EMPIRICAL COMPARISON OF VARIOUS CLUSTERING ALGORITHMS AND METHODS TO DETERMINE OPTIMAL CLUSTERS FOR REAL DATASETS

V Raviteja Kanakala[1],

Koneru Lakshmaiah Education Foundation, Vijayawada, Andhra Pradesh,

raviteja.kanakala@gmail.com,

K.Jagan Mohan[2],

Department of Information Technology, Annamalai University, Tamilnadu,

aucsejagan@gmail.com

V.Krishna Reddy[3],

Department of CSE, Gandhi Institute of Technology and Management, Andhra Pradesh,

kvuyyuru@gitam.edu

Y Jnapika[4],

Computer Science, Smt.Nps Govt. Degree College, Chittoor, Andhra Pradesh

jnapikagdl@gmail.com

Abstract:

In data mining, Clustering is one of the most powerful unsupervised learning technique to find the similar characteristics among the dataset and to separate dissimilar objects in different groups. As there are various number of clustering algorithms, and every clustering algorithm exhibits different results according to the conditions, the choice of selecting a suitable algorithm and suitable measure for evaluation depends on the clustering objectives and task. Hence the quality of clustering process is determined by the purity of the cluster, cluster analysis plays a important role. The main objective of this paper is to determine optimal number of clusters while working on real data sets with different clustering algorithms and evaluation methods. The work of this paper is concerned with the evaluation of various clustering algorithms like K-MEANS , GMM, Hierarchical clustering with different methods like Elbow method, Silhouette method, Gap Statistic Method, Calinski-Harabaz index to find through which evaluation method we can get optimal number of clusters when above mentioned clustering algorithms are used.

Keywords: Clustering, K-MEANS , GMM, Elbow method, Silhouette method, Gap Statistic Method, Calinski-Harabaz index.

