is. They believed that the rationale behind the method is the assumption that both normal and anomalous may appear from different clusters in the features space. Perhaps, the data may contain outliers which do not belong to a bigger cluster, yet this does not disturb the K-Means clustering as long the number of outliers is small.

Recently, Gerstenberger and Vogel (l2015) criticized the Standardized Weighted Distance in Equation [1] in which it was based on standard deviation to down weight the data. They noted that this method lacked robustness, because the calculation of standard deviation was based on the sample mean which was very sensitive to outliers.

A Standard deviation has 0% breakdown point as stated in Rousseeuw and Hubert (2011). It is susceptible to outliers and has low efficiency in heavy-tailed distributions (Gerstenberger & Vogel, 2015). In order to remedy this problem, we suggest to employing high break down point estimators of *Sn* and Q*n* where both have 50% breakdown points with efficiency of 58% and 82%, respectively (Rousseeuw & Croux, 1993).

The proposed methods are summarized as follows:

***Qn* Weighted Euclidean Distance Function.**

$$d\left(x_i, x_j\right)=\sqrt{\sum_{i=1}^{n}\left(\frac{x_i - x_j}{Qn}\right)^2} \qquad [3]$$

where:

$$Qn = c\left\{\left|x_i - x_j\right|; i < j\right\}, \qquad [4]$$

Rousseeuw and Croux (1993) suggested assigning c = 2.2219 for consistency in the Gaussian distribution.

***Sn* Weighted Euclidean Distance Function.**

$$d\left(x_i, x_j\right)=\sqrt{\sum_{i=1}^{n}\left(\frac{x_i - x_j}{Sn}\right)^2}\ , \qquad [5]$$

where *Sn* is the median of the *n* medians of the absolute difference between x_i and x_j:

$$Sn = c \times med_i \left\{ med_j \left| x_i - x_j \right|; j = 1,2,...,n \right\}, \quad [6]$$

Rousseeuw and Croux (1993) suggested assigning $c = 1.1926$ for consistency in the Gaussian distribution.

RESULTS AND DISCUSSION

Simulation Study

In this section, Monte Carlo simulation study is presented to compare the performance of existing methods, such as Euclidean distance (Loohach & Garg, 2012), and Standardized Weighted Euclidean distance (Xu & Tian, 2015) with our proposed methods *Qn* Weighted Euclidean distance, and *Sn* Weighted Euclidean distance.

Following Loohach and Garg (2012), and Xu and Tian (2015), two (x_1, x_2) and four (x_1, x_2, x_3, x_4) variables are generated such that each of the exploratory variables (x_1, x_2) and (x_1, x_2, x_3, x_4) are simulated from uniform distribution with range [-10, 10]. The variables are clustered into three classes (clusters, groups) as; cluster 1, cluster 2, and cluster 3. We consider a sample of size 50, 100 and 160. The basis for using different sample sizes is to ascertain the consistency, effectiveness and accuracy of the proposed methods compared to the existing methods. The conventional distance functions, Euclidean and Standardized Weighted Euclidean, and the *Qn* and *Sn* Weighted Euclidean Distance Functions were then applied to the data. Some external validity measures such as; purity (Hernandez-Torruco et al. 2014), fowlkes-mallow index (Velardi et al., 2012), rand index (Noorbehbahani et al., 2015: Rand, 1971;), f-measure (score) (Velardi et al., 2012), jaccard index (Velardi et al., 2012), recall (Velardi et al., 2012), f-measure (beta varied) (Velardi et al., 2012), geometric-mean (Tomar & Agarwal, 2015), precision (Kou et al., 2014: Rokach & Maimon, 2008)), specificity (Velardi et al., 2012), accuracy (Tomar & Agarwal, 2015) and sensitivity (Velardi et al., 2012), computing time, and maximum number of clusters clustered are recorded. In each of the experimental runs, there are 1000 replications. The performance of the four methods are evaluated based on average external validity measures for each distance functions, computational timing (minutes), and having three levels of cluster as; cluster 1, cluster 2, and cluster 3. The values in the parenthesis are unnormalized data and not in parenthesis for normalized data. A good method is one that has maximum external validity measure nearly equal to 1 (at maximum 1), less computing time and maximum numbers clustered in each cluster.

