Study of Internal Indices on Time Series Data

Shailesh Kumar Jha
Department of Computer Science
Georgia State University
Atlanta, GA, USA
sjha3@student.gsu.edu

Ruizhe Ma
Department of Computer Science
Georgia State University
Atlanta, GA, USA
rma1@student.gsu.edu

Rafal Angryk
Department of Computer Science
Georgia State University
Atlanta, GA, USA
rangryk@cs.gsu.edu

ABSTRACT

Clustering is an important concept used in unsupervised machine learning and data mining. There are many studies done to verify the goodness of the clustering. They are broadly divided into Internal and external indices. Where external indices are based on ground truth, internal indices work on barycenters and within cluster and between cluster distances. There are various indexes proposed by various authors for both internal and external indices quality index to determine the optimal clustering.

All these methods have used Euclidean distances, but they do not work with Time series data. When working with time series, Dynamic Time warping (DTW) is the accepted way of measuring distances between two time series.

In this paper, we study to understand if these internal indices work good with time series data using DTW distance instead of Euclidean distance and then also do a comparative study between Kmeans clustering, hierarchical clustering with single linkage, complete linkage and average linkage and the newly proposed DDC(Distance Density Clustering) algorithm[1]. First, we study to understand if we can efficiently replace Euclidean distance with DTW distance. Next, we consider ten UCR datasets and study seven internal indices to understand the performance of each clustering method and document the result.

Introduction

In machine learning and data mining, clustering has always been a challenge and there have been many methods proposed to help with clustering similar data points together. Clustering groups data into various clusters based on some form of similarity between points in the same cluster and dissimilarity between points of different clusters. In other words, the clustering methods are based on the concept of minimization of intra-cluster distance or the distance between each point in the cluster to its centroid which would translate into them being very similar to each other. The second part of it being the maximization of inter cluster distance between two clusters meaning, two different clusters should be far from each other or as dissimilar as possible.

Various methods of clustering have been proposed over the year but all of them use Euclidean distance on discrete datasets. However, this distance does not work on time series data. For this, we study and experiment by using DTW distance on time series data to explore if this is a good method to cluster time series data. Dynamic Time warping distance has been long accepted as the method of calculating distances between two time series. It is a useful distance like similarity measure between multiple time series sequence of varying lengths or speeds. DTW has been used on various time series methods and is recognized as the method to calculate the distance between two time series.

Internal indices use the intra-cluster and inter-cluster distance in various combinations to give us the goodness of the cluster and various authors have coined different indices based on their research. In this paper, we study seven of these indices which have been experimentally proven to be a good measure of the clustering quality using Euclidean distance as a way to measure the distance between members of the dataset. The indices under experimentation are Ball Hall Index, Banfeld Raftery Index, Calinsky-Harabasz Index, Davies Bouldin Index, LogSSI Index, PBM Index and Ray Turi Index. These indices would be discussed in detail in the subsequent sections. The experiment is divided in two parts - First, we experiment and find if they show positive result when we substitute their distance calculation by DTW and second, we then compare the four clustering methods to understand if we have can derive at a clear winner among the clustering methods under discussion namely – Kmeans clustering, hierarchical clustering with single linkage, complete linkage and average linkage and DDC clustering.

Calculating the centroid of the cluster is comparatively difficult for time series data in comparison to regular discrete data. The reason being time series data is often of varying size and speeds and regular averaging methods do not work on time series data. To calculate the average of timer series, Dynamic Time Warping Barycenter Averaging (DBA) was proposed by Petitjean et al. [2] as a method to calculate the average for time series data. It is used as the globally accepted method to calculate the average time series between a cluster of time series as it uses DTW and so the length does not affect the calculation. DBA is the averaging method that has been employed in this paper for calculating barycenter of the clusters.

The below figure [Fig 1] shows the intra-cluster and inter cluster distance between two clusters of data. We will be discussing how using these distances and various internal indices could help us determine which clustering method gives us the best clustering.

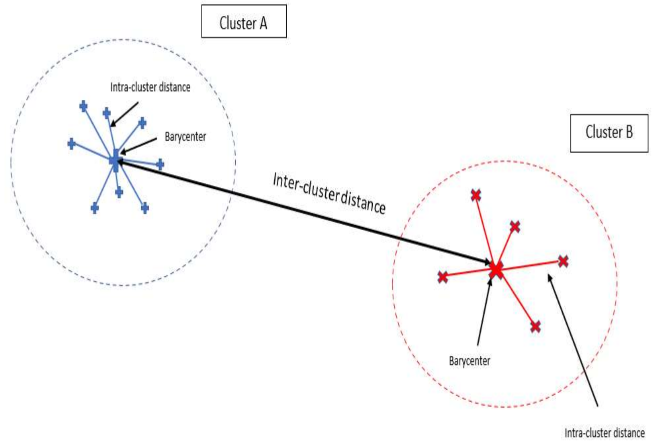

Fig 1 Defining the intra-cluster and inter-cluster distances in a cluster.

KEYWORDS

Time Series Clustering, Internal Indices, Dynamic Time Warping(DTW) distance, DBA(DTW Barycenter Averaging), Ball Hall Index, Banfeld Raftery Index, Calinsky-Harabasz Index, Davies Bouldin Index, LogSSI Index, PBM Index and Ray Turi Index, Kmeans clustering, hierarchical clustering with single linkage, complete linkage and average linkage and DDC(Dynamic Distance Clustering)

RELATED WORK, CLUSTERING METHODS AND INTERNAL INDICES

A lot of work has been done to find new ways to find the distance between time series sequences and to cluster them in a better way but there has been no significant work done in this area to find out if we validate the goodness of the clusters using existing internal indices by replacing certain parameters and recording the results.

