IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-10, Oct. 2018 E-ISSN: 2347-2693

Clustering Algorithms Validated Using Relative Index Validation

T. Senthil Selvi[1*], R. Parimala[2]

[1,2]Dept. of Computer Science, Periyar E.V.R. College, Tiruchirappalli, Tamilnadu, India

[*]*Corresponding Author: senthilselvikumar@yahoo.co.in, Tel.: +919443780284*

Available online at: www.ijcseonline.org

Accepted: 6/Oct/2018, Published: 31/Oct/2018

Abstract—Clustering pertains to the task of finding out groups of objects such that the objects of one group are dissimilar from other groups and is similar within the same group. This work uses feature selection technique like the Document frequency Feature selection (DFFS) and feature extraction techniques like Principal Component Analysis (PCA) and Kernel Principal Component Analysis (KPCA) were it constructs a small set of features from the original features. The newly constructed features run the K-Means algorithm without any loss of information. On several runs evaluate the accuracy for the clustering algorithms and record the results. For the obtained results, determine the cluster validation. Internal validation measures are employed to evaluate for cluster validation, based on these measures the relative validation measure is employed to determine the best clustering algorithm. Experiments are conducted for various benchmark datasets comprising of unlabelled documents and the final results prove to show that DFFS, KPCA followed by K-Means algorithm gives the best clustering results of accuracy.

Keywords—Clustering,RelativeValidityMeasures,PCA,KPCA.

I. INTRODUCTION

Clustering is the process of organizing objects into groups whose members are similar in some way. It is difficult to analyze whether the grouping is correct or not. A great challenge in clustering algorithm is that it should produce groups with distinct non-overlapping boundaries, although a perfect separation cannot typically be achieved in practice. There exists different problems with clustering, among them includes how to deal with the size of a corpus and the number of features. Researchers performed Distance based K-Means clustering algorithm for clustering unstructured corpus. The K-Means procedure can be viewed as a greedy algorithm for partitioning the n corpus into k clusters so as to minimize the sum of the squared distances to the cluster center. The measures of within-cluster (Intra) with lower values and between-cluster (Inter) separability with higher values are better. The common approach towards corpus clustering is the bag of words model (BOW), where words/terms are considered as features. TF-IDF is used to represent the corpus into Vector Space Model. The effectiveness of the method depends on the distance measure used and the time complexity of distance calculation. The K-Means clustering algorithm uses the Euclidean distance to measure the similarities between corpuses. The calculation of distance depends on the number of features. Unsupervised Feature selection is used to select the set of features whose size is less than the original feature size. Feature extraction maps corpus of high dimensionality space into lower dimensionality space. PCA and KPCA are used to reduce the dimensionality of the feature space and then run on K-Means clustering algorithm. This approach allows us to overcome most of the limitations imposed by K-Means. Cluster validation considers the quality of the clustering algorithm results which attempts to find a partition that best suits the intrinsic nature of the data. In reality, Clustering techniques are very sensitive to their input corpus and results obtained vary differently for each seed and run. Here, the focus is made on the Cluster validation Indexes (CVI's) based on the internal criteria and how well the compactness and separability holds good for a proposed work. The need for Relative Validation Index (RVI) decides which features best suites the clustering.

The paper is organized as follows: Section I introduces clustering and Cluster validation Index. Section II outlines the Literature Review. Section III presents the cluster validation category Section IV presents the various validity indices. Section V the Proposed Methodology. Section VII brings out the Corpus used for study. Section VIII outlines the used Environment and Libraries. Section IX discusses on the Experimental Results and finally Section X concludes.

II. LITERATURE REVIEW

K.P. Agrawal, S. Garg and P. Patel proposed to validate clustering structure especially for dense, sparse and arbitrary shaped clusters. In K-Means clustering the best number of

clusters is determined based upon the maximum choice which satisfies a given cluster validation index [1].
Y. Liu, Z. Li, H. Xiong, X. Gao and J. Wu proposed eleven CVI for crisp clustering based on five facts. They are monotonicity, noise, density, subclusters and skewed distributions and out of 11 indices S-dbw was the best to produce the best solution on all these five facts. These 11 indices were tested on four algorithms of four different categories namely K-Means of prototype-based, DBSCAN of density-based, Hierarchical Agglomerative of Average link based and Chameleon of graph-based [2] algorithms.