Table 1
Average external validity measures, computing time and maximum clusters for n = 50 (x_1, x_2)

Distance Functions	Euclidean	Weighted Euclidean	Qn W. Euclidean	Sn W. Euclidean
Purity	0.866(0.822)	0.869(0.825)	0.893(0.863)	0.893(0.863)
Fow. M. I.	0.867(0.825)	0.876(0.832)	0.892(0.852)	0.892(0.862)
Rand Index	0.895(0.852)	0.894(0.860)	0.930(0.900)	0.931(0.901)
F-M. (Score)	0.854(0.810)	0.857(0.813)	0.882(0.851)	0.881(0.851)
Jaccard Index	0.789(0.745)	0.792(0.768)	0.813(0.790)	0.800(0.793)
Recall	0.866(0.822)	0.867(0.823)	0.897(0.864)	0.891(0.862)
F-M. (varied)	0.852(0.814)	0.866(0.822)	0.887(0.854)	0.885(0.852)
G. Means	0.872(0.837)	0.879(0.846)	0.892(0.865)	0.892(0.867)
Precision	0.878(0.834)	0.879(0.835)	0.897(0.874)	0.896(0.863)
Specificity	0.899(0.855)	0.900(0.890)	0.980(0.953)	0.947(0.914)
Accuracy	0.874(0.831)	0.894(0.850)	0.961(0.930)	0.965(0.932)
Sensitivity	0.866(0.822)	0.867(0.823)	0.898(0.865)	0.899(0.866)
Average	0.865(0.822)	0.870(0.832)	0.902(0.891)	0.907(0.886)
Compt. Time	35(38)	35(37)	28(30)	28(30)
Clust.1(max. 15)	12(11)	11(11)	14(12)	13(12)
Clust.2 (max.15)	10(10)	11(9)	12(11)	12(12)
Clust.3 (max.20)	15(14)	(15)	17(16)	17(16)

Table 2
Average external validity measures, computing time and maximum clusters for n = 50 (x_1, x_2, x_3, x_4)

Distance Functions	Euclidean	Weighted Euclidean	Qn W. Euclidean	Sn W. Euclidean
Purity	0.877(0.844)	0.878(0.845)	0.930(0.901)	0.921(0.916)
Fow. M. I.	0.871(0.846)	0.892(0.863)	0.920(0.911)	0.920(0.911)
Rand Index	0.894(0.861)	0.893(0.860)	0.939(0.922)	0.931(0.924)
F-M. (Score)	0.861(0.832)	0.863(0.830)	0.937(0.914)	0.936(0.903)
Jaccard Index	0.834(0.801)	0.837(0.804)	0.913(0.900)	0.913(0.901)
Recall	0.871(0.843)	0.872(0.845)	0.907(0.890)	0.901(0.890)
F-M. (varied)	0.897(0.864)	0.891(0.865)	0.960(0.930)	0.944(0.911)
G. Means	0.929(0.910)	0.929(0.910)	0.951(0.944)	0.969(0.937)
Precision	0.893(0.864)	0.896(0.864)	0.947(0.914)	0.956(0.923)
Specificity	0.895(0.862)	0.896(0.863)	0.946(0.913)	0.947(0.914)
Accuracy	0.894(0.871)	0.898(0.865)	0.952(0.936)	0.965(0.942)
Sensitivity	0.894(0.871)	0.892(0.870)	0.978(0.945)	0.969(0.936)
Average	0.893(0.855)	0.886(0.857)	0.940(0.920)	0.939(0.917)
Compt. Time	42(45)	42(45)	37(40)	37(40)
Clust.1(max. 15)	11(9)	10(10)	13(13)	13(13)
Clust.2 (max.15)	9(9)	9(9)	13(12)	12(11)
Clust.3 (max.20)	14(13)	15(12)	17(15)	18(16)

Table 3
Average external validity measures, computing time and maximum clusters for n = 100 (x_1, x_2)

