International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 11 (2018) pp. 9364-9376

Study on Different Cluster Validity Indices

Shyam Kumar K[1], Dr. Raju G[2]

[1]*Department of Computer Application, NSS College Rajakumari, Idukki, Kerala, India.*
[2]*Department of Information Technology, Kannur University, Kannur, Kerala, India.*

Abstract

Data clustering algorithms produce a specified number of clusters based on inherent similarities existing in the dataset. The study on what constitutes a suitable number of clusters for a dataset and the assessment of quality of generated clusters has been a challenging issue in research for decades. Our study is aimed at clustering the data using variant forms of K-Means algorithm and assessing the quality of clusters generated using a set of well-defined cluster validity indices. The study with the intension of getting an assumption on optimum number of clusters for the dataset was performed on a set of standard datasets. The study recommends the development of suitable measures for finding an ideal clustering algorithm and providing some hindsight on optimal number of clusters for the dataset.

Keywords: Cluster Validity Indices, clustering, Dunn Index, Gamma Index.

INTRODUCTION

Information retrieval from voluminous stored data requires advanced and sophisticated data handling capabilities of Data Mining. Data Clustering, a technique in Data Mining, is often used in obtaining a partitional subset of the data, based on similarity of objects in the dataset. The resultant clusters ideally should have maximum intra cluster similarity [1]. Clustering should preferably result in a set of well partitioned subset with no overlapping objects in the partition. Finding an optimal clustering algorithm is still a challenge in the research field in spite of abundance of methods and algorithms available for clustering.

Hierarchical clustering and partitional clustering are the prominent methods of clustering the data. Irrespective of the method of clustering adopted, the output generated is always a user specified number of clusters of the given dataset. The performance of the clustering algorithm depends on the required parameters of clustering. In partitional clustering, the K-Means algorithm requires the number of clusters (k) to be formed, the maximum number of iterations to be performed etc. [2] as parameters before the clusters are generated. It is known that the same clustering algorithm forms different clusters for the same value of k, based on the initial position of seed points and number of iterations. Specification of initial seed point position and the way in which it is specified (whether k random points are selected or the user chooses them) is also known to affect the clusters. In hierarchical clustering too, the results vary on the basis of distance measure chosen and the number of clusters to be formed. Also, the level at which hierarchical clustering is stopped affects the clusters. This necessitates analysis of clusters to come out with clusters of better quality.

The quality of clusters [3] formed is of utmost importance, the clusters needs to be analyzed to see how closely the elements are bound within the clusters and also how well separated the clusters are. The clusters should represent the dataset well and should have maximum similarity within. The final set of clusters should be stable, with no intra cluster movement of objects. Thus, cluster quality evaluation becomes essential in all research applications involving data clustering.

EVALUATION OF CLUSTERS

The clustering algorithms produces user specified number of clusters. Based on the distance measure used and the parameters provided, in most of the clustering algorithms, clusters formed will be drastically different. For example, the popular K-Means algorithm has many variant forms like Lloyd, Forgy, Mac Queen and Hartigan-Wong etc. The final clusters is known to be dependent on how the cluster centers are modified across iterations and on the usage of commonly used distance calculation methods like Euclidian distance, Manhattan distance, Chebyshev distance or Minkowski distance [4]. Hence the clustered results need to be evaluated for quality as well as usefulness.

Theoretically, four categories of cluster validations exist [5]. They are,

1. Relative Cluster Validation: Here we evaluate the clustered results by varying the parameters and re-execute the algorithm. This method is often helpful in finding the optimal number of clusters, by varying the user specified parameter K.
2. External cluster validation: Here we compare the clustering results obtained against known properties of the dataset like the number of classes existing and the method is useful in selecting the right clustering algorithm for a specific dataset about which prior knowledge is available.
3. Internal Cluster Validation: Here we use the clusters generated and for each cluster, we consider only the objects belonging there to evaluate the goodness of the cluster. The method helps in identifying the number of clusters and the suitable algorithm for clustering the dataset.
4. Cluster stability validation: Here the stability of clustered result is ascertained by re-clustering the data by removing one attribute at a time and see the impact of each attribute on clustering.

In this paper, we focus on relevance of external and internal measures for cluster validation and optimization by applying different clustering methods on popular UCI datasets.

EXTERNAL CLUSTER VALIDATION

Evaluating clusters externally involves known properties of the dataset. Hence, these measures are applicable only if the ground truths are available. Usually, there are many external criteria [6] for assessing clustering quality. Purity and entropy are well received external evaluation measure.

PURITY

Purity measures the ability of clustering approach, using the known labels even when the number of clusters differ from the given number of class labels. Purity can be measured in terms of majority of available class labels in a cluster and cluster size. The term purity can be defined as,

$$\text{Purity} = 1/N \sum_{i=1}^{k} \max |c_i \cap t_j| \quad (1)$$

Where, N is the number of objects (data points) in the data set, k, the number of clusters, c_i cluster in C, and t_j is the classification label which has the max count for cluster c_i. If all the points in the cluster have the same label, then the purity is 1. Purity ranges from 0 to 1, the higher values imply better purity.

ENTROPY

Entropy assess the purity of the clusters using the known class labels. In a cluster, if all the objects do have the same class label, then entropy is said to be zero. Entropy increases as the number of class labels in a cluster increase. For each cluster j, compute the probability (P_{ij}) that a member of a cluster j belongs to a class i. Then the entropy of all objects in cluster j is calculated. The total entropy of all the clusters is calculated as the weighted sum of all entropies of all clusters in the dataset.

$$E_j = \sum_i p_{ij} \log(p_{ij}) \quad (2)$$

$$E = \sum_{j=1}^{m} \frac{n_j}{n} E_j \quad (3)$$

INTERNAL CLUSTER VALIDATION

The internal cluster validation depends on compactness, cohesion, separation and how the objects in the dataset are connected. Compactness gives information on how closely the objects in a cluster are connected. A lower within cluster variation is ideal for a good compact cluster. Cohesion is the measure of sum of weights of all links within a cluster and ideally a high value for cohesion indicates a good cluster. Separation is measured in terms of how separated the cluster centers are and sometimes the pair wise distance between objects in different clusters is taken as a measure of separation. Separation needs to be high for a good partition. Connectivity is another measure showing the quality of clustering. It shows how closely objects in a cluster are placed when compared with objects of another cluster. Connectivity of objects across clusters needs to be minimized.

For measuring internal cluster quality, a large number of cluster validity indices are proposed. In this work, we selected the following popular cluster validation measures for detailed study.

BH GAMMA INDEX

The BH Gamma index uses an adaptation of Goodman and Kriskal's Gamma Statistic [7] in clustering environment. Here the comparison is between all within cluster dissimilarities and all between cluster dissimilarities. A comparison is said to be concordant if the within cluster dissimilarity is strictly less than the between cluster dissimilarity and it is discordant if within cluster dissimilarity is greater than the between cluster dissimilarity [Gordon 1999]. Ignoring the possible equality, the index is defined as,

$$Gamma = \frac{s(+) - s(-)}{s(+) + s(-)} \quad (4)$$

Where, s (+) is the number of concordant comparisons, and s (-) is the number of discordant comparisons.

DUNN INDEX

The Dunn index [8] defines the ratio between the minimal inter-cluster distance to maximal intra-cluster distance. This index is given by the equation,

$$Dunn = \frac{\min\limits_{1<i<j\leq q} d(C_i, C_j)}{\max\limits_{1\leq k\leq q} diam(C_k)} \quad (5)$$

The dissimilarity function between two clusters C_i and C_j, $d(C_i, C_j)$, is defined as,

$$d(C_i, C_j) = \min_{x\in C_i, y\in C_j} d(x, y) \quad (6)$$

and $iam(C)$, the diameter of a cluster, can be considered as a measure of cluster dispersion. The diameter of a cluster C is defined as,

$$diam(C) = \max_{x,y\in C} d(x, y) \quad (7)$$

CALINSKI-HARABASZ (CH) INDEX

The Calinski-Harabasz (CH) [9] is calculated as the ratio of Between Cluster Dispersion to the Within Cluster Dispersion. The calculation is done in two steps.