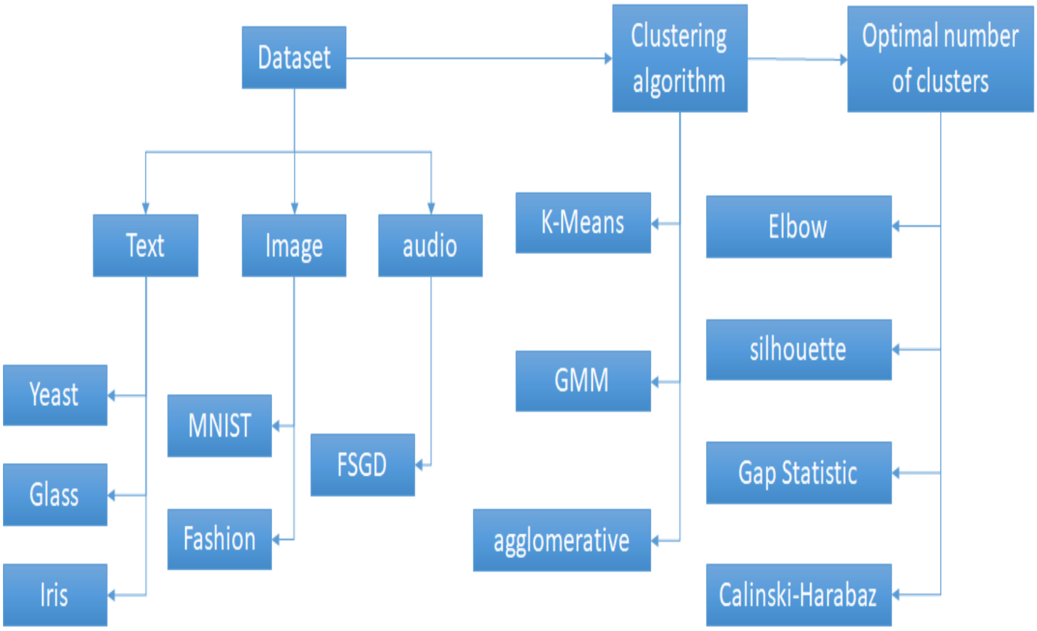
INTRODUCTION

Machine Learning uses some programmed algorithms which can receive and analyses the input data to do predictive modelling , Selecting the right algorithm is the key part of any machine learning project .Learning algorithms are majorly classified into different categories supervised, unsupervised, reinforcement and semi-supervised learning. The amount of available data significantly affects the performance of these learning algorithms. While supervised learning algorithms are being used extensively in many fields they possess the inherent limitation of availability of data. In supervised learning the data used in training the algorithm need to be prelabelled. Such labelled processed data is scarce when compared to the amount of raw unlabeled data generated today in today's digital world. Unsupervised learning algorithm is used to draw the inferences from the dataset that is neither labelled nor Classified. In Unsupervised algorithms, Clustering is one such exploratory technique used in organizing the data into sensible groups which is essential for learning and understanding. K-means, GMM and Hierarchical clustering are one of the most prominently used clustering algorithms. Since clustering uses the untagged data it is crucial to determine the number of groups and clusters exists in the dataset. In datasets such as handwritten digits such as MNIST it is obvious for that there could be only ten clusters or groups but for many multivariant datasets it not so apparent. Over years there are many techniques proposed to determine the number of such clusters such as Elbow, Silhouette, Gap statistic method, Calinski-Harabaz Index and information criteria are being widely used.

RELATED WORK

More than clustering algorithms there are methods that determine number of clusters. There exists no rule of thumb in picking a right combination of these for unsupervised learning. Many works on clustering are done using artificially generated datasets, but very few papers discuss about how these algorithms in combination with the methods to determine the optimal number of clusters work on real-world datasets. Baarsch.et. al[1] discusses how the methods perform using K-mean clustering on artificially generated datasets. Another similar review paper by kodinariya.et.al[2] in 2013 discusses some of the methods for K-mean clustering. Another notable work on determining number of clusters using a clustering algorithm which is based on scale-space theory is proposed by nakamura.et.al[3] in 1998.

METHODOLOGY



Dataset Selection:

Initially a diverse group of real-world datasets are chosen avoiding any computer generated datapoints. The datasets chosen among the classes of text, image and audio. Here, the yeast, Glass and Iris are some popular text datasets, MNIST and Fashion are widely used image datasets, and Free Sound audio datasets recently introduced large multiclass audio dataset. During this step we also enforced necessary preprocessing and feature extraction. Dimensionality reduction using AutoEncoder is utilized in order to achieve computationally inexpensive implementation while retaining as much information as possible. The table [ref table] best summarizes about the dataset with number of classes and their dimensions.

Datasets:

The following are the datasets used in the paper.

Yeast dataset:

Yeast dataset is a publicly available dataset [Ref the dataset]. This is a multiclass classification dataset with 1484 data points. The objective of this dataset is to predict the cellular localization of the yeast protein. The number of classification classes are ten and each data point has eight attributes.

Glass dataset:

Glass dataset is a publicly available dataset [? ref the dataset]. This is a multiclass classification dataset with 214 data points. This objective of this dataset is to classify the type of glass based on the given attributes. The number of classification classes are seven and number of attributes for each data point is nine.

Iris dataset:

This is one of the most popular publicly available datasets in the machine learning community. This is a multivariant dataset introduced by the biologist Ronald Fisher in his paper. The dataset 150 records of three types of iris flowers with four attributes which include petal length, petal width, sepal length and sepal width.

Free sound audio dataset:

This dataset is introduced in the "General-purpose audio tagging of Freesound content with Audio Set labels" challenge hosted on Kaggle. The goal of this challenge is to build an audio tagging system. This dataset consists of 41 diverse sounds. There are 9473 audio files which belong on to either of 41 classes.

MNIST:

The MNIST is the image datasets of 10 digits. This is popular dataset consists of grey scale images of size 28x28 pixels, used for benchmarking various algorithms. This dataset consist of 60,000 train images 10,000 test images.

Fashion:

The Fashion is another popular image dataset with 10 classes. This is introduced as an alternative to MNIT dataset for benchmarking algorithms. The dataset include 28x28 pixel gray scale images of fashion items such as shirts, trousers, shoes, bags etc. This dataset consists of 60,000 train images and 10,000 test images.

Dataset	No. of features	No. of classes	No. of datapoints

Yeast	8	10	1484
Glass	9	7	214
Iris	4	3	150
MNIST	49	10	60000
Fashion	49	10	60000
Free sound dataset	131	41	9473

Feature extraction and Dimensionality reduction:

For extracting features from the Audio dataset, we have used Mel-frequency cepstral coefficients (MFCC). In total of 131 features were extracted from audio files. In case of image dataset, namely MNIST and Fashion, the image dataset has 28x28 pixel values. If each pixel is represented as a feature, then there are 784 features. Such a large number of features are computationally expensive for clustering. Thus, for reducing the dimension of this data we used another unsupervised learning method, Auto Encoding. In this process using a single encoding layer we have reduced the feature size from 784 to 49.