S. Saitta, B. Raphael and I.F.C. Smith proposed a score functions to estimate the number of clusters in a dataset. This Score Function is based on inter and intra-cluster related distances. The score function (SF) was based on two terms: "the distance between clusters and the distance inside a cluster" and "within class distance". The SF score is between [0, 1] and was tested for artificial and multidimensional real data sets using K-Means algorithm. The SF and four CVI's Dunn, Davies-Bouldin, Silhouette and Maulik-Bandyopadhyay were tested and found that SF outperformed all these indices and was best suited to determine perfect and unique clusters. The SF was able to estimate correctly the number of clusters in several artificial and real-life datasets [3].

Mustakim proposed K-Means and PCA algorithm for 3 indices. Three indices were Davies Bouldin Index (DBI), Silhouette Index (SI) and Dunn Index (DI). Here PCA algorithm performed well than K-Means. PCA K-Means is capable to determine a lower value for DBI, and for SI and DI they found that the patterns change continuously. They also showed how many clusters and features can be used [4].
C. Legany, S. Juhasz and A. Babos have validated on Dunn, S_dbw and SD and found that Dunn and S_dbw can find well separated clusters and found that the Dunn index is the most time-consuming and also reported that the SD index is the fastest [5].

T. Karkkainen and S. Jauhiainen proposed a new index KCE (k times the Clustering Error) and compared this with four indices namely Calinski-Harabasz, Silhouette, Pakhira-Bandyopadhyay-Maulik (PBM) and Wemmert-Gancarski. These indices suggested best clusters. The new index KCE suggested zero or one cluster and found that KCE worked for clusters of spherical shapes [6].

L.J.Deborah, R.Baskaran and A.Kannan concentrated on the sensitive characteristics of original data set especially noise and dimension. Hence, the performance of partitioning data sets, whether high dimension or with outliers, shall work better when considering the geometry of clusters formed for multiple dimensions and also for mixed type datasets [7].

S.Jauhiainen, J. Hamalainen and T.Karkkainen proposed the framework for prototype-based clustering. The indices used are KCE, WB, Calinski–Harabasz (CH), Wemmert–Gançarski (WG), Davies–Bouldin (DB), Pakhira, Bandyopadhyay and Maulik (PBM), and Ray–Turi (RT) index. The index values are calculated for the clustering algorithm with the specified distance. They used city-block distance for K-Medians, Squared Euclidean distance for the K-Means, and Euclidean distance for the K-Spatial Medians. From their experiments and results S_Dbw proved to produce the correct number of clusters. The CVI results are based and dependent on the datasets. Each CVI is suited best for a certain type of data. Varying datasets have varying results [8].

M. Charrad, Y. Lechevallier, M.B. Ahmed and G. Saporta proposed a framework for block clustering where rows and columns pair is considered. Let r represents the number of rows and c represents the number of columns. When r=c Dunn, BH, CH, DB and Sihouette indices identify the best cluster pair (r,c) and when r≠c BH index was found to be the best index [9].

J.Baarsch and M. E. Celebi in their paper used the most widely used technique for clustering the K-Means algorithm, dependent on the choice of the number of clusters k. In unsupervised situations, experiments were conducted to evaluate commonly used cluster validity measures, including Dunn, Davies-Bouldin, Calinski-Harabasz, Silhouette, Point Bi-serial, PBM, and Sum-of-Squares. These measures were applied to K-Means clustering and found that the Sum-of-Squares method was found to be the most effective followed by Silhouette whereas Calinski-Harabasz and Davies-Bouldin both showed moderate results and other indices performed poorly [10].

III. CLUSTER VALIDATION CATEGORY

CVI is categorized into Internal, External and Relative validation. Internal validation uses only the internal intrinsic properties of corpus with unknown class labels like the clustering algorithm (Eg. SSE, Dunn, Silhouette etc.) Internal indices are used to measure the goodness of a clustering structure without external information [11]. External validation uses the similarity measures of clusters for well-known class labels (Eg. The Czekanowski-Dice index, The Folkes-Mallows index, The Hubert index, The Jaccard index, The Phi index, The Rand index, Entropy and F-measures etc. Relative Validation compares the clustering structure of different clustering algorithm using either external or internal indices measures. Indices from this group are used for deciding which clustering scheme fits the data best.