For calculating the Between Cluster Dispersion (BCD), the mean value of each cluster is found. The sum of squares of deviations from the mean value of data set and the cluster mean is calculated for all the clusters. Multiply the obtained value with the number of elements in that cluster. The process is continued for all the clusters in the partition.

The Within Cluster Dispersion (WCD) is calculated as the sum of squares of deviation of each element in a cluster with the corresponding cluster mean. The WCD calculation is performed for all the clusters in the partition.

The CH index is calculated as the ratio of BCD to WCD. The ideal value for CH index is in the range zero to infinity and it is ideally the maximum for the optimal number of cluster for the dataset.

$$CH_q = \frac{trace(B_q)/(q-1)}{trace(W_q)/(n-q)} \tag{8}$$

GDI INDEX

The GDI [10] stands for the Generalized Dunn's Index and it is a measure depending on the between cluster distance and within cluster distance. Denoting the between cluster distance using the symbol δ and the cluster diameter using Δ, the equation 4 is modified as,

$$C = \frac{\min\limits_{k \neq k'} \delta(C_k, C_{k'})}{\max\limits_{k} \Delta(C_{k'})} \tag{9}$$

With 1≤k≤K and 1≤k'≤K

RATKOWSKY LANCE INDEX

Ratkowsky and Lance [11] proposed the index for identifying the optimal number of clusters for datasets. The measure is calculated using sum of squares between the clusters for each object and the total sum of squares of each object within the cluster. The index is calculated as,

$$Ratkowsky = \frac{\bar{S}}{q^{1/2}} \tag{10}$$

Where,

$$\bar{S}^2 = \frac{1}{p}\sum_{j=1}^{p} \frac{BGSS_j}{TSS_j} \tag{11}$$

$$BGSS_j = \sum_{k=1}^{q} n_k \left(c_{kj} - \bar{x}_j\right)^2 \tag{12}$$

$$TSS_j = \sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j\right)^2 \tag{13}$$

TAU INDEX

The Tau index [12] in the context of clustering validation is defined as,

$$Tau = \frac{s(+) - s(-)}{[(N_t\,(N_t-1)/2 - t)(N_t(N_t-1)/2)]^{1/2}} \tag{14}$$

where $s+$ and $s-$ are defined the same as in the Gamma index, N_t is the total number of distances, t is the number of comparisons of two pairs of points when both pairs are within-cluster or in different cluster.

PBI INDEX

The Point Biserial Index popularly known as PBI [13] [14] [15] is calculated with the help of two matrices. The first is the dissimilarity matrix and the second is matrix with entries 0 or 1 depending on whether the two objects are clustered together (0) or in different clusters (1). The index is calculated as follows,

$$Ptbiserial = \frac{[\bar{S}_b - \bar{s}_w][N_w N_b / N_t^2]^{1/2}}{s_d} \tag{15}$$

Where N_t is the total number of pairs of observations in the data set.

$$N_t = \frac{n(n-1)}{2} \tag{16}$$

N_w is the total number of pairs of observations belonging to the same cluster:

$$N_w = \sum_{k=1}^{q} \frac{n_k(n_k-1)}{2} \tag{17}$$

N_b is the total number of pairs of observations belonging to different clusters:

$$N_b = N_t - N_w \tag{18}$$

S_w is the sum of within cluster distances:

$$S_w = \sum_{k=1}^{q} \sum_{\substack{i,j \in c_k \\ i<j}} d(x_i, x_j) \tag{19}$$

S_b is the sum of between cluster distances:

$$S_b = \sum_{k=1}^{q-1} \sum_{l=k+1}^{q} \sum_{\substack{i,j \in c_k \\ i<j}} d(x_i, x_j) \tag{19}$$

PBM INDEX

The PBM [16] index is calculated by using the distance between the points and the new centroids (barycenter) and the distances between centroids themselves. The equation is,

$$C = \left(\frac{1}{K} \times \frac{E_T}{E_W} \times D_B\right)^2 \tag{20}$$

Where D_B is the distance largest distance between two cluster barycenters calculated as,

$$D_B = \max_{k<k'} d\left(G^{\{k\}}, G^{\{k'\}}\right) \quad (21)$$

E_W is the sum of the distance of the points of each cluster to their barycentre.

$$E_W = \sum_{k=1}^{K} \sum_{i \in I_k} d\left(M_i, G^{\{k\}}\right) \quad (22)$$

And E_T the sum of all distances of all the points to the barycenter G of the entire dataset.

$$E_T = \sum_{i=1}^{N} d(M_i, G) \quad (23)$$

WEMMERT GANCARSKI INDEX

The Wemmert Gancarski [17] index is also calculated using the concept of distances between the points and the barycenters of all the clusters.

For calculating the index, we use the following terms. If M is a point in a cluster C_k, the term R(M) is defined as the quotient of distance of this point to the barycenter of the cluster to which it is a member and the smallest distance of this point to the barycenters of other clusters,

$$R(M) = \frac{\left\|M - G^{\{k\}}\right\|}{\min_{k \neq k'} \left\|M - G^{\{k'\}}\right\|} \quad (24)$$

The mean value of this quotient is then calculated for each cluster, if the mean value is greater than 1, the value is ignored, otherwise its complement to 1 is calculated. It is done using the equation

$$J_k = max\left\{0, \quad 1 - \frac{1}{n_k} \sum_{i \in I_k} R(M_i)\right\} \quad (25)$$

The Wemmert Gancarski index is calculated as the weighted mean of J_K for all the clusters.

$$C = \frac{1}{N} \sum_{k=1}^{K} n_k J_k \quad (26)$$

The complete equation can be rewritten as,

$$C = \frac{1}{N} \sum_{k=1}^{K} max\left\{0, \; n_k - \sum_{i \in I_k} R(M_i)\right\} \quad (27)$$

LITERATURE REVIEW

- The study was made on four synthetic datasets and one real time dataset. K-Means and CURE were considered as the clustering algorithms. The study revealed that the essence of clustering was not resolved with the Cluster Validation Indexes (CVI) but found to be helpful in suggesting some indication on which further research to be carried out [18].

Table 1: Summary of the literature review

No	Authors	Year of publication	CVI measures studied
1	Maria Halkidi, Michalis Vazirgiannis and Yannis Batistakis [18]	2001	R Squared (RS), Root Mean Squared Standard Deviation (RMSSTD), Davies-Bouldin (DB), SD
2	Maria Halkidi, Michalis Vazirgiannis and Yannis Batistakis [19]	2002	Rand Statistic, Jaccard coefficient Folkes-Mallows(F_M) index, Hubert Γ statistic normalized Γ statistic
3	Maria Halkidi, Michalis Vazirgiannis and Yannis Batistakis [20]	2002	Modified Hubert Γstatistic, Dunn indices, DB index, RMSSTD, RS, Distance between Two clusters(CD), Semi Partial R Squared (SPR), SD index, Sdbw index
4	Ujjwal Maulik and Sanghamitra Bandyopadhyay [21]	2002	DB, Dunn,Callinski _Harabasz (CH) , I Index
5	Hui Xiong, Junjie Wu and Jian Chen [22]	2006	Entropy
6	Yanch Liu, Zhongmou Li, Hui Xiong, Xuedong Ga and Junjie Wu [23]	2010	RMSSTD, RS, Modified Hubert Γstatistic, CH, I Index, Dunn, Silhouette, DB, Xie _Beni (XB) SD, SDBW
7	Jonathan Baarsch and M. Emre [24]	2012	Dunn, DB, CH, Silhouette, Point Biserial (PBI, PBM, Sum of Squares
8	L. Guerra, V. Robles, C. Bielza and P. Larrañaga. [25]	2012	Silhouette, CH, C index, DB, Gamma and Adjusted Rand Index
9	Olatz Arbelaitz, Ibai Gurrutxaga, Javier Muguerzaa and Jesús M. Pérez [3]	2013	Rand, Adjusted Rand, Jaccard, F-M and 30 CV index for internal evaluation
10	M. Hassani and T. Seidl [26]	2017	RMSSTD, RS, Modified Hubert Γstatistic, CH, I index, SDBw, DB, Dunn, SD, Silhouette, XB

- The experimental study on synthetic dataset suggested that CV indices were not capable of handling arbitrary shaped clusters and they suggested further research to develop an integrated data mining results quality assessment model to deal with various types of clustering. Stressed on the importance of having additional measures to be introduced for assessing cluster quality[19].