1. Clustering Methods

Kmeans was first introduces 50 years ago [3] and is one of the most popular methods used for clustering in unsupervised learning. But Kmeans suffers from not working well with global clusters, different initial partitions resulting in different final clusters and adversely being affected by outliers.

Hierarchical clustering is a methods of building hierarchy of clusters based on the approach taken. The bottom up approach is known as agglomerative approach of clustering into dendrograms and we consider top down approach, it is known as divisive approach. Fig 2 describes both the methods.

In agglomerative approach, we consider each point as its own cluster and compute the similarity based on distance. Then we join the most similar points into a cluster and in this way we computer a hierarchy of clusters until we form a big cluster of the complete set of observations.

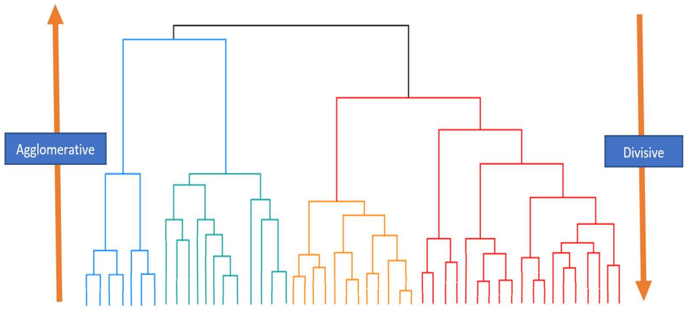

Fig 2: Different approaches for Hierarchical Clustering

In divisive approach, we divide the entire dataset into two dissimilar clusters and continue this until we have one cluster for each observation.

There are three different types of hierarchical clustering that has been taken into consideration in this experiment and are discussed in detail.

In hierarchical clustering with single linkage, the distance between two clusters is defined by the smallest distance between two points in each cluster. For example, the distance between cluster r and s is defined as the distance between the two points highlighted in Fig 3.

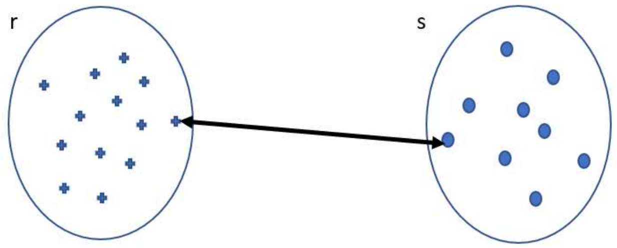

Fig 3: Single Linkage distance

In hierarchical clustering with complete linkage, the distance between two clusters is defined by the largest distance between two points in each cluster. For example, the distance between cluster r and s is defined as the distance between the two points highlighted in Fig 4.

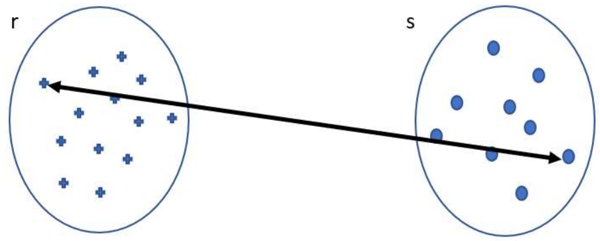

Fig 4: Complete Linkage distance

In hierarchical clustering with complete linkage, the distance between two clusters is defined by the average distance between each point in one cluster to every point in the other cluster. For example, the distance between cluster r and s is defined as the average distance between points of one cluster to another in Fig 5.

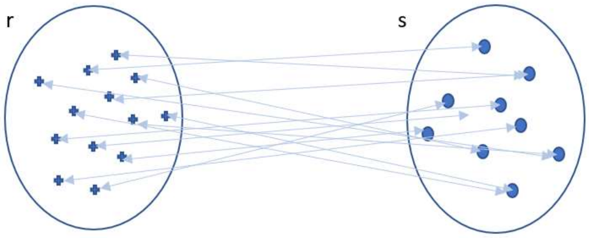

Fig 5

In Distance Density Clustering or DDC proposed by Ma et al [1], they first calculate the distance matrix and select the nearest and furthest point based on majority voting. They plot the distance in an ascending fashion on the initial seed and check the region of bend. This can be considered the region of biggest bend. Based on this, the cluster is again divided into two clusters and rebalancing is done using DBA. Now, they treat each cluster separately and repeat the same process. They select the seed of the cluster with a stronger bend of more clear distance to avoid doubling the number of clusters each time.

2. Internal Indices

Internal Indices measure the goodness of your clustering based on WGSS (Within Group Scatter) and BGSS (Between Group Scatter). Based on the calculation of these two in a cluster, many researchers have proposed various internal indices to figure out the goodness of the cluster and we focused on seven of them in our experiment.