Distance Functions	Euclidean	Weighted Euclidean	Qn W. Euclidean	Sn W. Euclidean
Purity	0.894(0.861)	0.894(0.861)	0.960(0.930)	0.963(0.932)
Fow. M. I.	0.895(0.862)	0.895(0.862)	0.950(0.920)	0.950(0.920)
Rand Index	0.897(0.864)	0.898(0.865)	0.970(0.940)	0.964(0.931)
F-M. (Score)	0.896(0.863)	0.897(0.864)	0.967(0.934)	0.969(0.936)
Jaccard Index	0.834(0.801)	0.837(0.804)	0.913(0.900)	0.913(0.900)
Recall	0.894(0.861)	0.895(0.862)	0.937(0.904)	0.937(0.904)
F-M. (varied)	0.897(0.864)	0.895(0.862)	0.933(0.900)	0.937(0.904)
G. Means	0.896(0.863)	0.899(0.866)	0.954(0.921)	0.963(0.930)
Precision	0.895(0.863)	0.959(0.926)	0.970(0.940)	0.970(0.940)
Specificity	0.898(0.865)	0.899(0.866)	0.970(0.940)	0.980(0.950)
Accuracy	0.897(0.864)	0.897(0.864)	0.945(0.912)	0.948(0.915)
Sensitivity	0.894(0.861)	0.894(0.861)	0.960(0.932)	0.960(0.931)
Average	0.891(0.858)	0.897(0.864)	0.952(0.923)	0.955(0.924)
Compt. Time	47(52)	46(52)	41(43)	41(43)
Clust.1(max. 30)	25(24)	26(23)	26(26)	27(25)
Clust.2 (max.30)	20(20)	22(21)	26(23)	26(24)
Clust.3 (max.40)	33(30)	31(30)	35(34)	34(34)

Table 4
Average external validity measures, computing time and maximum clusters for n = 100 (x_1, x_2, x_3, x_4)

Distance Functions	Euclidean	Weighted Euclidean	Qn W. Euclidean	Sn W. Euclidean
Purity	0.896(0.863)	0.899(0.866)	0.950(0.920)	0.954(0.931)
Fow. M. I.	0.896(0.863)	0.898(0.865)	0.953(0.920)	0.953(0.920)
Rand Index	0.899(0.866)	0.898(0.865)	0.970(0.940)	0.964(0.931)
F-M. (Score)	0.894(0.861)	0.893(0.860)	0.970(0.940)	0.969(0.936)
Jaccard Index	0.867(0.834)	0.870(0.840)	0.947(0.914)	0.946(0.913)
Recall	0.896(0.863)	0.894(0.861)	0.941(0.911)	0.935(0.901)
F-M. (varied)	0.893(0.860)	0.893(0.860)	0.933(0.900)	0.937(0.904)
G. Means	0.895(0.862)	0.895(0.863)	0.954(0.0.921)	0.960(0.930)
Precision	0.897(0.864)	0.895(0.862)	0.970(0.940)	0.970(0.941)
Specificity	0.898(0.865)	0.900(0.890)	0.973(0.941)	0.980(0.950)
Accuracy	0.897(0.864)	0.897(0.864)	0.945(0.912)	0.949(0.916)
Sensitivity	0.894(0.861)	0.895(0.863)	0.960(0.930)	0.960(0.930)
Average	0.894(0.861)	0.894(0.861)	0.947(0.924)	0.956(0.925)
Compt. Time	53(55)	53(55)	46(49)	46(49)
Clust.1(max. 30)	23(22)	24(23)	25(24)	26(25)
Clust.2 (max.30)	23(21)	23(20)	25(24)	25(23)
Clust.3 (max.40)	31(28)	30(28)	35(33)	34(33)

Table 5
Average external validity measures, computing time and maximum clusters for n = 160 (x_1, x_2)