INTERNAL CLUSTERING VALIDATION MEASURES

Internal indices operate on the proximity measures which should be non-negative, symmetrical and fulfil the criteria such as compactness, separation and well connectives. Internal validation measures are often based on the following three criteria. Compactness defines the closeness of objects within the same cluster and are measured based upon their variance. Lower variance indicates better and good compactness. Separation defines how distinct or well-separated a cluster is from other clusters. It maximizes the inter cluster distance, minimizes the intra cluster distances and maximizes / minimizes the measures based on density. Connectivity defines to what extent the objects are placed in the same cluster as their nearest neighbours. The connectivity measure has a value ranging between zero and infinity.

IV. VALIDITY INDICES

The goal of cluster analysis is that the objects within a group be similar to one another and different from the objects in other groups. The greater the similarity within a group and the greater the difference between groups, the better or more distinct is the clustering. So, the measure of goodness of the K-Means clustering accuracy has been defined by the ratio BetweenSS (BSS) to TotalSS (TSS), where SS stands for Sum of Squares. The proposed algorithms are executed for different feature set on the same corpus and the results are compared with the clustering accuracy. The best clustering accuracy with dimensionality reduction is recorded.

Internal Cluster Validation (ICV)

i) C-Index (CI)

C-Index is the ratio of the difference between the sum of distances over all pairs of patterns from the same cluster S and the sum of the l smallest distances S_{min} out of all pairs, divided by the difference between the sum of the l largest distances S_{max} out of all pairs and the sum of the l smallest distances out of all pairs S_{min} [12]. Small values of C indicate good clustering.

This index is defined as follows:

$$CI = \frac{S - S_{\min}}{S_{\max} - S_{\min}} \qquad (1)$$

ii) Davies-Bouldin index (DB)

The Davies-Bouldin index [13] computes the Inter-cluster distance and Intra-cluster distance. It computes the similarities for each cluster and assigns the highest value obtained to C as its cluster similarity. Then the DB index can be obtained by averaging all the cluster similarities. By obtaining a small index value it achieves a better clustering and by minimizing this index value it achieves a better partition.

Let μ_i denote the mean of cluster

$$\mu_i = \frac{1}{|C_i|} \sum_{x_j \in C_i} x_j \qquad (2)$$

and σ_i denote the dispersion of the points in the cluster C_i around it's mean μ_i.

$$\sigma_i = \sqrt{\frac{\sum_{x_i \in C_i} \alpha(x_j, u_i)}{|C_i|}} \qquad (3)$$

The Davies-Bouldin measure for pair of clusters C_i is defined as follows:

$$\mathrm{DBI}_{ij} = \frac{\sigma_i + \sigma_j}{\alpha(\mu_i, \mu_j)} \qquad (4)$$

DBIij measures the compactness of clusters compared to the distance between the cluster means.

$$DB = \frac{1}{|C|} \sum_{i=1}^{|C|} \max\{DBI_{ij}\} \qquad (5)$$

That is, for each cluster C_i pick another cluster C_j which produces the largest value of DBI_{ij} ratio. The smaller the DB value the better the clustering, because this means that clusters are well-separated (the distance between cluster means is large) and each cluster is compact (has a small spread). For a good clustering Davies-Bouldin index will have a small value.

iii) Sd_Scat Index(SS)

This index [5,14] measures the homogeneity and compactness of the clusters. It defines two quantities:

i) S: Average scattering for the clusters.

The vector of variances for each variable in the data set and cluster is given below. It is a vector V of size p defined by:

Variance of the dataset

$$\sigma_x^P = \frac{1}{n} \sum_{k=1}^{n} (x_k^P - \bar{x}^P)^2$$

$$\sigma(x) = \begin{bmatrix} \sigma_x \\ .. \\ .. \\ \sigma_x^d \end{bmatrix}$$

Variance of the cluster

$$\sigma_n^P = \frac{1}{\|C_i\|} \sum_{k=1}^{n} (x_k^P - \bar{v}_i^P)^2$$

$$\sigma(v_i) = \begin{bmatrix} \sigma_{v_i} \\ .. \\ .. \\ \sigma_{v_i}^d \end{bmatrix}$$