Let us assume there are N observations partitioned in K clusters. Let us assume Ck is cluster k of K clusters. Then, the within-cluster dispersion, noted WGSS$^{\{k\}}$ is the sum of the squared distances between the observations $M_i^{\{k\}}$ and the barycenter $G^{\{k\}}$ of the cluster. Finally, the pooled within-cluster sum of squares WGSS is the sum of the within-cluster dispersions for all the clusters:

$$\text{WGSS} = \sum_{k=0}^{K} WGSS^{k}$$

Equation 1

The between-group dispersion measures the dispersion of the clusters between each other. Precisely it is defined as the dispersion of the barycenters $G^{\{k\}}$ of each cluster with respect to the barycenter G of the whole set of data. Geometrically, this sum is the weighted sum of the squared distances between the $G^{\{k\}}$ and G, the weight being the number n_k of elements in the cluster C_k:

$$\text{BGSS} = \sum_{k=0}^{K} n_k \left\| G^{\{k\}} - G \right\|^2$$

Equation 2

In the subsequent sections, we discuss about the various internal indices used in our experiments.

2.1 Ball Hall Index

Ball hall index is defined as the mean through all the clusters of the mean dispersion of a cluster. The mean dispersion of a cluster is the mean of the squared distances of the point of the cluster with respect to their barycenter. Ball Hall index considers only WGSS. This index was proposed in 1965[4]. If K is the number of clusters that N observations are partitioned into, n_k is the number of observations in cluster k, M is a point in the cluster and, G is the barycenter of the cluster, then Ball Hall Index(C) is given as

$$C = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i \in I_k} \left\| M_i^{\{k\}} - G^{\{k\}} \right\|^2$$

Equation 3

2.2 Banfeld-Raftery Index

This index proposed in 1974[5] is the weighted sum of the logarithms of the traces of the variance-covariance matrix of each cluster. The quantity $\frac{Tr(WG^{\{k\}})}{n_k}$ can be interpreted as the mean of the squared distances between the points in cluster Ck and their barycenter. If K is the number of clusters that N observations are partitioned into, n_k is the number of observations in cluster k, M is a point in the cluster and, G is the barycenter of the cluster The Banfeld-Raftery Index (C) can be written as:

$$C = \sum_{k=1}^{K} n_k \log(\frac{Tr(WG^{\{k\}})}{n_k})$$

Equation 4

2.3 The Calinski-Harabasz index

This index was proposed in 1974[6] and is given as the below ratio:

$$C = \frac{BGSS}{WGSS} \frac{N-K}{K-1}$$

Equation 5

where K is the number of clusters, N is the total number of observations, $BGSS$ is weighted sum of the squared distances between the $G^{\{k\}}$ and G, the weight being the number n_k of elements in the cluster C_k, and WGSS is the sum of the squared distances between the observations $M^{\{k\}}$ i and the barycenter $G^{\{k\}}$ of the cluster.

2.4 Davies Bouldin Index

This index proposed in 1979[7] considers both intra-cluster distance and inter-cluster distance. Let us denote by δ_k the mean distance of the points belonging to cluster C_k to their barycenter $G^{\{k\}}$:

$$\delta_k = \frac{1}{n_k} \sum_{i \in I_k} \left\| M_i^{\{k\}} - G^{\{k\}} \right\|$$

Equation 6

Let us also denote distance between the barycenters $G^{\{k\}}$ and $G^{\{k'\}}$ of clusters C_k and $C_{k'}$

$$\Delta_{kk'} = d\left(G^{\{k\}}, G^{\{k'\}}\right) = \left\| G^{\{k'\}} - G^{\{k\}} \right\|$$

Equation 7

For each cluster k, the maximum M_k of the quotients $\frac{\delta_k + \delta_{k'}}{\Delta kk}$ for all indices k' ≠ k. The Davies- Bouldin index is given as the mean value among all the clusters as:

$$C = \frac{1}{K} \sum_{k=1}^{K} M_k = \frac{1}{K} \sum_{k=1}^{K} \max_{k\prime \neq k} \left(\frac{\delta_k + \delta_{k'}}{\Delta kk'} \right)$$

Equation 8

2.5 Log SS Ratio index

Log SS ratio was proposed in 1975[8] and is the ratio between Between-Group scatter spread BGSS and Within-Group Scatter Spread WGSS. It is defined as:

$$C = \sum_{k=1}^{K} \log\left(\frac{BGSS}{WGSS}\right)$$

Equation 9

2.6 PBM Index

Named after the initials of the three researchers who proposed it (Pakhira, Bandyopadhyay and Maulik)) is calculated calculated using the distances between the points and their barycenters and the distances between the barycenters themselves. It was proposed in 2004[9]

Let us denote by DB the largest distance between two cluster barycenters:

$$D_B = \max_{k<k\prime} d(G^{\{k\}} - G^{\{k\prime\}})$$

Equation 10

On the other hand, let us denote by EW the sum of the distances of the points of each cluster to their barycenter and ET the sum of the distances of all the points to the barycenter G of the entire data set:

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

Equation 11

$$E_W = \sum_{k=1}^{K} d(M_i, G)$$

Equation 12

The PBM index would be defined as

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

Equation 13

where K is the number of clusters, E_T is the sum of the distances of all the points to the barycenter G of the entire data set, E_W is the sum of the distances of the points of each cluster to their barycenter, D_B is the largest distance between two barycenters.

2.7 Ray Turi Index

Ray Turi Index was proposed in 1999[10] and is defined as the quotient.The numerator is the mean of the squared distances of all the points with respect to the barycenter of the cluster they belong to:

$$\frac{1}{N} \sum_{k=1}^{K} WGSS^{\{k\}} = \frac{1}{N} WGSS$$

Equation 14

the denominator is the minimum of the squared distances $\Delta_{kk'}$, between all the cluster barycenters:

$$\min_{k<k\prime} || G^{\{k\}} - G^{\{k'\}} ||^2$$

Equation 15

The Ray Turi Index is given as:

$$\frac{1}{N} \frac{WGSS}{\min_{k<k\prime} \Delta_{kk}^2}$$

Equation 16

3. Datasets

There were ten datasets used in this experiment and they were sourced from UCR repository [11] and a summarized description is mentioned in the following section:

3.1 ArrowHead: The arrowhead data consists of outlines of the images of arrowheads. The shapes of the projectile points are converted into a time series using the angle-based method.