Distance Functions	Euclidean	Weighted Euclidean	Qn W. Euclidean	Sn W. Euclidean
Purity	0.911(0.894)	0.914(0.897)	0.963(0.930)	0.963(0.931)
Fow. M. I.	0.912(0.896)	0.912(0.895)	0.963(0.930)	0.953(0.932)
Rand Index	0.940(0.910)	0.948(0.915)	0.970(0.940)	0.965(0.932)
F-M. (Score)	0.910(0.891)	0.913(0.894)	0.970(0.940)	0.970(0.940)
Jaccard Index	0.0.834(0.797)	0.837(0.785)	0.947(0.914)	0.947(0.914)
Recall	0.911(0.894)	0.912(0.892)	0.941(0.911)	0.935(0.902)
F-M. (varied)	0.907(0.900)	0.911(0.891)	0.945(0.914)	0.937(0.910)
G. Means	0.929(0.910)	0.929(0.911)	0.954(0.923)	0.962(0.932)
Precision	0.923(0.890)	0.926(0.913)	0.970(0.941)	0.969(0.936)
Specificity	0.955(0.932)	0.961(0.931)	0.970(0.952)	0.980(0.943)
Accuracy	0.940(0.913)	0.948(0.935)	0.952(0.934)	0.965(0.943)
Sensitivity	0.911(0.893)	0.912(0.890)	0.972(0.941)	0.960(0.934)
Average	0.915(0.893)	0.919(0.896)	0.960(0.931)	0.959(0.932)
Compt. Time	63(64)	63(64)	56(60)	56(60)
Clust.1(max. 50)	46(46)	45(44)	46(46)	46(45)
Clust.2 (max.50)	47(45)	46(44)	47(45)	47(46)
Clust.3 (max.60)	53(53)	54(54)	58(57)	58(58)

Table 6
Average external validity measures, computing time and maximum clusters for n = 160 (x_1, x_2, x_3, x_4)

Distance Functions	Euclidean	Weighted Euclidean	Qn W. Euclidean	Sn W. Euclidean
Purity	0.916(0.912)	0.919(0.915)	0.953(0.943)	0.953(0.943)
Fow. M. I.	0.917(0.915)	0.926(0.922)	0.948(0.932)	0.942(0.932)
Rand Index	0.935(0.922)	0.944(0.930)	0.960(0.948)	0.951(0.940)
F-M. (Score)	0.924(0.911)	0.927(0.913)	0.952(0.945)	0.958(0.945)
Jaccard Index	0.893(0.884)	0.902(0.890)	0.943(0.932)	0.950(0.939)
Recall	0.916(0.902)	0.920(0.913)	0.947(0.934)	0.942(0.936)
F-M. (varied)	0.905(0.891)	0.911(0.892)	0.937(0.924)	0.935(0.922)
G. Means	0.912(0.907)	0.919(0.906)	0.952(0.945)	0.957(0.947)
Precision	0.908(0.894)	0.919(0.895)	0.947(0.934)	0.946(0.936)
Specificity	0.929(0.915)	0.929(0.914)	0.950(0.941)	0.957(0.944)
Accuracy	0.914(0.901)	0.924(0.910)	0.951(0.940)	0.955(0.942)
Sensitivity	0.916(0.902)	0.920(0.913)	0.947(0.934)	0.942(0.936)
Average	0.915(0.905)	0.921(0.909)	0.949(0.938)	0.949(0.939)
Compt. Time	70(72)	70(72)	65(68)	65(68)
Clust.1(max. 50)	45(44)	45(44)	46(45)	47(46)
Clust.2 (max.50)	44(44)	45(43)	47(46)	47(46)
Clust.3 (max.60)	54(53)	55(52)	56(55)	56(55)

Tables 1, 2, 3, 4, 5, and 6 present the average values of 1000 replications of average external validity measures, maximum number of samples in each cluster, and the computing time (minutes).

It is evidently clear that all the two proposed methods have shown impressive performance by achieving the highest average maximum external validity measures and recording lowest computational timing. The proposed methods have also been clustered to nearly the maximum numbers required to be clustered in each cluster (group). This indicates that the performance of the proposed methods is more accurate and efficient compared to the existing methods.

Table 7 presents the average external validity measures and computing time based on 1000 simulation runs, for 5% and 10% contaminated data generated from uniform

Table 7
Average external validity measures and computing time (minutes) for n = 50, 100, and 160