- The main focus was on cluster quality based on relative measures for crisp clustering. K-Means and CURE algorithms were used for partitional clustering. The Hierarchical approach and Fuzzy clustering were also used in experimentation, they studied the impact of CV indices on giving an indication on suitable clustering algorithm for a specific dataset. The study covered four two dimensional and a six-dimensional synthetic dataset containing three clusters each and concluded that CV indices worked better when the clusters were compact but not highly recommended for clusters of arbitrary shapes. Suggested more research to be carried out on introduction of more CV indices[20].

- For clustering the datasets, K Means, Single linkage hierarchical clustering and Simulated Annealing (SA) were used. For experimentation, three two-dimensional, one three-dimensional synthetic dataset, and two real time datasets were used. The study revealed I index as the best measure for five datasets but recommended an extensive study on CV measures and on convergence speed of algorithms [21].

- The study was on impact of entropy on K-Means like algorithms and used the K-Means implementation in CLUTO package. Real time datasets on documentation, Biological datasets and UCI datasets were clustered and analyzed. Found that the results of K-Means on high variant datasets to be far away from the true clusters, but generally entropy was recommended as a suitable measure for K-Means clustering [22].

- Detailed study on 11 internal measures based on five aspects of clustering (noise, density, monotonicity, sub clusters and skewed distribution) found the measure SDBW to outperform other measures [23].

- The study undertook on 125 artificially generated datasets, found the sum of squares measure as the most productive one for predicting an appropriate value for the number of clusters. The silhouette index was recommended as a good alternative, followed by CH and DB indices. K-Means with farthest first approach was used in clustering and they found the irregularities of K-Means (clusters do not reflect natural structure), variation of indices in measuring compaction and separation, contributing heavily in analyzing the clustering results. Suggested more work to be carried out on using validity measures as interpretive tools in clustering data which are ambiguous and difficult to visualize[24].

- The study covered 36 datasets with strong and distinct clusters, 48 datasets with noise and a few real time datasets. The algorithms used are K Means, Hierarchical clustering (ward's method) and Expectation Maximization (EM). The measure CH was found to be better for low noise data and for lesser number of dimensions. Gamma index was not found to be affected by noisy data, silhouette was less susceptible to noise but was affected by outliers. Hierarchical clustering produced better clustering than K-Means and EM. C index was the measure found to be less effective. They recommended the study on better CV indices to handle outliers and noise in data and also missing data [25].

- This study was found to be the most comprehensive one in terms of datasets, CVI measures and clustering approaches. Studied three external validity measures (Jaccard, Adjusted Rand and VI) and 30 internal cluster validation measures. The study was on K-Means clustering, Hierarchical clustering (ward and average linkage) on 10 synthetic datasets with 72 different configurations and 20 real-time datasets from the UCI repository. The study revealed that no CVI measure was universally acceptable and heavily dependent on the dataset. The measures behaved unpredictably for overlapping clusters. In the case of synthetic datasets, most of the measures failed for K means, though CH index was found to be comparatively better. The silhouette index was found to be the best for hierarchical clustering. For lesser number of clusters to be formed, CVI indices provided better results and the number of dimensions did not have much significance. For real time datasets, Dunn index and its variants provided the best possible results. In the case of external measures, the Jaccard and VI measures behaved in a similar manner, but adjusted rand index behaved in contradictory manner. Results for synthetic and real time datasets were qualitatively similar, although some disagreements were there for some measures [3].

- One of the recent studies on CVI measures concentrated on streaming input data. K-Means and Density based clustering algorithms were used and the results analyzed with the help of 11 CV indices. CH index was found to be suitable for K-Means and revised validity index for density-based approach. For static datasets, CH and I index gave poor performance and CH was found to be vulnerable to noise. SDBW index behaved well for static datasets but failed to handle common errors of stream clusters. The revised SDBW showed adaptability to density-based clustering. For static data, XB index was also found to be dependable to some extent [26].

CLUSTERING ALGORITHMS

Clustering is one of the primary knowledge discovery techniques helpful in discovering meaningful structures in datasets. Clustering results in partitioning datasets into subgroups having maximum intra group similarity with groups ideally well separated. In this analysis, two partitional clustering algorithms named K-Means and K-Medoids are implemented.

K-MEANS CLUSTERING

The K-Means algorithm [27] begins its execution by choosing K random points from the dataset or the user has the freedom of selecting any K points to form the set of initial seed points. Considering these points as the initial cluster centroids, the algorithm calculates the deviation of each data point from the centroids. Points are assigned to belong to subgroups based on their adjacency to the selected centroids. The formed subgroups are then refined by calculating the mean value of all points in respective sub groups as the new centroid. The deviation of all data points in the dataset to the new cluster centroids is again calculated and clustering is done again. The procedure continues till all the clusters become stable, ensuring that there is no across movement of points belonging to a cluster [28].

The K-Means algorithm is known to have different variants like K-Means (Forgy), K-Means(Lloyd), K-Means (Hartigon) and K-Means (Mc Queen). The variation is critical on the distance measure used to find the adjacency of data points and the centroid.

The K-Means (Lloyd) method [29]: This is the classical K-Means method, in which given some initial centroids, repeatedly compute the mean value and assign points to the closest centers. Cluster assignments are re-estimated and centroids are updated on each iteration. The algorithm repeatedly computes the distance measures every time, though points assigned to a cluster rarely changes its cluster membership.

The K-Means (Hartigon) Method [30] provides a more sophisticated approach to perform K-Means. The algorithm proceeds point by point to determine its optimal cluster assignment. The method takes into account the movement of the means resulting from the reassignment—for obtaining the optimal cluster, it may reassign a point to another cluster, even if it is already assigned to the closest center. This may guarantee a better cluster, but computationally expensive.

The K-Means (Forgy) method [31], the Forgy and Lloyds methods are both batch offline models, but basic difference is that the Lloyd's algorithm considers the data discrete distribution while the Forgy algorithm considers the continuous distribution. Otherwise, they use the same procedure for computation of clusters.

The K-Means (Mc Queen method) [27]: The Mc Queen algorithm (1967) is also an iterative algorithm. On comparing with Forgy/Lloyd's algorithm, here the centroids are recalculated every time a data point changes the subgroup to which it was a member. Considering each data point one by one, if the centroid of the subgroup it currently belongs to is the nearest, no change is made in the subgroup. If another centroid is the closest, the data point is reassigned to the other centroid and the centroids for both the old and new subgroups are recalculated as the mean.

EMPIRICAL EVALUATION

R LANGUAGE

R is a programming language introduced and monitored by the R foundation for statistical computing [32]. Initially coded and developed by Ross Ihaka and Robert Gentleman at the Department of Statistics of the University of Auckland, New Zealand, R got introduced in 1993. It is an open source free software popularly used by statisticians and data miners due to the facilities for data analysis, high quality graphical support and reporting. R is a GNU package, freely available under GNU general public license, whose source code is written in C and Fortran.

Being a programming language, it has all the fundamental capabilities for decision making, constructing loops, creating user defined functions and it supports high end powers for input and output of various types of data. R allows users to add their own features and facilities to the system library and codes from other programming languages can be linked and executed at run time. R is thus a fully planned coherent programming system as compared to many of the available data analysis software.

DATASET

For our studies, we have taken some of the standard datasets available in the UCI data repository. The properties of the data set can be summarized as in the following table.