3.2 BeetleFly: It is a database of binary images developed for testing MPEG-7 shape descriptors. It is used for testing contour/image and skeleton-based descriptors

3.3 BirdChicken: It is a database of binary images developed for testing MPEG-7 shape descriptors. It is used for testing contour/image and skeleton-based descriptors

3.4 Coffee: Food spectrographs are used in chemometrics to classify food types, a task that has obvious applications in food safety and quality assurance. The coffee data set is a two-class problem to distinguish between Robusta and Aribica coffee beans.

3.5 GunPoint: This dataset involves one female actor and one male actor making a motion with their hand. The two classes are: Gun-Draw and Point

3.6 Ham: The ham data includes measurements from 19 Spanish and 18 French dry-cured hams.

3.7 Meat: Food spectrographs are used in chemometrics to classify food types, a task that has obvious applications in food safety and quality assurance. The classes are chicken, pork and turkey.

3.8 SonyAIBORobotSurface: The robot has roll/pitch/yaw accelerometers. The task is to detect the surface being walked on which is cement or carpet.

3.9 ToeSegmentation2: Motions in the database containing the keyword walk are classified by their motion descriptions into two categories. The first is the normal walk, with only walk in the motion descriptions. The other is the abnormal walk, with the motion descriptions containing hobble walk, walk wounded leg, walk on toes bent forward, hurt leg walk, drag bad leg walk, or hurt stomach walk

3.10 Wine: This dataset involves wine classification by spectrograph

4. Experimental Results

We considered recording the results of running the internal indices on Hierarchical clustering with single linkage, average linkage and complete linkage, DDC and average values of 10 rounds of randomized Kmeans clustering observations and recorded their results. We also considered computational processing time to consider which internal indices to be explored for our experimentation.

4.1 Ball Hall Index

Here higher the difference between the cluster values, better is the clustering. Based on this DDC algorithm emerged as the best algorithm in our experiments. Fig 6 shows the results of DDC on ArrowHead dataset and Fig 7 describes the result of experimentation on all the dataset for Ball Hall index.

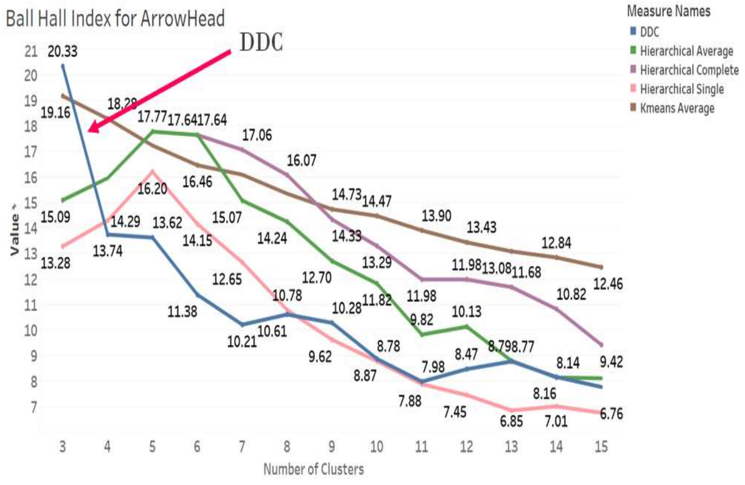

Fig 6: Performance of all algorithms for ArrowHead dataset

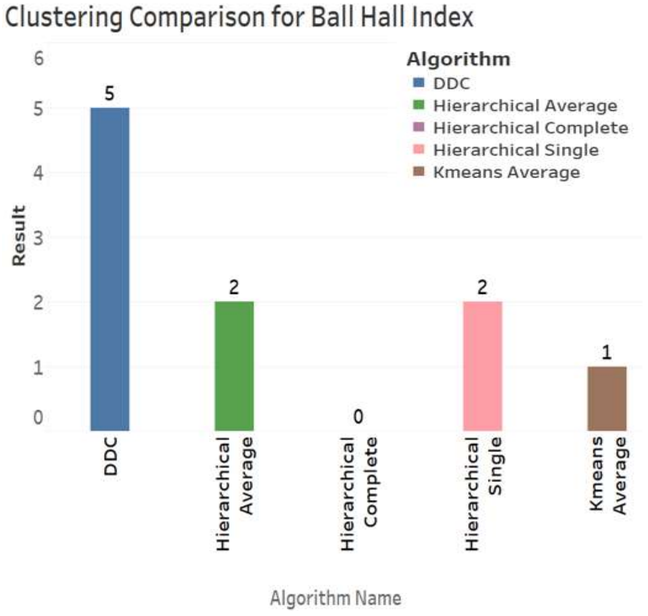

Fig 7: Comparison of all the datasets for Ball Hall Index

4.2 Banfeld-Raftery Index

Since it considers the WGSS values only, the lower value on this internal index means a better clustering. Once again DDC emerges the winner. Fig 8 shows the results of DDC on ArrowHead dataset and Fig 9 describes the result of experimentation on all the dataset for Banfeld Raftery index.