n	Contaminated	Method	x_1, x_2		x_1, x_2, x_3, x_4	
			Av. Ext. Val.	Comp. Time	Av. Ext. Val.	Comp. Time
50	5%	Euclidean	0.6752	53	0.6634	56
		W.ted Eu.	0.6947	52	0.6815	54
		Qn W. Eu.	0.7427	47	0.7141	50
		Sn W. Eu.	0.7503	46	0.7223	49
	10%	Euclidean	0.6178	60	0.5912	61
		W.ted Eu.	0.6331	59	0.6152	62
		Qn W. Eu.	0.6792	54	0.6573	56
		Sn W. Eu.	0.6801	53	0.6633	55
100	5%	Euclidean	0.6973	62	0.6864	64
		W.ted Eu.	0.6978	62	0.6982	65
		Qn W. Eu.	0.7433	59	0.7117	61
		Sn W. Eu.	0.7392	59	0.7114	63
	10%	Euclidean	0.6191	68	0.6043	70
		W.ted Eu.	0.6186	68	0.6128	71
		Qn W. Eu.	0.6707	64	0.6672	66
		Sn W. Eu.	0.6718	64	0.6595	66
160	5%	Euclidean	0.5984	68	0.5873	70
		W.ted Eu.	0.6112	67	0.6004	69
		Qn W. Eu.	0.6673	64	0.6433	67
		Sn W. Eu.	0.6621	63	0.6403	67
	10%	Euclidean	0.5546	69	0.5334	73
		W.ted Eu.	0.5722	70	0.5471	74
		Qn W. Eu.	0.6371	66	0.6013	68
		Sn W. Eu.	0.6494	65	0.6105	67

distribution with range [15, 16]. From the table it is exciting to note that despite the contamination of the data, the two proposed methods; *Qn* weighted Euclidean and *Sn* weighted Euclidean outperformed the two existing methods. Hence, these findings show that the two proposed methods can perform fairly well even in the presence of contamination.

Real Data Applications

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

Iris Dataset. The iris dataset has been used by many researchers, such as Galili (2015), Jayalakshmi and Santhakumaran (2011), Benson-Putnins et al. (2011), and Han et al. (2011). 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 (Bache and Lichman, 2013).

Hayes-Roth Dataset. The Hayes-Roth dataset has also been used by many researchers, such as Uddin et al. (2017), Han et al. (2011), Jayalakshmi and Santhakumaran, (2011), 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 (Bache and Lichman, 2013).

The average external validity measures and computing time under each distance function for Iris and Hayes-Roth datasets are presented in Tables 8 and 9, respectively.

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

It is clear that all the proposed methods have achieved better performance in the two datasets used. It is important to note that Iris data set in Table 8 has recorded the following results; average external validity measures for Euclidean (0.89387), Standardized Weighted Euclidean (0.88002), *Qn* Weighted Euclidean (0.90379) and *Sn* Weighted Euclidean (90262). While, the computing time (minutes) for Euclidean (44), Standardized Weighted Euclidean (44), *Qn* Weighted Euclidean (42) and *Sn* Weighted Euclidean (42). The Hayes-Roth data set in Table 9 has the following results as; external validity measures for Euclidean (0.66190), Standardized Weighted Euclidean (0.66100). *Qn* Weighted Euclidean (0.67031) and *Sn* Weighted Euclidean (0.67166). While, also its computing time (minutes) for Euclidean (45), Standardized Weighted Euclidean (45), *Qn* Weighted Euclidean (43) and *Sn* Weighted Euclidean (43). Generally, on the average, the two datasets indicated that the two proposed methods had shown impressive performance. Therefore, the results based

on the two datasets applied confirmed that the real numbers used in iris dataset provided higher quality performance in the external validity measures compared to integer numbers used in the Hayes-Roth dataset.

Table 8
Average external validity measures and computing time under each Distance Functions for Iris Dataset

Distance Functions	Euclidean	Weighted Euclidean	Qn Weighted Euc.	Sn Weighted Euc.
Purity	0.88667	0.85333	0.89230	0.89107
Fow. M. I.	0.88876	0.85412	0.89117	0.89217
Rand Index	0.92444	0.90222	0.93657	0.93556
F-M. (Score)	0.88609	0.86327	0.89333	0.89111
Jaccard Index	0.80454	0.79793	0.82271	0.82172
Recall	0.88667	0.88533	0.89667	0.89333
F-M. (varied)	0.88528	0.88358	0.89219	0.89229
G. Means	0.91161	0.89848	0.93440	0.93392
Precision	0.89786	0.88599	0.91631	0.91476
Specificity	0.94333	0.93667	0.94743	0.94667
Accuracy	0.92444	0.91222	0.92577	0.92556
Sensitivity	0.88667	0.88533	0.89667	0.89333
Average	0.89387	0.88002	0.90379	0.90262
Compt. Time	44	44	42	42