Table 2: Details of the selected dataset

Slno	Name of dataset	Number of instances	Number of attributes	Number of known classes
1	*Breast tissue*	*106*	10	6
2	*Ecoli*	*336*	8	8
3	*Iris*	*150*	4	3
4	*Parkinson*	*195*	23	2
5	*Glass*	*214*	10	7

RESULTS

In this section, we narrate the detailed analysis conducted to investigate the cluster validity indices discussed in the previous section. The comparative analysis includes the implementation of traditional partitional clustering algorithm namely K-Means and its variants on different UCI datasets and the analysis of different cluster validation measures on the generated clusters.

The relative analysis used in this experiment differs from the conventional approaches, where the number of clusters has to be mentioned in advance for each dataset. Here all the algorithms are executed with the number of clusters ranging from 2 to $\sqrt{n}$, where n is the number of elements in each dataset, resulting in the formation of different partitions in

different iterations. Then the different measures are calculated for every partition and detailed analysis is carried out to attain information on the ideal clustering for that particular dataset.

In *Fig. 1.* the performance of the discussed measures is plotted for the dataset Iris using K-Means algorithm. It is known that the iris dataset contains 3 classes. From the graphs, it can be noted that most of the measures behave in a similar manner as the number of clusters are increasing. The PBI index is giving better value as the number of clusters increase, touching the peak when k = 12. Dunn index is fluctuating as the number of clusters are increasing. CH, PBM, Tau and GDI index (except Hartigon method) are found to behave in a similar manner, gradually decreasing after k=4. The PBI index and Ratkowsky index (to some extent) performed in a contradictory manner to all other indices. Analyzing the graphs, a value of 4 or less can be the ideal for the number of clusters for the Iris dataset.

E coli dataset

The Ecoli dataset has eight classes. In Fig.2. BH Gamma Index shows indifferent behavior, fluctuating until it reaches k = 10 and after that it gradually increases with the increase in number of clusters. The PBI Index is gradually increasing with the increase in the number of clusters while CH Index, GDI Index, Wemmert Index, Tau Index and PBM Index are gradually decreasing with the number of clusters. Looking at the graph, no index attains a maximum value in the neighborhood of eight, the number of classes. Measures are not helping in suggesting optimal number of clusters for the Ecoli dataset.

In the case of Breast tissue dataset, the available number of classes is six. The performance of different cluster validity measures is given in Fig. 3. In this case, the clustering obtained with the method Mac Queen is far apart from the other K-Means variants and the performance of measures based on this particular clustering behaves differently from others. Analyzing the graphs lead to the following conclusions. For the Lloyd and Forgy methods, Dunn index and BH gamma index attain the maximum value in the neighborhood of six. The same indices are giving contradictory values for Hartigon and Macqueen methods. PBI index for Forgy has a maximum value at k=6. The Wemmert index for Lloyd method has a maximum value at k= 6 and the same index for Forgy has its peak value at k=7. For Forgy and Hartigon methods, the Ratkowsky index attains its peak k=6. On overall analysis, no CVI index is found to be suitable for suggesting optimum value of clusters for the Breast tissue dataset.

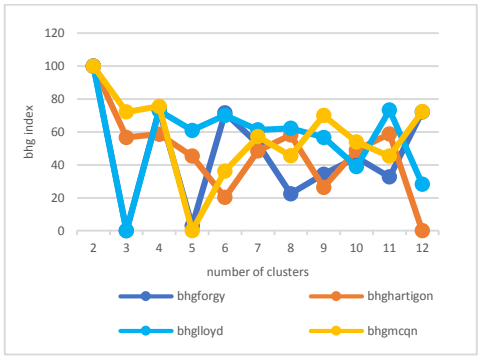
The glass dataset has seven classes. On analyzing the graph, no index is found to be acceptable across all the methods of K Means. BH Gamma and PBM index are observed to attain the peak value in the neighborhood of seven, but for only some methods. PBI and Ratkowsky indices has their maximum value at seven for the Lloyd method. These indices behave indifferently for other methods.

The Parkinsons dataset has two classes. The CH index, PBM index and Wemmert index are almost identical and has their peak value at two. The Tau index almost resemble their performance. The BH index and GDI index are found to be fluctuating throughout. PBI and Ratkowski index behave in contradiction to other measures.

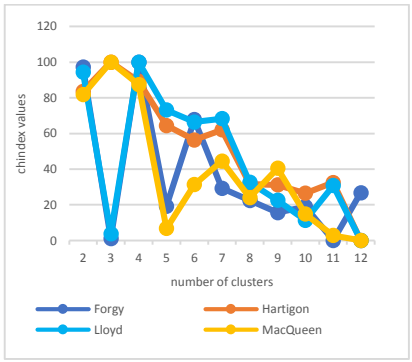
CONCLUSION

Cluster analysis is one of the major areas in Data Mining. There exists a large number of clustering algorithms available but analyzing the cluster quality is a challenging task. There are many number of cluster validity indices available, but none of them can be used as a universal solution. The cluster validity indices are behaving differently on different datasets. In the era of big data, analyzing the quality of the clusters formed from the real-time data is a major research challenge which needs to be further explored by the researchers. The supervised methods including classification incorporates the class labels along with the attributes in the analysis. Since, clustering is purely based on the attributes, we cannot predict clustering algorithms to perform in a similar manner. Hence, assessing the cluster quality is very significant. We recommend that there is much scope for further research in this area for developing an internal cluster validity measure which will estimate the quality of clusters accurately.

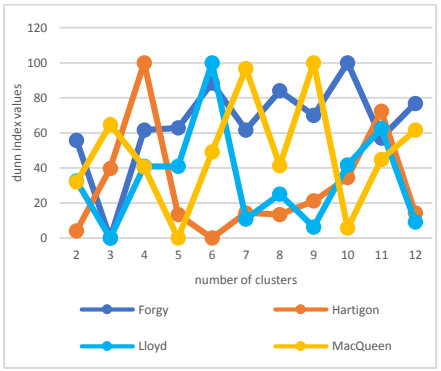
The K-Means algorithm is a robust clustering algorithm capable of handling any type of data, irrespective of the number of attributes, their type and number of observations in the dataset. The algorithm cluster the data into user specified number of clusters. A conclusion on optimal number of clusters for the dataset is difficult and it is practically impossible to arrive at a globally acceptable generalization. The study reveals that no particular cluster validation index is competent enough to predict the suitable number of clusters for any dataset, though there are some measures found to be suitable for some datasets.