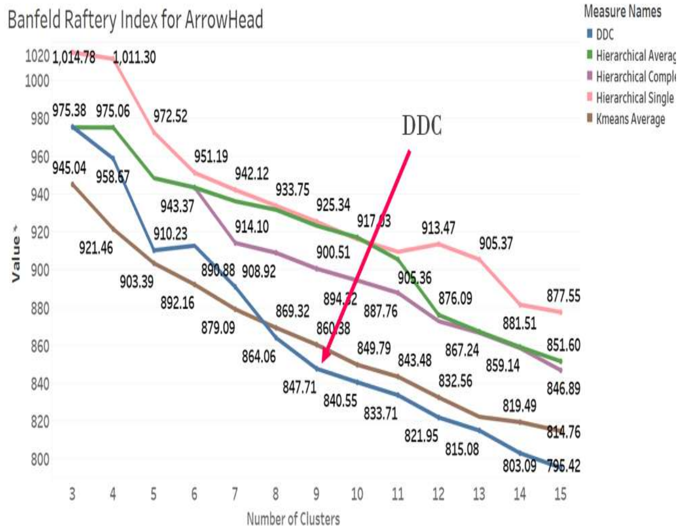

Fig 8: Performance of all algorithms for ArrowHead dataset

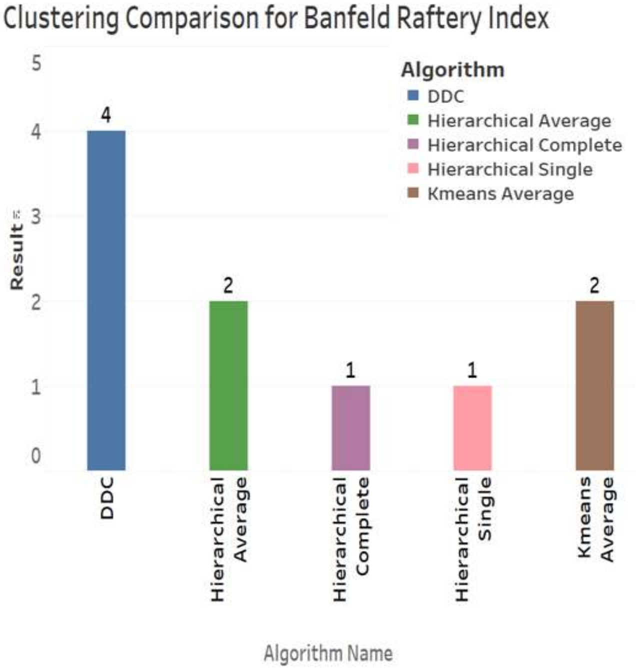

Fig 9: Comparison of all the datasets for Banfeld Raftery Index

4.3 Calinski-Harabasz Index

Almost all algorithms have similar performance here. It is a ratio of both BGSS and WGSS. They perform neck to neck but DDC has a slight edge over here. Here, higher the value, better is the clustering. Fig 10 shows the results of DDC on BirdChicken dataset and Fig 11 describes the result of experimentation on all the dataset for Calinski-Harabasz index.

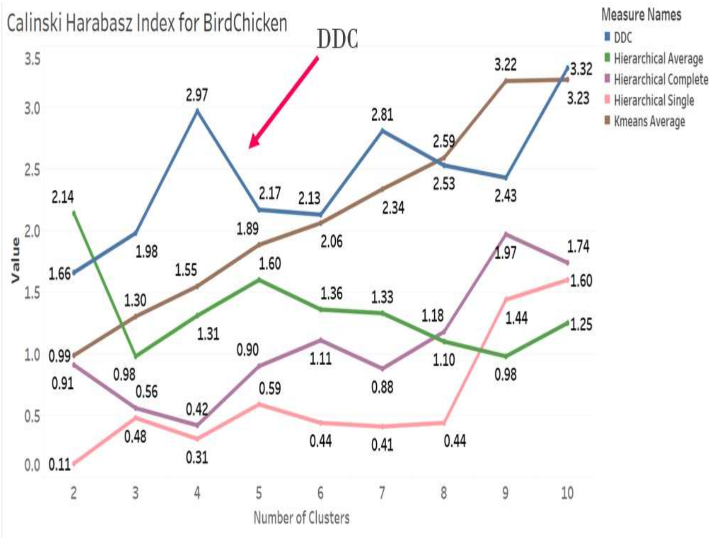

Fig 8: Performance of all algorithms for BirdChicken dataset

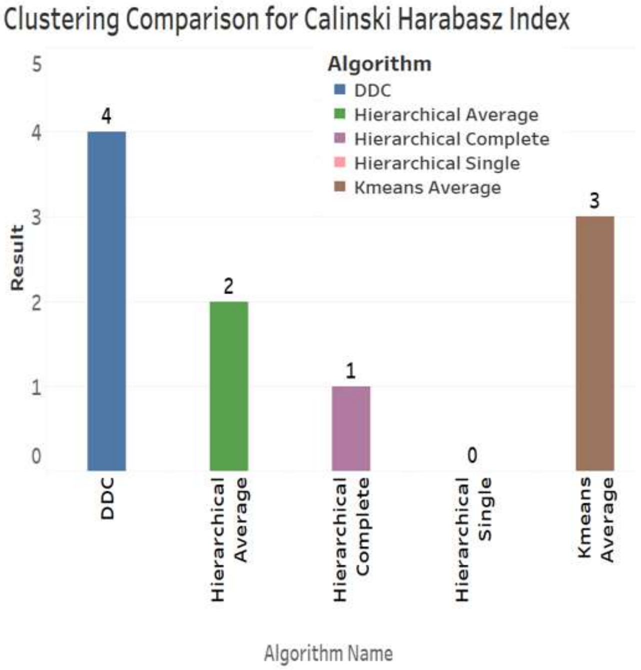

Fig 11: Comparison of all the datasets for Calinski Harabasz Index

4.4 Davies Bouldin Index

Davies Boulding index considers sum of mean distances between points and their barycenters and hence lower values implies a better clustering. Here, Kmeans outperforms every other clustering method by a big margin. Fig 12 shows the results of DDC on BeetleFly dataset and Fig 13 describes the result of experimentation on all the dataset for Davies Bouldin index.