Table 9
Average external validity measures and computing time under each Distance Functions for Hayes-Roth Dataset

Distance Functions	Euclidean	Weighted Euclidean	Qn Weighted Euc.	Sn Weighted Euc.
Purity	0.61250	0.50375	0.62125	0.62625
Fow. M. I.	0.60666	NaN	0.62372	0.62873
Rand Index	0.77500	0.75002	0.78417	0.77939
F-M. (Score)	0.61236	NaN	0.63216	0.63792
Jaccard Index	0.43881	0.42504	0.44223	0.44135
Recall	0.65132	0.65221	0.65399	0.65333
F-M. (varied)	0.65055	NaN	0.65993	0.66148
G. Means	0.71888	0.70453	0.72183	0.72519
Precision	0.60256	NaN	0.61529	0.61447
Specificity	0.84783	0.83767	0.84861	0.85251
Accuracy	0.77500	0.76250	0.78654	0.78607
Sensitivity	0.65132	0.65221	0.65399	0.65333
Average	0.66190	0.66100	0.67031	0.67166
Compt. Time	45	45	43	43

Note: NaN = Not a Number

CONCLUSION

In this paper, we proposed two methods to overcome the weakness of Standard Weighted Euclidean Distance method, whereby it has 0% breakdown point characteristics (Rousseeuw & Hubert, 2011), a lack of robustness, is susceptible to outliers, and its low efficiency at heavy-tailed distributions (Gerstenberger and Vogel, 2015). The proposed methods are called *Qn* Weighted Euclidean and *Sn* Weighted Euclidean distance functions. These methods are based on the increase of accuracy and efficiency in performance using of high breakdown estimators as *Qn* and *Sn* both have 50% breakdown points, and their efficiency *Sn* is 58% and *Qn* is 82% (Rousseeuw and Croux, 1993). Therefore, to improve the accuracy and efficiency of Standardized Weighted Euclidean (Xu & Tian, 2015), we employed and adopted the ideas of Rousseeuw and Hubert (2011) to make the distribution more symmetric.

Furthermore, we also presented average external validity measures and computing time (minutes) based on 1000 simulation runs for contaminated data. From the outcome, it is exciting to observe that despite the contamination of the data, the two suggested methods had performed better compared to the existing methods. To investigate the performance of our proposed methods, a simulation study and real data were considered. The results indicate that the Euclidean distance function has the least performance. This is due to the fact that the Euclidean distance has not applied any of the existing estimators to down weight the datasets. However, the two proposed methods have good performance; evidently, by achieving nearly maximum points in the average external validity measures, lower computational timing and clustering the object points to almost all their maximum number of cluster centers.

From the results, it can be concluded that the two proposed methods are much better in performance compared to the existing methods.

REFERENCES

Bache, K., & Lichman, M. (2013). *UCI machine learning repository*. Irvine, CA: University of California, School of Information and Computer Science.

Benson-Putnins, D. A. V. I. D., Bonfardin, M., Magnoni, M. E., & Martin, D. (2011). Spectral clustering and visualization: a novel clustering of Fisher's Iris data set. *SIAM Undergraduate Research Online, 4*, 1-15.

Cao, F., Liang, J., & Jiang, G. (2009). An initialization method for the K-Means algorithm using neighborhood model. *Computers & Mathematics with Applications*, *58*(3), 474-483.

Galili, T. (2015). Dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics*, *31*(22), 3718-3720.

Gerstenberger, C., & Vogel, D. (2015). On the efficiency of Gini's mean difference. *Statistical Methods & Applications*, *24*(4), 569-596.

Giancarlo, R., Bosco, G. L., & Pinello, L. (2010, January). Distance functions, clustering algorithms and microarray data analysis. In *International Conference on Learning and Intelligent Optimization* (pp. 125-138). Berlin, Heidelberg: Springer.

Gnanadesikan, R., Kettenring, J. R., & Tsao, S. L. (1995). Weighting and selection of variables for cluster analysis. *Journal of Classification*, *12*(1), 113-136.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and techniques*. *Sabre Foundation.* Waltham, USA: Morgan Kaufmann Publishers.

Hernández-Torruco, J., Canul-Reich, J., Frausto-Solís, J., & Méndez-Castillo, J. J. (2014). Feature selection for better identification of subtypes of Guillain-Barré syndrome. *Computational and Mathematical Methods in Medicine, 2014*, 1-9.