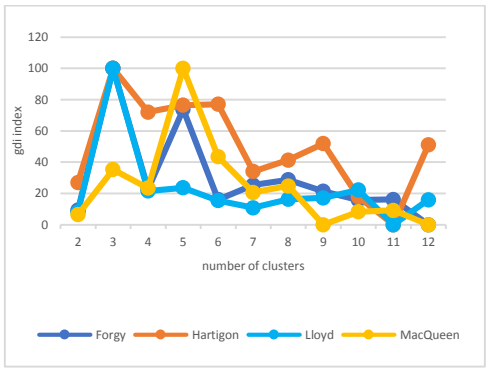
BH gamma index



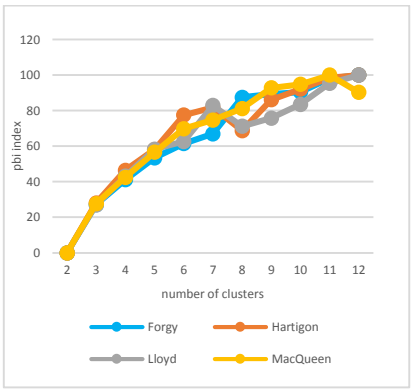
CH index



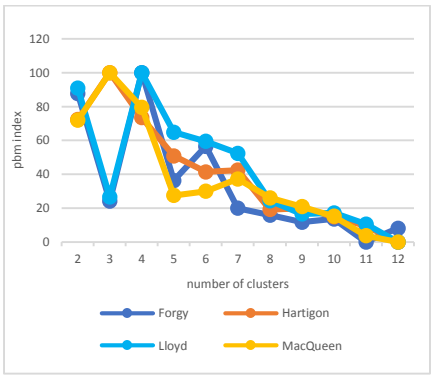
Dunn index



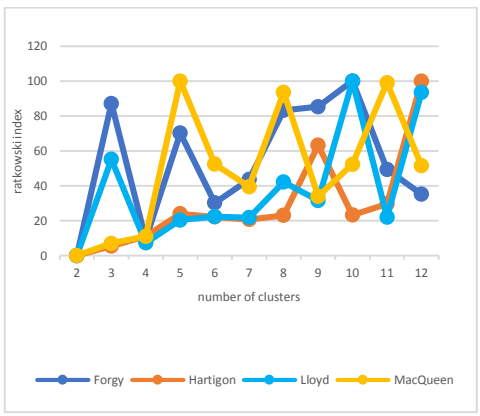
GDI Index



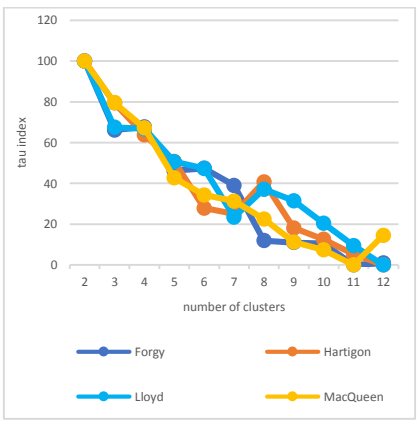
PBI index



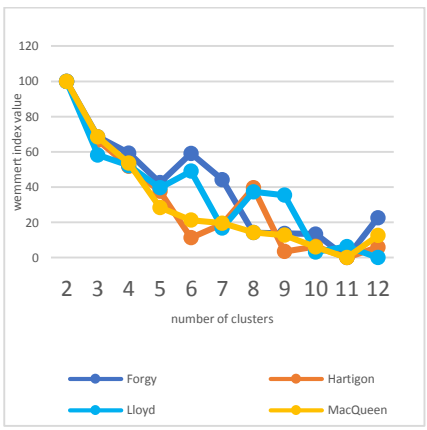
PBM index



Ratkowski index

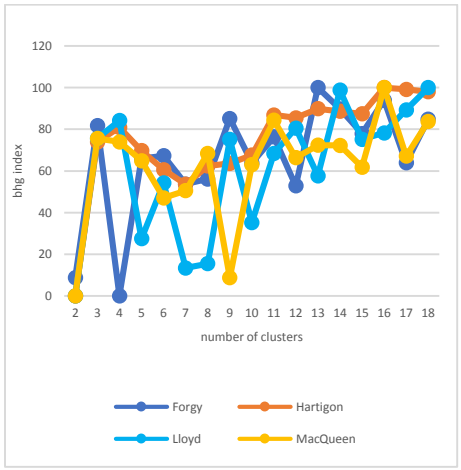


Tau index

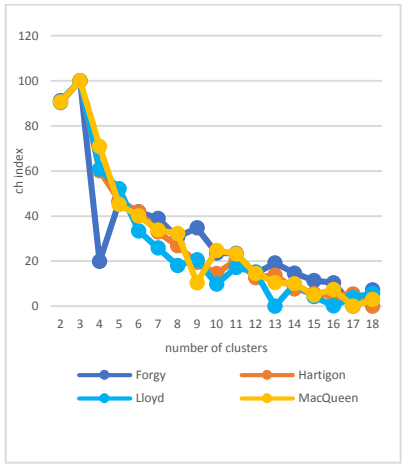


Wemmert index

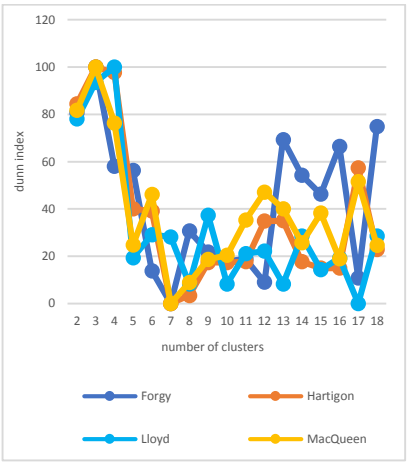
Figure 1. Performance of different cluster validity indices on Iris dataset using K-Means