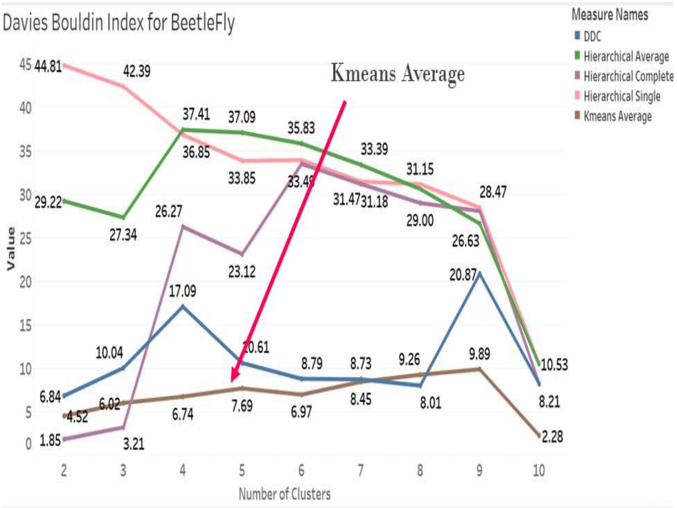

Fig 12: Performance of all algorithms for BeetleFly dataset

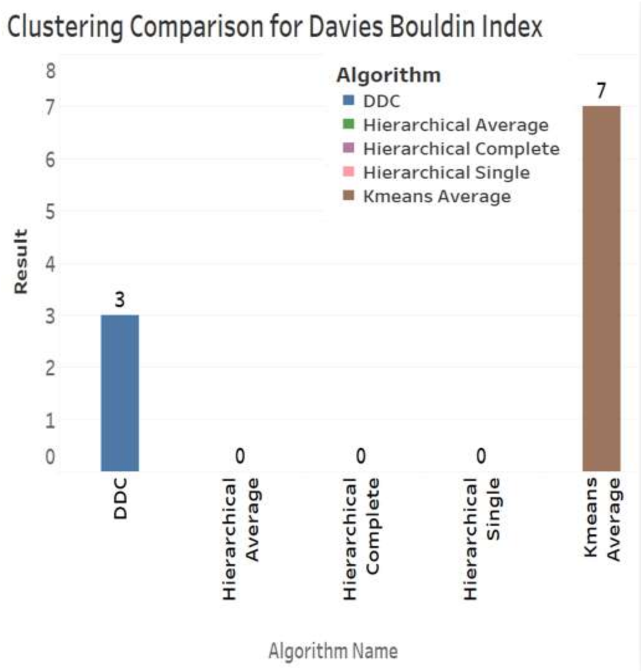

Fig 13: Comparison of all the datasets for Davies Bouldin Index

4.5 Log SS Ratio Index

It is the natural logarithm of ratio of BGSS and WGSS. Here, DDC and Hierarchical Average linkage perform equally well and are way ahead of the others clustering algorithms. Fig 14 shows the results of DDC on BeetleFly dataset and Fig 15 describes the result of experimentation on all the dataset for Log SS Ratio index. Here, the flatter the curve is, the better the clustering is.