A) Applying clustering algorithm:

The next step include using clustering algorithms K-means, GMM , and Agglomerative clustering on the chosen datasets. After obtaining the clusters using the above algorithms we move to step-3 which aims at determining the optimal number of clusters. The Various Clustering Algorithms used in this paper are:

i)K-Means:

K-means is one of the most popular and widely used partition clustering algorithms known for its simplicity and speed. This algorithm has been in use for over 60 years Jain, A. K et.al[7]. This algorithm requires three parameters from user prior to its execution: number of clusters(K), cluster initial positions, and distance metrics. There is no perfect method to determine number of clusters required (the methods are discussed in latter sections). While different initializations produce different clustering since k-mean converges to local optimum, it is noted in a study that large probability K-mean could converge to global optimum if the clusters are well separated Meilă e.t al [8].Euclidean metric is generally used as the distance metric hence producing spherical clustering. Apart from this, Mahalanobis distance metric Mao,J e.t al [9] , Itakura–Saito distance Linde, Y e.t al [10] can also be used.

ii)GMM:

Gaussian mixture Model is another widely used in pattern recognition and statistical pattern recognition \ci McLachlan, G e.t al [20]. This is a model-based clustering algorithm where it is assumed that the data are generated by a mixture of underlying probability distributions. In this method we maximize the posterior probability that a data point belongs to its clusters. While K-means produces hard assignment, GMM produces soft assignments since we are calculating the probabilities of each data point belong all given clusters which makes GMM more flexible than K-means. In this algorithm, the clusters are assumed to have gaussian distributions and we determine the parameters to determine the agaussian distribution which best fit the given data. The tool used for determining the parameters (weight, means, covariances) for each gaussian cluster is Expectation-Maximization (EM) algorithm Dempster, A. P e.t al [12], Redner, R. A [13], In the first step, the expectation step, we compute the probability thateach data point is generated by k gaussians and in the second step, the Maximization step we update our weights, means, and covariances. This approach is similar to K-means, in fact K-means is often called as a special case of Gaussian Mixture Model. The GMM model takes number of clusters (k) as user input before performing clustering analysis.

iii)Hierarchical Clustering:

Hierarchical clustering is often portrayed as clustering approach but is limited because of its quadratic time complexity. In this approach data is grouped over a variety of scales by creating a dendrogram or tree. This is multilevel hierarchical approach where clusters at a level are joined as cluster in the next level. There are two main strategies in this approach: Agglomerative and Divisive. In Agglomerative, also known as bottom-up approach, each observation is started as its own cluster and they merge while moving up in the hierarchy in a pair-wise manner. In Divisive, also known as top-down approach, all observations are treated as one single cluster, and are split recursively while moving down the hierarchy. The Hierarchical methods are practically feasible if number of possible splits are restricted. In Agglomerative approach, the number of stages is bounded by the number of groups in the initial partition. The splitting and merging operations in Hierarchical clustering is based on some heuristic criteria such as single link, complete link or sum of squares Kaufman, L e.t al [17] Maximum-likelihood criteria is used case of model-based methods when merging the groups Banfield, J. D e.t al [15].

B) Determining optimal number of clusters:

In this step we determine the number of optimal clusters using the different criteria – Elbow, Silhouette, Gap Statistic and Calinski-Harabaz index. Using the information obtained from this step observation is presented.

Methods

Algorithms such as K-means and GMM require user to specify the number of clusters (k) prior to clustering. Hence, the determination of optimal number of clusters in a dataset is a fundamental issue in clustering. Over years many indices and methods have been published for determining the number of clusters. Among these Elbow, Silhouette and Gap Statistic methods are some popular methods. These methods can be classified as direct and statistical testing methods; In direct method, a criteria is optimized such as sum of squares and in statistical testing methods evidences are compared with null hypothesis \cite{charrad2012nbclust}. Elbow, and Silhouette are examples of Direct methods and Gaps Statistic is a statistical testing method. Apart from these, a simple and popular approach is using hierarchical clustering which produces dendrogram or tree to determine number of clusters.

i)Elbow method:

This method is popular in determining the number of clusters, especially when using K-Means algorithm. The intra-cluster variation or within-cluster sum of square(WSS) is minimized in K-means algorithm, measures the compactness of a cluster. This value should be as small as possible for better cluster performance. In Elbow method the WSS value is plotted for different number of clusters forming a curve. The optimal number of clusters is determined using Elbow method when adding another cluster number does not produce significant improvement in WSS value. In the graph, this number is usually located at the bend (Elbow) shaped position of the curve.