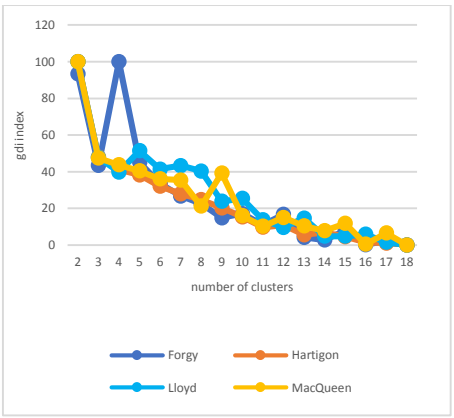
BH gamma index



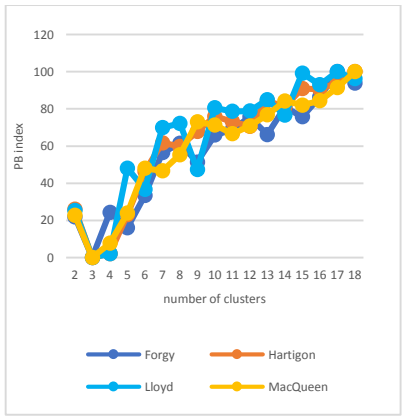
CH index



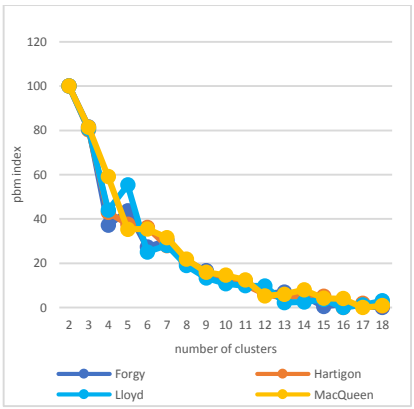
Dunn index



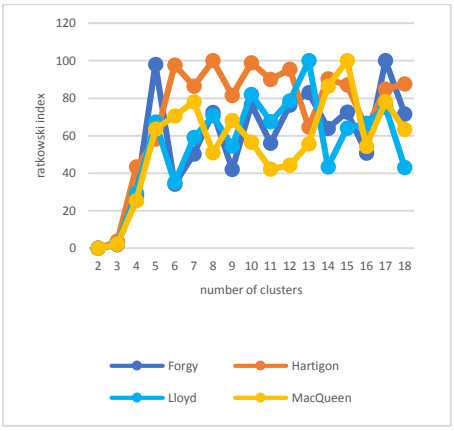
GDI Index



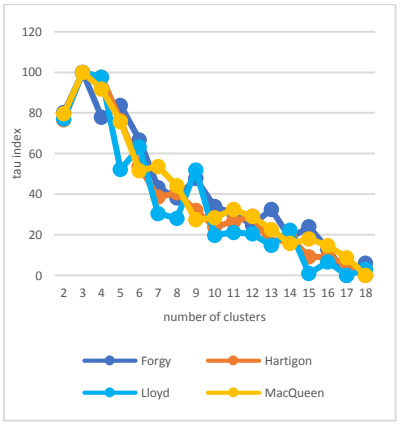
PBI index



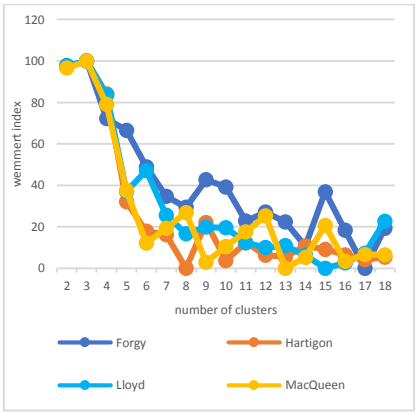
PBM index



Ratkowski index

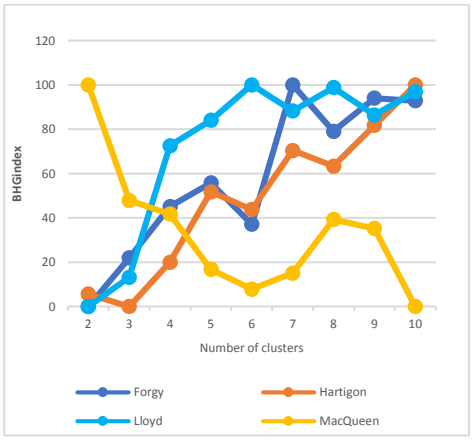


Tau index

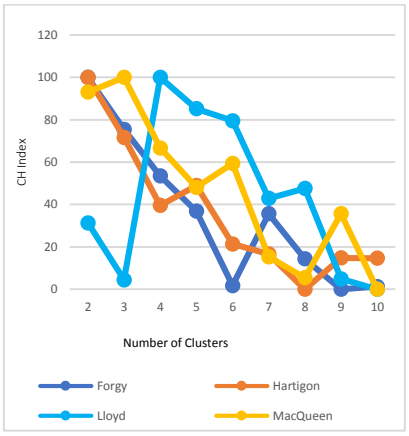


Wemmert index

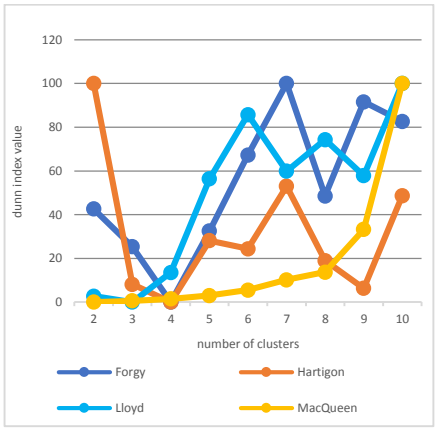
Figure 2. Performance of different cluster validity indices on Ecoli dataset using K-Means



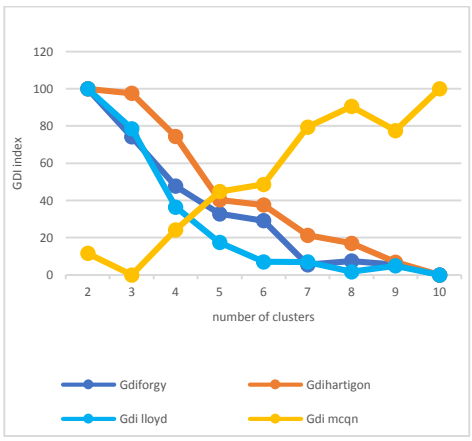
BH gamma index



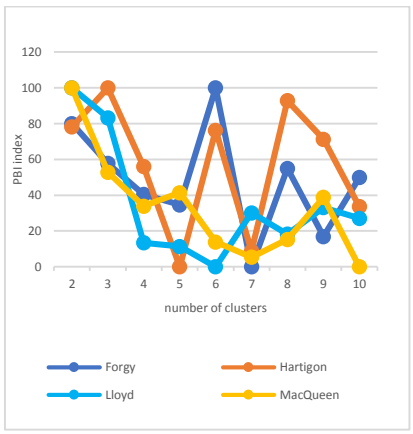
CH index



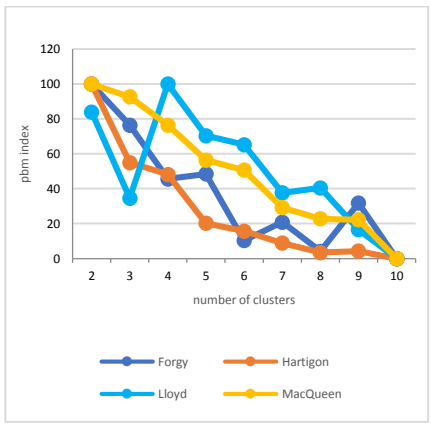
Dunn index



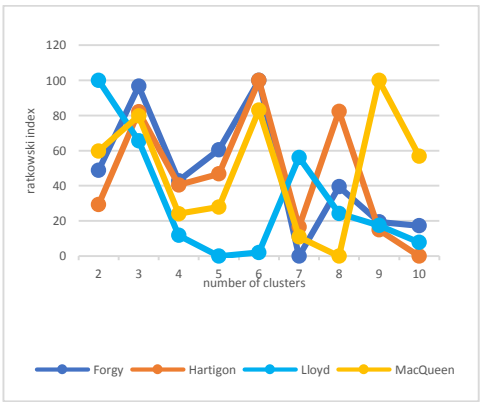
GDI Index



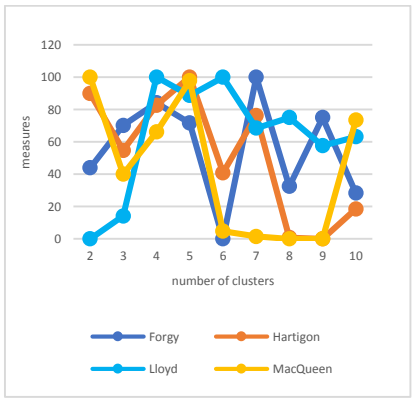
PBI index



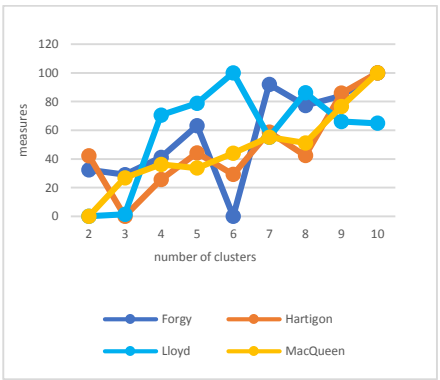
PBM index



Ratkowski index

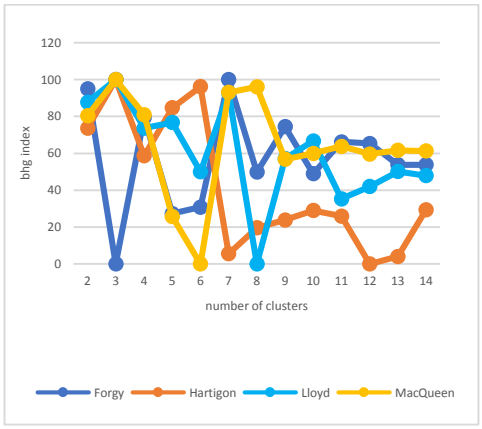


Tau index

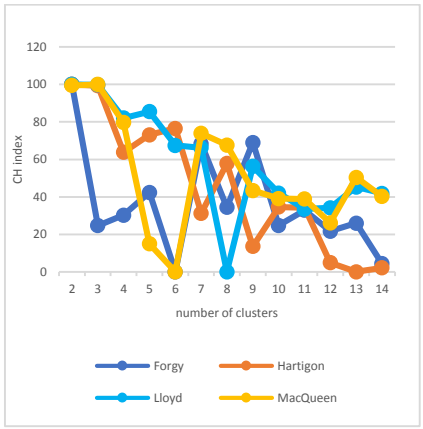


Wemmert index

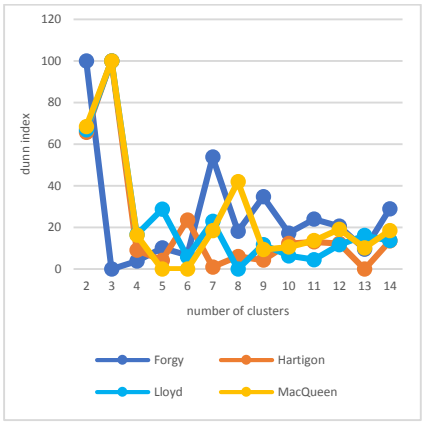
Figure 3. Performance of different cluster validity indices on Breast Tissue dataset using K-Means