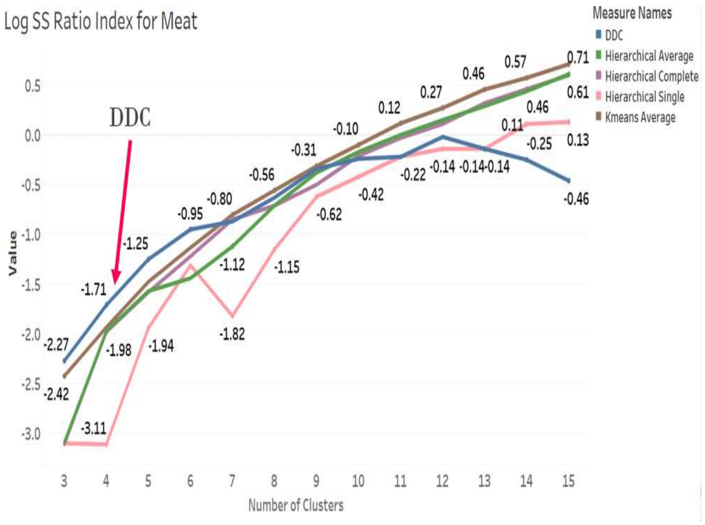

Fig 14: Performance of all algorithms for Meat dataset

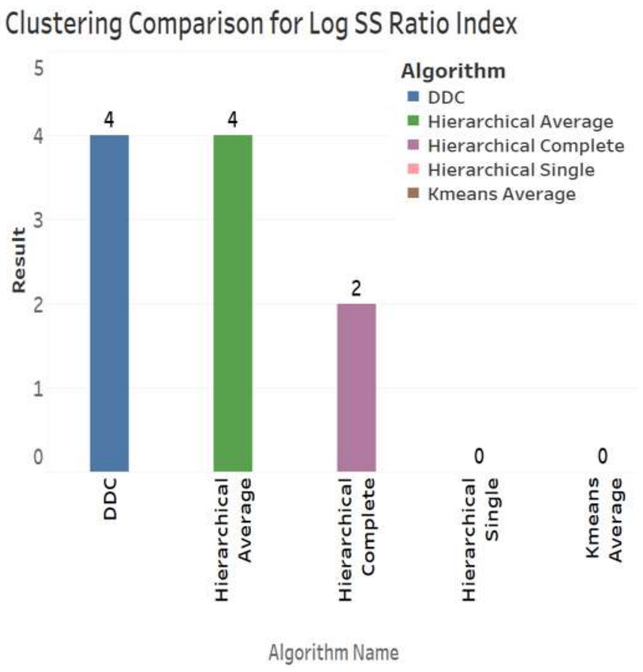

Fig 15: Comparison of all the datasets for Log SS ratio Index

4.6 PBM Index

This index talks about the largest of the distances between two barycenters and hence the higher value implies a better clustering. DDC outperforms other clustering algorithms here whereas the various flavors of hierarchical clustering are equal on performance. The result is shown in Fig 16 which shows the results of DDC on ArrowHead dataset and Fig 17 describes the result of experimentation on all the dataset for PBM index

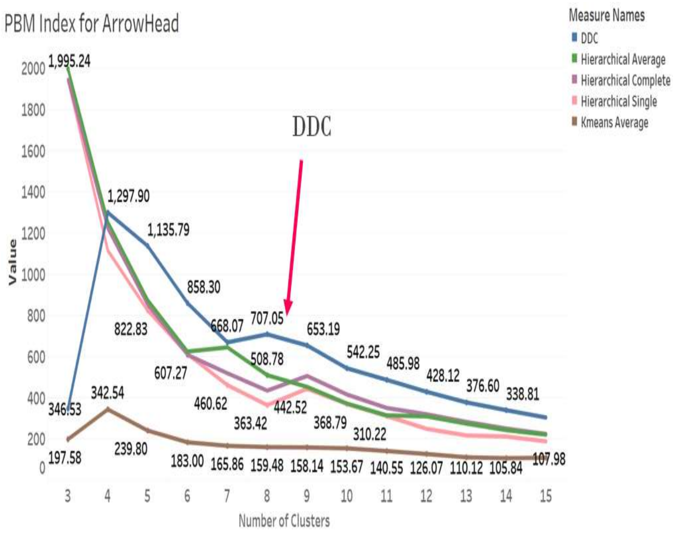

Fig 16: Performance of all algorithms for ArrowHead dataset

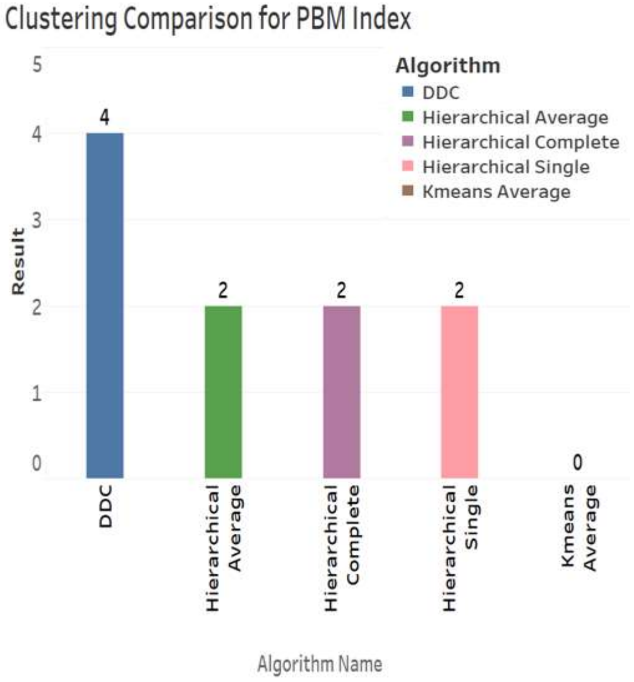

Fig 17: Comparison of all the datasets for PBM Index

4.7 Ray Turi Index

Ray Turi index deals with ratio of WGSS and minimum distance between two cluster barycenters. Here, the lower value we have, the better is the clustering. Once again DDC algorithm performs comparatively better which hierarchical single linkage is a distance second in performance. Fig 18 shows the results of DDC on ArrowHead dataset and Fig 19 describes the result of experimentation on all the dataset for Ray Turi index.