ii)Silhouette method:

Since Elbow method is ambiguous sometimes when its hard to find the clear Elbow position in the graph, Silhouette method is used in its stead. This method determines the quality of clustering based on how each object lies within its cluster squares Kaufman, L e.t al [17]. The higher the value of Average Silhouette value the better is the clustering. The optimal number of cluster, is determined when it maximizes the average silhouette value over a range of values for k squares Kaufman, L e.t al [17].

iii)Gap Statistic Method:

This widely used Statistical testing method since it could be applied to any clustering algorithm. This approach compares total within-cluster variation for different values of k with their expectation maximum under null reference distribution of data. The optimal cluster number is the one with maximum gap statistic value cite{TrevorV63N2}.

iv)Calinski-Harabaz index:

Among various independent methods proposed in finding number of clusters Calinski-Harabaz index Caliński, T e.t al [19] has been one of the most successful ones. This value is determined by the relationship between ‘between cluster scatter matrix’ and a ‘within cluster scatter matrix’, shown in the equation.Here, trace BCSM is the weighted sum of square distances between centroid of entire data points and individual cluster centers. Trace WCSM is simply the distance between each datapoint within a cluster and the center of the cluster. Unlike other methods Calinski-Harabaz index uses the distance between cluster center and centroid of entire dataset as a measure of separation in the given dataset. From the equation, due to normalizing ration (N-K)/(K-1) it is evident that the score decreases as k value increases.

I. RESULTS AND DISCUSSION

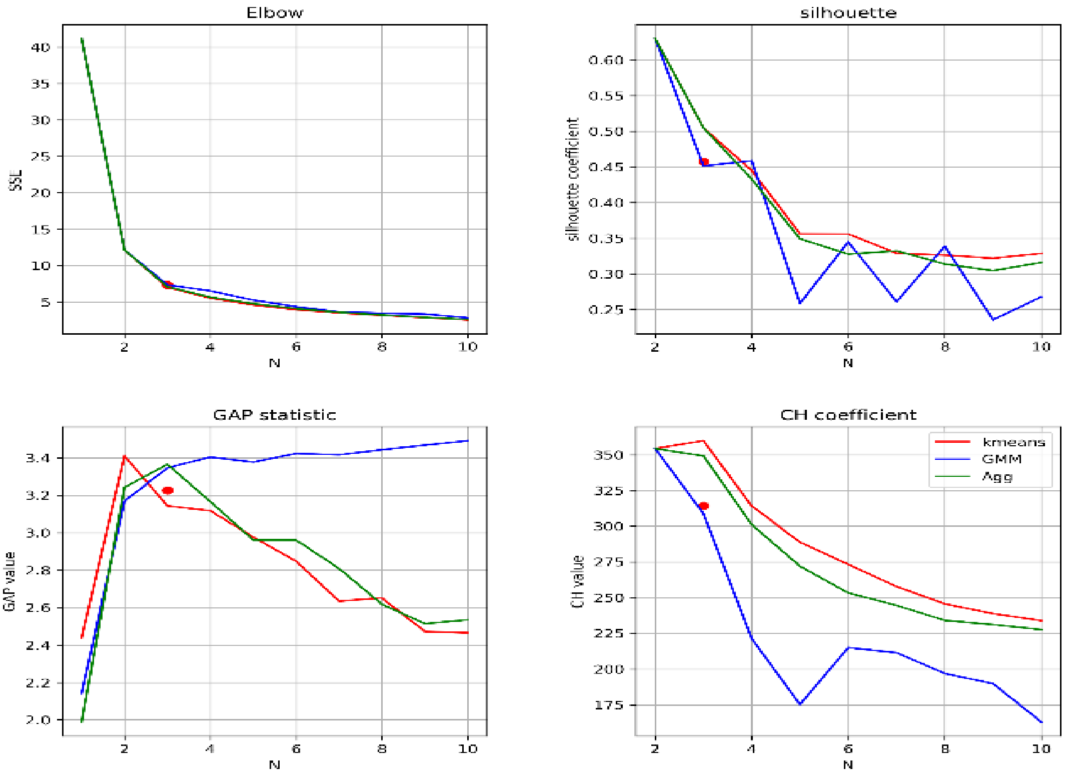


Fig 4: IRIS

All the methods should posit the number of clusters equal to number of classes in the datasets. The figure [ref iris] show the values of the coefficients and index values for determining the optimal number of clustering for Iris dataset. From the fig[iris] it is evident that, since Iris data is well clustered dataset the Elbow methods accurately shows the number of clusters equal to number of classes. For the same dataset, the silhouette coefficient fails to determine the optimal number of clusters as it determines the number of clusters to be two for all the used algorithms. Using GAP statistic method proven successful only in the case of Agglomerative clustering as it fails to determine the number of clusters using the clusters obtained by K-means and GMM. Using Calinski-Harabaz index is only proved useful in case of K-means as Calinski-Harabaz index could not determine the number of clusters obtained using GMM and Agglomerative clustering.