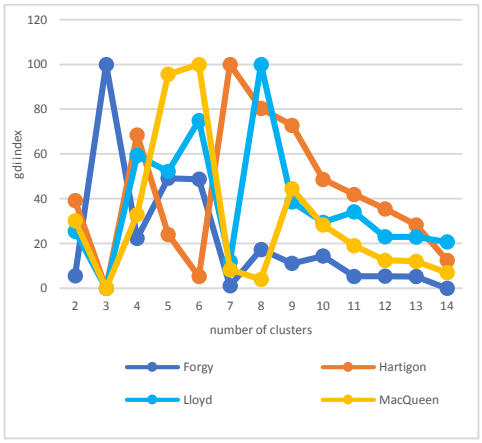
BH gamma index



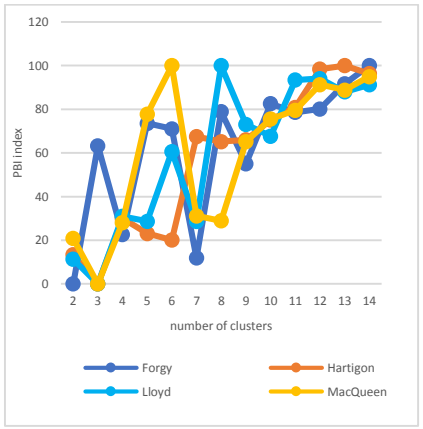
CH index



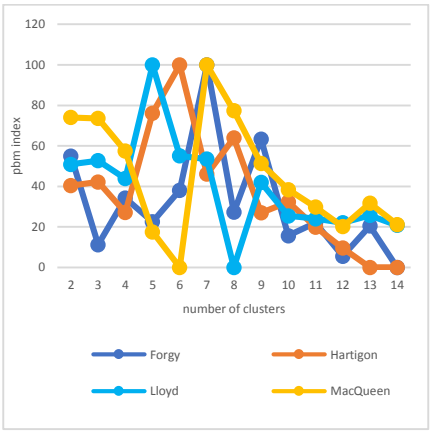
Dunn index



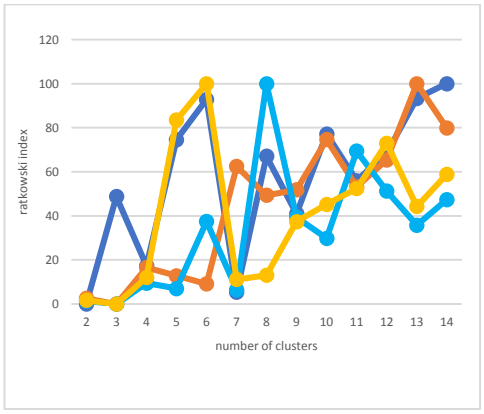
GDI Index



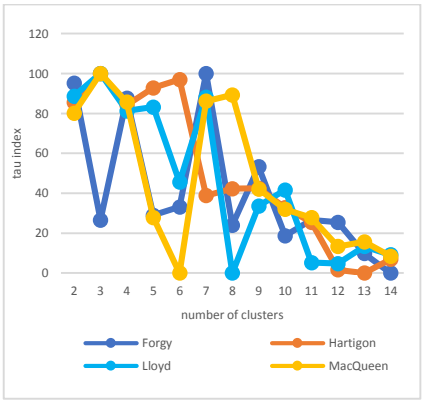
PBI index



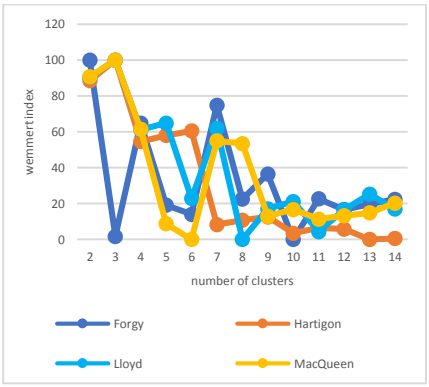
PBM index



Ratkowski index

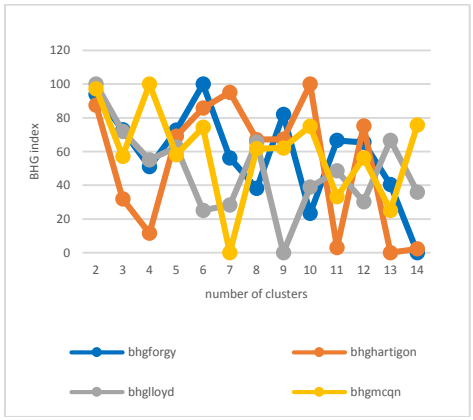


Tau index

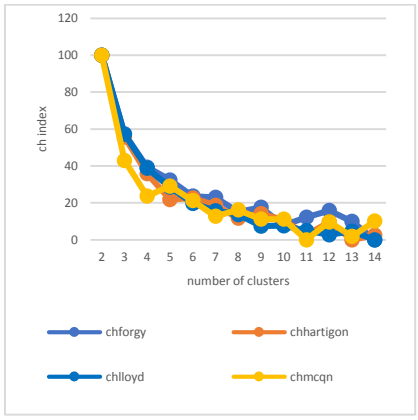


Wemmert index

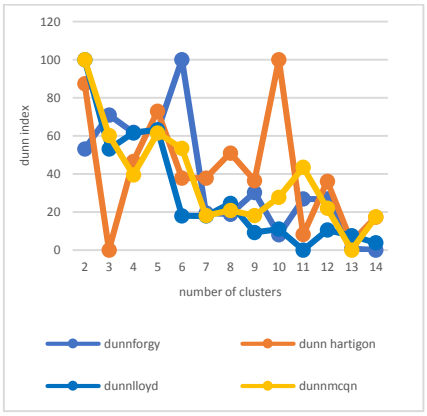
Figure 4. Performance of different cluster validity indices on glass dataset using K-Means