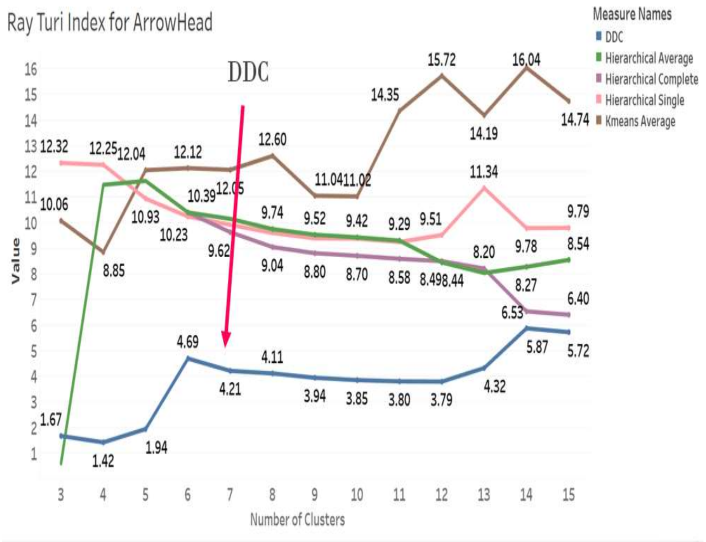

Fig 18: Performance of all algorithms for ArrowHead dataset

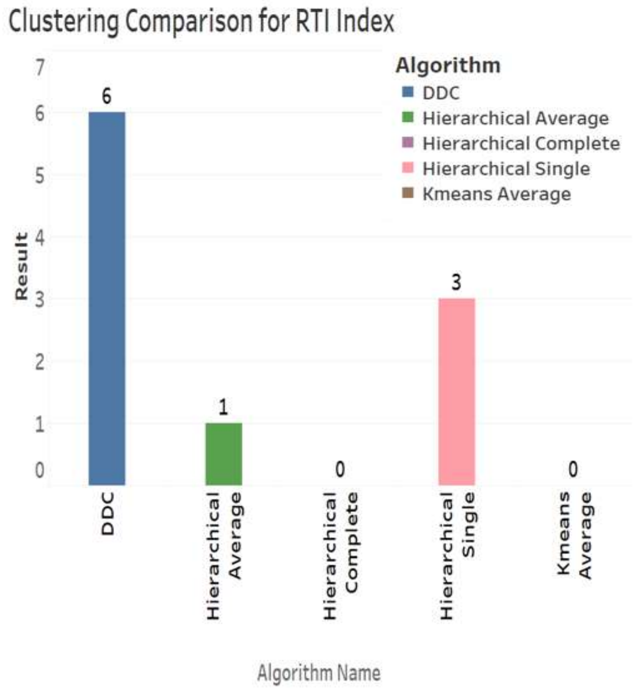

Fig 17: Comparison of all the datasets for Ray Turi Index

Based on the performance and results of the various algorithms on the ten time series datasets that we used from UCR repository has been summarized in table 1.

Internal Indices	Intra cluster distance	Inter cluster distance	Ratio of inter and intra cluster distance	Lower Index value is better	Higher Index value is better	Lower value difference is better	Higher value difference is better	DDC	Hierarchical Single	Hierarchical Average	Hierarchical Complete	Kmeans Average
Ball Hall Index	X						X	X				
Banfeld-Raftery Index	X			X				X				
Calinski-Harabasz Index			X		X			X				
Davies-Bouldin Index			X	X								X
Log SS Ratio Index			X			X		X		X		
PBM Index	X				X			X				
Ray-Turi Index			X	X				X				

Table1: Summarized result

5. Conclusion and future works

Based on the experiments, DDC emerges as the clear winner in comparison to all the other clustering algorithms that we have employed. The conclusions are:

- Internal indices can be used for gauging the goodness of clustering in timer series data.
- DDC outperforms all other methods when we consider intra-cluster distances.
- When there is a ratio of BGSS and WGSS, all of them have almost equal performance in our datasets however, DDC has a slight edge over the others.
- On analysis, for Davies-Bouldin Index K-means random outperforms DDC because it has many single element clusters and due to which the mean distance increases hence the increase in the Index value which denotes a bad performance by DDC.
- A lower value of k for DDC clustering for Calinski-Harabasz Index, Davies- Bouldin Index can augment its performance.

Though the initial results have been encouraging in favor of DDC algorithm, we plan to do a more extensive research on more datasets with more internal indexes to give us a better comparison between the clustering algorithms. We plan to also investigate other clustering algorithms.

ACKNOWLEDGMENTS

This project has been supported in part by funding from CISE, MPS and GEO Directorates under NSF award #1443061, and by funding from the LWS Program, under NASA award #NNX15AF39G.

REFERENCES

[1] Ruizhe Ma and Rafal Angryk, 2017. 2375-9259/17 © 2017 IEEE DOI 10.1109/ICDMW.2017.11.

[2] Petitjean F, Ketterlin A, Ganarski P. A global averaging method for dynamic time warping, with applications to clustering[J]. Pattern Recognition, 2011, 44(3): 678-693.

[3] Nagy, George. "State of the art in pattern recognition." Proceedings of the IEEE 56.5 (1968): 836-863.

[4] G. H. Ball and D. J. Hall. Isodata: A novel method of data analysis and pattern classi_cation. Menlo Park: Stanford Research Institute. (NTIS No. AD 699616), 1965...

[5] J.D. Ban_eld and A.E. Raftery. Model-based gaussian and non-gaussian clustering. Biometrics, 49:803{821, 1993.

[6] T. Calinski and J. Harabasz. A dendrite method for cluster analysis. Communications in Statistics, 3, no. 1:1{27, 1974.

[7] D. L. Davies and D. W. Bouldin. A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1, no. 2:224{227, 1979.

[8] J. A. Hartigan. Clustering algorithms. New York: Wiley, 1975.

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

[10] S. Ray and Rose H. Turi. Determination of number of clusters in k-means clustering and application in colour image segmentation. in Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques, pages 137{143, 1999.

[11] https://www.cs.ucr.edu/~eamonn/time_series_data_2018/