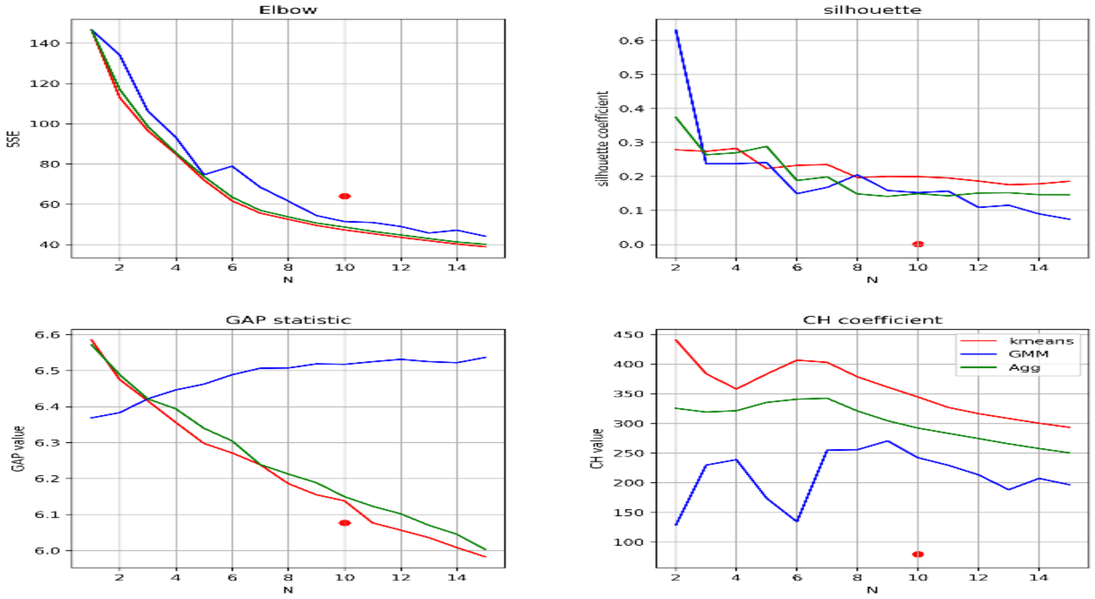


Fig 3: Yeast

The figure [ref yeast] shows the trends of values of the methods used in combination with the clustering algorithm for yeast dataset. From the figure it is hard for determining the location of elbow hence it is not suitable for to determine the number of clusters. Using Silhouette coefficient, we are unable to determine the optimal number of clusters. This is also the same case in GAP statistic, since it either is continuously increasing or decreasing function. Among other only Calinski-Harabaz index produces convincing number of clustering showing number of clusters as seven with agglomerative clustering and nine as number of clustering when GMM clustering algorithm is applied.