Jayalakshmi, T., & Santhakumaran, A. (2011). Statistical Normalization and Back Propagation for Classification. *International Journal of Computer Theory and Engineering*, *3*(1), 1793-8201.

Kou, G., Peng, Y., & Wang, G. (2014). Evaluation of clustering algorithms for financial risk analysis using MCDM methods. *Information Sciences*, *275*, 1-12.

Liu, W., Liang, Y., Fan, J., Fen, Z., & Cai, Y. (2014). Improved hierarchical K-Means clustering algorithm without iteration based on distance measurement. In *International Conference on Intelligent Information Processing* (pp. 38- 46). Berlin, Heidelberg: Springer.

Loohach, R., & Garg, K. (2012). Effect of distance functions on simple k-means clustering algorithm. *International Journal of Computer Applications*, *49*(6), 7-9.

Ma, M., Luo, Q., Zhou, Y., Chen, X., & Li, L. (2015). An improved animal migration optimization algorithm for clustering analysis. *Discrete Dynamic in Nature and Society, 2015*, 1-12.

Matthews, A. (1979). Standardization of Measures Prior to Cluster-Analysis. B*iometrics, 35*(4), 765-773.

Milligan, G. W., & Cooper, M. C. (1988). A study of standardization of variables in cluster analysis. *Journal of Classification*, *5*(2), 181-204.

Mohamad, I. B., & Usman, D. (2013). Standardization and its effects on K-means clustering algorithm. *Research Journal of Applied Sciences, Engineering and Technology*, *6*(17), 3299-3303.

Münz, G., Li, S., & Carle, G. (2007, September). Traffic anomaly detection using K-eans clustering. In *Proceedings of Performance, Reliability and Dependability Evaluation of Communication Networks and Distributed Systems, GI/ITG Workshop MMBnet* (pp. 13-14). Hamburg, Germany.

Noorbehbahani, F., Mousavi, S. R., & Mirzaei, A. (2015). An incremental mixed data clustering method using a new distance measure. *Soft Computing*, *19*(3), 731-743.

Orloci, L. (1967). An agglomerative method for classification of plant Communities. *The Journal of Ecology, 55*(1), 193-206.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*(336), 846-850.

Reddy, D., Jana, P. K., & Member, I. S. (2012). Initialization for k-means clustering using voronoi diagram. *Procedia Technology, 4*, 395-400.

Rokach, L., & Maimon, O. Z. (2008). *Data mining with decision trees: theory and applications* (Vol. 69). Singapore: World scientific Publishing.

Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, *88*(424), 1273-1283.

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

Ryu, T. W., & Eick, C. F. (2005). A database clustering methodology and tool. *Information Sciences*, *171*(1), 29-59.

Sarstedt, M., & Mooi, E. (2014). *A concise guide to market research: The Process, Data, and Methods using IBM SPSS Statistics* (pp. 273-318). Berlin Heidelberg: Springer.

Shirkhorshidi, A. S., Aghabozorgi, S., & Wah, T. Y. (2015). A comparison study on similarity and dissimilarity measures in clustering continuous data. *PloS one*, *10*(12), e0144059.

Tomar, D., & Agarwal, S. (2015). Hybrid feature selection based weighted least squares twin support vector machine approach for diagnosing breast cancer, hepatitis, and diabetes. *Advances in Artificial Neural Systems*, *2015*, 1-10.

Uddin, J., Ghazali, R., & Deris, M. M. (2017). An Empirical Analysis of Rough Set Categorical Clustering Techniques. *PloS one*, *12*(1), e0164803.

Velardi, P., Navigli, R., Faralli, S., & Ruiz-Martinez, J. M. (2012, May). A New Method for Evaluating Automatically Learned Terminological Taxonomies. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC* (pp.1498-1504). Istanbul, Turkey.

Velmurugan, T., & Santhanam, T. (2011). An experimental approach. *Information Technology Journal*, *10*(3), 478-484.

Visalakshi, N. K., & Thangavel, K. (2009). Impact of normalization in distributed K-means clustering. *International Journal of Soft Computing*, *4*(4), 168-172.

Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, *2*(2), 165-193.