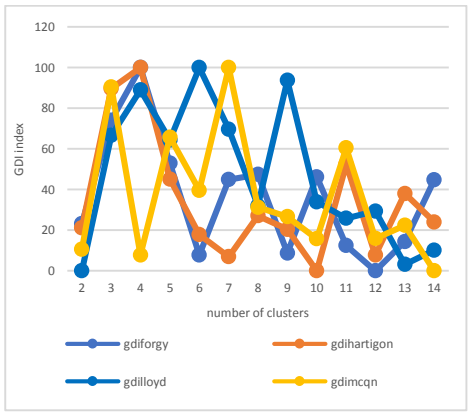
BH gamma index



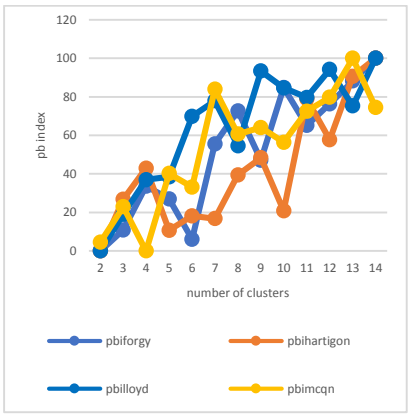
CH index



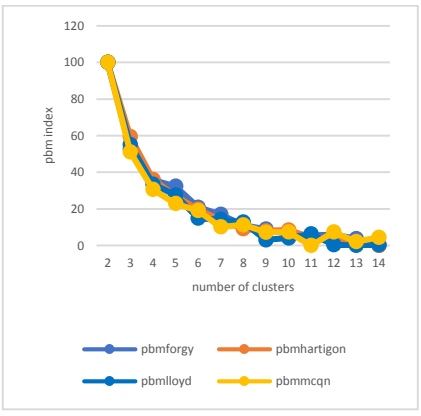
Dunn index



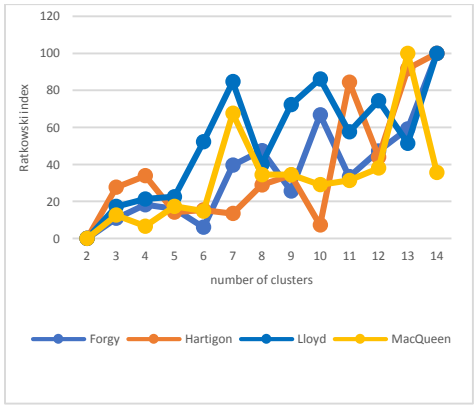
GDI Index



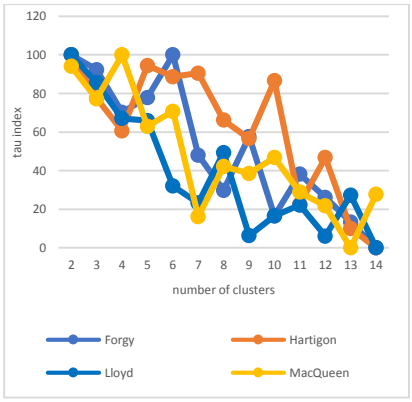
PBI index



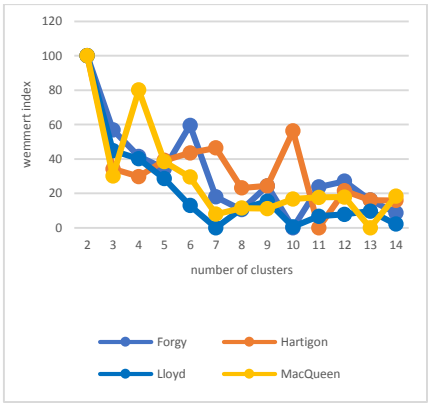
PBM index



Ratkowski index



Tau index



Wemmert index

Figure 5. Performance of different cluster validity indices on parkinson dataset using K-Means

REFERENCES

[1] G. K. Gupta. Introduction to Data Mining With Case Studies.3rd ed., PHI Learning Pvt. Ltd, 2014.

[2] A. K. Jain, M. N. Murty, P. J. Flynn. Data clustering: a review, ACM Computing Surveys (CSUR), vol.31, no.3, pp. 264-323, Sept. 1999.

[3] Olatz . Arbelaitz, Ibai. Gurrutxaga, Javier. Muguerza, Jesús. M. Pérez, Iñigo. Perona. "An extensive comparative study of cluster validity indices". Pattern Recognition.vol. 46, no.1, pp. 243-256, Jan 2013.

[4] Shirkhorshidi AS, Aghabozorgi S, Wah TY (2015) "A Comparison Study on Similarity and Dissimilarity

Measures in Clustering Continuous Data". PLoS ONE 10(12): e0144059. https://doi.org/10.1371/journal.pone.0144059

[5] Sergios. Theodoridis and Konstantinos. Koutroumbas, Pattern Recognition.4 th ed., Elsevier, 2008.

[6] Eréndira.Rendón, Itzel.Abundez, Alejandra. Arizmendi, Elvia M. Quiroz. "Internal versus External cluster validation indexes". International journal of computers and communications, vol.5, no.1, pp.27-34, 2011.

[7] F. B. Baker and L. J. Hubert. "Measuring the power of hierarchical cluster analysis". Journal of the American Statistical Association, 70:31-38, 1975.

[8] J. C. Dunn . "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters". Journal of C1ybernetics. Vol. 3, no.3, pp.32-57, Sep 1973.

[9] T. Calinski, and J. Harabasz. "A dendrite method for cluster analysis". Journal of Communications in Statistics.vol. 3, pp. 1-27, 1974.

[10] Maria. Halkidi, Michalis. Vazirgiannis, Yannis. Batistakis. "Quality Scheme Assessment in the Clustering Process", Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery. pp.265-276, 2000.

[11] Ratkowsky DA, Lance GN (1978). "A Criterion for Determining the Number of Groups in a Classiffication." Australian Computer Journal, 10(3), 115-117.

[12] F. James Rohlf "Methods of comparing classifications. Annual Review of Ecology and Systematics" 1974 5:1, 101-113

[13] Milligan GW (1980). "An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms." Psyc hometrika, 45(3), 325–342.

[14] Milligan GW (1981). "A Monte Carlo Study of Thirty Internal Criterion Measures for Cluster Analysis." Psychometrika, 46(2), 187–199.

[15] Kraemer HC (1982). Biserial Correlation. John Wiley & Sons. Reference taken from a SAS note about the BISERIAL macro on this Web Site: http://support.sas.com/kb/ 24/991.html.

[16] Bandyopadhyay S. Pakhira M. K. and Maulik U. "Validity index for crisp and fuzzy clusters". Pattern Recognition, 37:487{501, 2004.

[17] Desgraupes, B. "ClusterCrit: Clustering Indices". R Package Version 1.2.3., 2013. Available online: https://cran.r-proje ct.org/web/packages/clusterCrit/ (accessed on 6 September 2017).

[18] Maria. Halkidi, Yannis. Batistakis, Michalis. Vazirgiannis. "On Clustering Validation Techniques". Journal of Intelligent Information Systems. Vol.17, no 3, pp. 107–145, Dec 2001.

[19] Maria. Halkidi, Yannis. Batistakis, Michalis. Vazirgiannis. "Cluster Validity Methods: Part I", ACM SIGMOD Record, Vol. 31, No. 2, pp. 40-45, Jun 2002.

[20] Maria. Halkidi, Yannis. Batistakis, Michalis. Vazirgiannis. "Cluster Validity Methods: Part II", ACM SIGMOD Record, Vol. 31, No. 3, pp. 19-27, Sep 2002.

[21] Ujjwal. Maulik and Sanghamitra. Bandyopadhyay. "Performance Evaluation of Some Clustering Algorithms and Validity Indices". IEEE Transactions On Pattern Analysis And Machine Intelligence.vol. 24, no. 12, pp. 1650-1654, DEC 2002.

[22] Hui. Xiong, Junjie .Wu , Jian. Chen. "K-means clustering versus validation measures: a data distribution perspective". Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 779-784, Aug 2006.

[23] Yanch. Liu, Zhongmou. Li, Hui. Xiong, Xuedong. Ga, Junjie Wu. "Understanding of Internal Clustering Validation Measures". IEEE International Conference on Data Mining, pp. 911-916, 2010.

[24] Jonathan. Baarsch and M. Emre. Celebi. "Investigation of Internal Validity Measures for K-Means Clustering". International MultiConference of Engineers and Computer Scientists. Vol. 1, pp.14-16, March 2012.

[25] L.Guerra, V. Robles, C. Bielza, P. Larrañaga. "A comparison of clustering quality indices using outliers and noise". Journal of Intelligent Data Analysis, vol. 16, no. 4, pp. 703-715, July 2012.

[26] M. Hassani and T. Seidl. "Using internal evaluation measures to validate the quality of diverse stream clustering algorithms". Vietnam Journal of Computer Science. Vol. 4, no. 3, pp. 171–183, 2017.

[27] MacQueen, J. B. "Some Methods for classification and Analysis of Multivariate Observations". Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. **1**. University of California Press. pp. 281–297. (1967). MR 0214227. Zbl 0214.46201. Retrieved 2009-04-07 .

[28] Krista Rizman Zalik, "An efficient k-means clustering algorithm". Pattern Recognition Letters. Vol. 29, no. 9, pp. 1385-1391, 2008.

[29] Telgarsky, Matus, Vattani, Andrea. "Hartigan's Method: k-means Clustering without Voronoi". Journal of Machine Learning Research - Proceedings Track. Pp. 820-827, 2010.

[30] Noam. Slonim, Ehud. Aharoni, Koby. Crammer, "Hartigan's K-Means Versus Lloyd's K-Means — Is It Time for a Change?" ACM Proceedings of the Twenty-Third International International Joint

Conference on Artificial Intelligence (IJCAI). pp. 1677-1684, 2013.

[31] Laurence. Morissette and Sylvain. Chartier. "The k-means clustering technique: General considerations and implementation in Mathematica". Tutorials in Quantitative Methods for Psychology. Vol. 9, no. 1, pp. 15-24, 2013.

[32] Hornik and Kurt, "The Comprehensive R Archive Network. 2.13 What is the R Foundation? "2