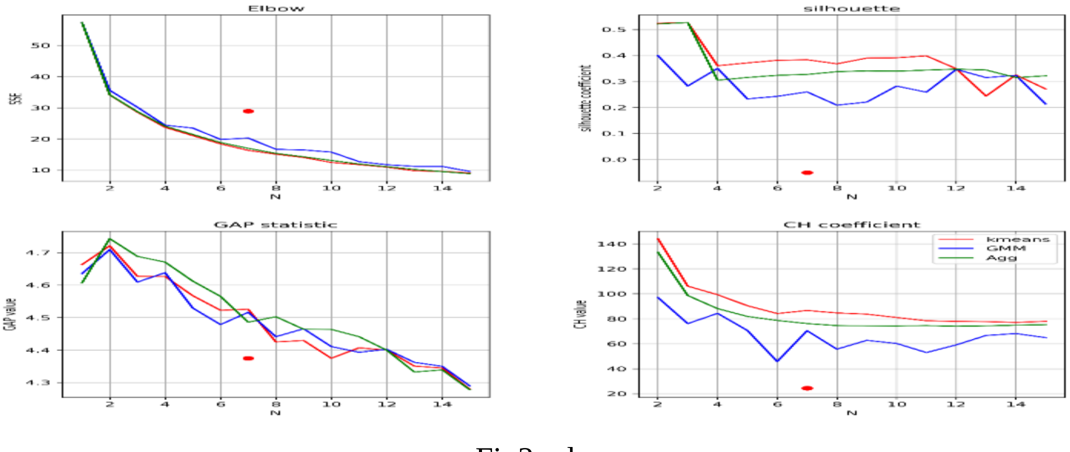


Fig2: glass

The figure [ref glass] shows the trends of values of the methods across various number of clusters using different algorithms for glass dataset. From the figure it can determined that the Elbow method does not provide clear location of the elbow to determine the optimal number of clusters. The Silhouette coefficient is does not produce highest coefficient value for the cluster number equal to number of classes for the dataset.GAP statistic does not show proper cluster number as the GAP value either increases or decreases with number of clusters. The Calinski-Harabaz index could not determine the optimal number of clusters using any of the clustering algorithm

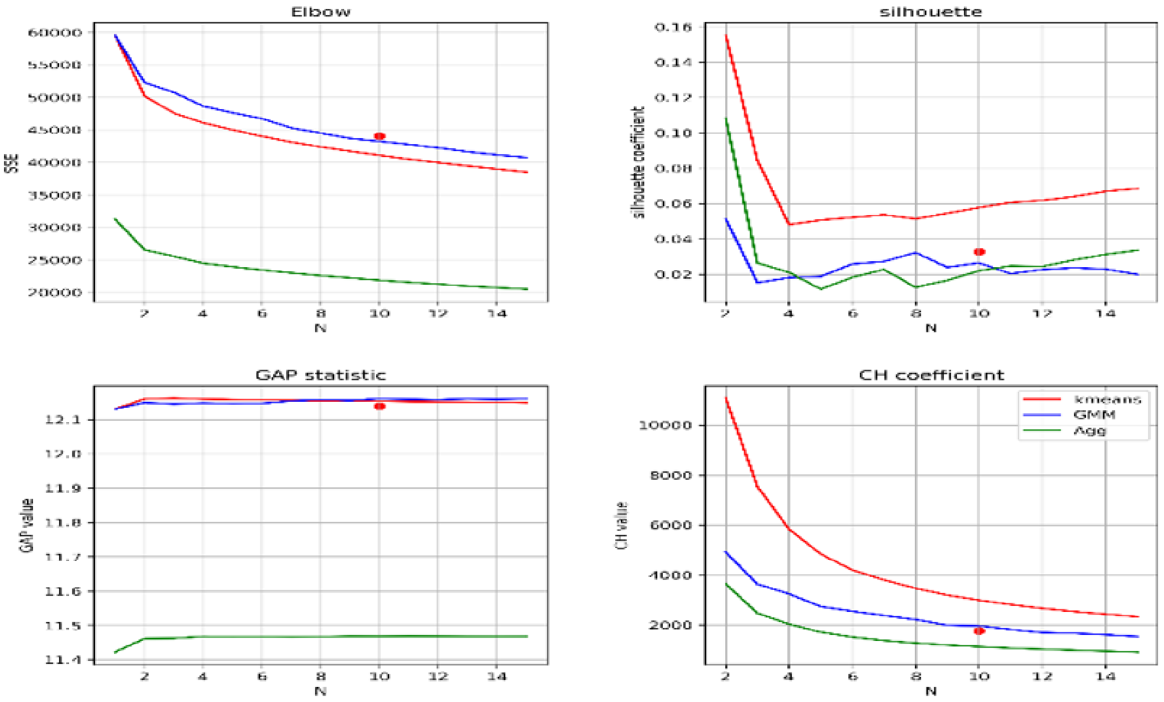


Fig 5:Fashion

The figure [ref fashion] shows the trends of values of the methods across various number of clusters using different algorithms for Fashion dataset. From the figure it is evident that it is hard to determine the number of clusters using Elbow as it lacks a clear elbow shaped curve. The Silhouette coefficient able to posit good optimal number of clusters of eight for K-means, and nine for agglomerative clustering and fail to produce convincing optimal cluster value using GMM clustering algorithm. Calinski-Harabaz could not produce good optimal clustering value using any of the mentioned clustering algorithm.

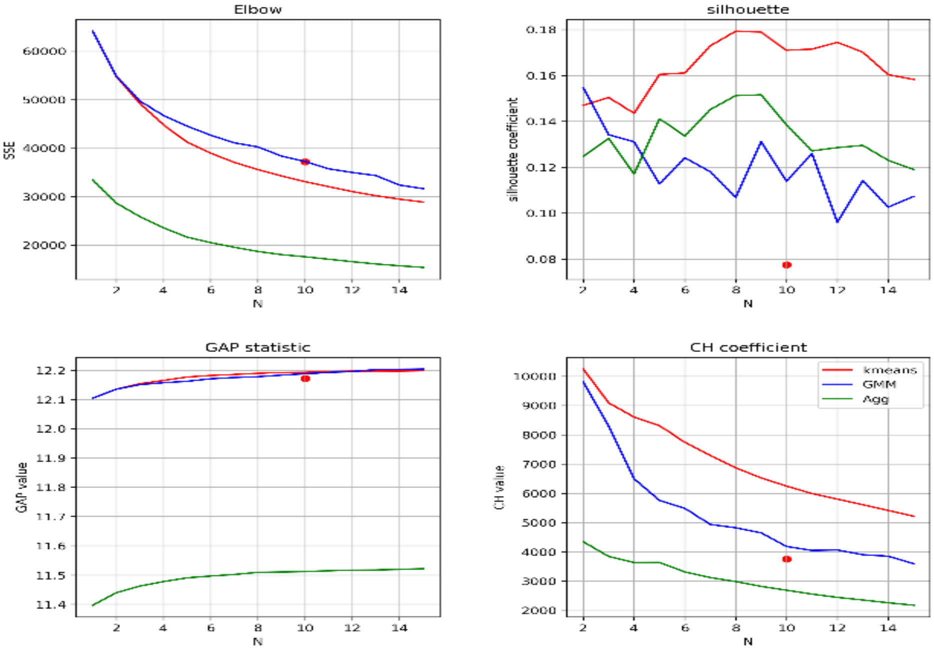


Fig6 : MNIST

The figure [ref mnist] shows the trends of values of the methods across various number of clusters using different algorithms for MNIST dataset. Elbow does not produce the right cluster number for the dataset using the clustering algorithms. On the other hand, silhouette produces relatively better results only with GMM as the second highest value of silhouette coefficient is obtained at the cluster number eight. Using GAP statistic, we are able obtain the number of clusters as eleven for agglomerative clustering but fail to determine the require optimal cluster number using K-means and GMM. The Calinski-Harabaz could not produce indicate the required optimal number of clusters which is the number of classes in the dataset i.e., ten clusters.

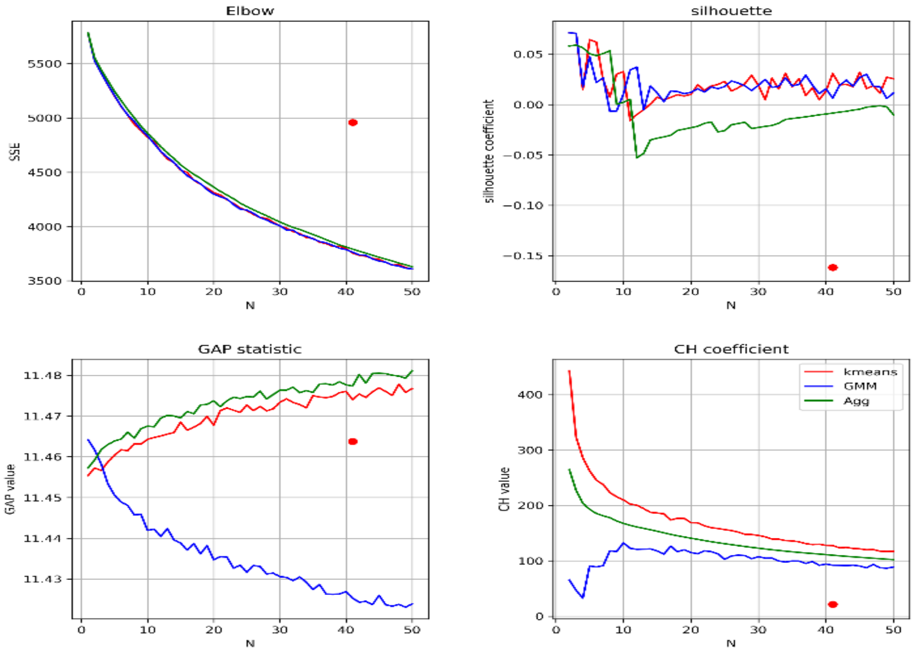


Fig7 : Audio

The figure [ref audio] show the trends of values of the methods across various number of clusters using different clustering algorithms for Free Sound Audio dataset. Elbow does not provide any clear elbow structure to determine the optimal number of clusters. Using Silhouette coefficient produces optimal number of clusters as 45 between the interval 10 to 50 using K-means, and number of clusters as 46 between the interval 11 to 50 using GMM clustering. Using Agglomerative clustering the silhouette produces the optimal number of clusters as 48. GAP statistics could not produce proper results with any of the clustering algorithm. This is also the same case with Calinski-Harabaz index. According to the index the optimal number of clusters is 10 using GMM clustering algorithm. It could be further noted that as the number datapoints increases in a cluster the Elbow method is less likely to be useful in determining the optimal number of clustering as the curve becomes more and more smooth and elbow shape turns to be more like a curve.

Table 2

	K-means	**GMM**	**Agglomerative clustering**
Yeast	Eb(7), CH(7)	-	Eb(7), CH(9)
Glass	Eb(4)	-	-

Iris	Eb,CH	Eb	Eb,GAP
MNIST	-	Silh(8)	GAP(11)
Fashion	Silh(8)	-	Silh(9),GAP(11)
Free sound dataset	Silh(45)-[10:50]	Silh(46)-[11:50]	Silh(48) – [10:50]

Conclusion:

As most of the times the number of clusters for real-world datasets cannot be determined directly But from the above results for **tabular datasets**, we found Elbow and Calinski-Harabaz index are producing better number of cluster predictions when using K-means and Agglomerative clustering. When considering **image datasets** we find GAP statistics is able to produce better approximation of number of clusters when used in combination with Agglomerative clustering. Due to high dimensionality of **Audio dataset** we found that all methods are predicting number of clusters less than 10. But using **Silhouette method** we are able to find better approximation of number of clusters when considering values Silhouette coefficient for clusters greater than 10.

REFERNCES

[1] Baarsch, J., & Celebi, M. E. (2012, March). Investigation of internal validity measures for K-means clustering. In Proceedings of the international multiconference of engineers and computer scientists (Vol. 1, pp. 14-16). sn.

[2] Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. International Journal, 1(6), 90-95.

[3] Nakamura, E., & Kehtarnavaz, N. (1998). Determining number of clusters and prototype locations via multi-scale clustering. Pattern Recognition Letters, 19(14), 1265-1283.

[4] Bholowalia, P., & Kumar, A. (2014). EBK-means: A clustering technique based on elbow method and k-means in WSN. International Journal of Computer Applications, 105(9).

[5] Park, G. Y., Kim, H., Jeong, H. W., & Youn, H. Y. (2013, March). A novel cluster head selection method based on K-means algorithm for energy efficient wireless

sensor network. In 2013 27th International conference on advanced information networking and applications workshops (pp. 910-915). IEEE.

[6] Pérez, J., Pazos, R., Cruz, L., Reyes, G., Basave, R., & Fraire, H. (2007, August). Improving the efficiency and efficacy of the k-means clustering algorithm through a new convergence condition. In International Conference on Computational Science and Its Applications (pp. 674-682). Springer, Berlin, Heidelberg.

[7] Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern recognition letters, 31(8), 651-666.

[8] Meilă, M. (2006, June). The uniqueness of a good optimum for k-means. In Proceedings of the 23rd international conference on Machine learning (pp. 625-632). ACM.

[9] Mao, J., & Jain, A. K. (1996). A self-organizing network for hyperellipsoidal clustering (HEC). Ieee transactions on neural networks, 7(1), 16-29.

[10] Linde, Y., Buzo, A., & Gray, R. (1980). An algorithm for vector quantizer design. IEEE Transactions on communications, 28(1), 84-95.

[11] Rasmussen, C. E. (2000). The infinite Gaussian mixture model. In Advances in neural information processing systems(pp. 554-560).

[12] Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1), 1-22.

[13] Redner, R. A., & Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. SIAM review, 26(2), 195-239.

[14] Rousseeuw, P. J., & Kaufman, L. (1990). Finding groups in data. Hoboken: Wiley Online Library.

[15] Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. Biometrics, 803-821.

[16] Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2012). NbClust Package: finding the relevant number of clusters in a dataset. UseR! 2012.

[17] Kaufman, L., & Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis (Vol. 344). John Wiley & Sons.