(6)

The average scattering for clusters is defined as

$$S = \frac{1}{n_c} \sum_{i=1}^{n_c} \frac{\|\sigma(v_i)\|}{\|\sigma(x)\|} \qquad (7)$$

D: Total separation between clusters. It is defined as follows. Let $\underbrace{\max}_{i,j=1,..n} \left(\|v_j - v_i\| \right)$ and $\underbrace{\min}_{,j=1,..n} \left(\|v_j - v_i\| \right)$ denote the largest and the smallest distance respectively between the

centres of the cluster. The total separation between clusters is calculated as follows:

$$D = \frac{\max\limits_{i,j=1,..n}\left(\|v_j - v_i\|\right)}{\min\limits_{,j=1,..n}\left(\|v_j - v_i\|\right)} \sum_{t=1}^{k}\left(\sum_{z=1,z\neq t}\|v_t - v_z\|\right)^{-1} \quad (8)$$

$$SS = \alpha S + D \quad (9)$$

the SS index is defined as were α is a weight equal to the value of D obtained for the partition with the greatest number of clusters. The value of this index is the summation of these two terms and the optimal number of clusters can be obtained by minimizing the value of SS.

iv) S_Dbw Index (SD)

The S_Dbw index [15] measures the intra-cluster variance and the inter-cluster variance. This index relies on the notion of density of points belonging to two clusters which defines a limit value σ, Square root of the sum of the norms of the variance vectors/Number of clusters.

The intra cluster variance measures the average scattering of clusters and it is described by equation (10).

$$\sigma = \frac{1}{K}\sqrt{\sum_{k=1}^{K}\|V^k\|} \quad (10)$$

The density k_0 for a given point, relative to two clusters C_k and C_{k0} , is equal to the number of points in these two clusters whose distance to this point is less than σ. Geometrically, this amounts to considering the radius σ centred at the given point and counting the number of points of C_k located in this centre.

The inter–cluster density is defined as follows.

$$BCD = \frac{1}{n_c(n_c-1)}\sum_{i=1}^{n_c}\left\{\sum_{j=1,i\neq j}\frac{density(u_{ij})}{\max\{density(v_i), density(v_j)\}}\right\} \quad (11)$$

S-Dbw is the Sum of the mean dispersion in the clusters (S) and Between-cluster density (BCD). The S_Dbw index indicates "good" clustering and reliable evaluation of clustering results. In S-dbw, lower index value indicates better clustering.

v) Calinski and Harabasz Index (CH)

The CH index [16] is defined as:

$$C = \frac{BGSS/(K-1)}{WGSS/(N-K)} = \frac{N-K}{K-1}\frac{BGSS}{WGSS} \quad (12)$$

where $BGSS = \sum_{i=1}^{k}|C_i|\|\mu_i - \mu\|^2$ is the sum of squares among the clusters. μ_i is the mean of the i[th] cluster, μ is an overall mean of sample data and $\|\mu_i - \mu\|$ is the Euclidean distance between the two vectors and $WGSS = \sum_{i=1}^{k}\sum_{x_j\in C_i}\|x_j - \mu_i\|^2$ is the sum of squares within the clusters, N is the number of data points and K is the number of clusters and measures the compactness based on the sum of distances between objects and their cluster centre. The larger the value of Calinski-Harabasz index, the better the quality of the clustering scheme. Good clustering has large between-cluster variance BGSS and a small within-cluster variance WGSS.

vi) Dunn Index (DU)

Researchers used Dunn index [17], which is the ratio of minimal cluster distance and maximal cluster diameter.

$$D = \min_{1=1,..n_c}\left\{\min_{j=i+1,..n_c}\left(\frac{d(c_i,c_j)}{\max\limits_{k=1,..n_c}(diam(c_k))}\right)\right\} where$$

$$d(c_i,c_j) = \min_{x\in C_i, y\in c_j}\{d(x,y)\} \text{ and } \mathrm{diam}(c_i) = \max_{x.y\in c_i}\{d(x,y)\} \quad (13)$$

Larger the value of index presents the more compact and well separated clusters.

vii) Gamma Index (GA)

Gamma index is defined by [18]:

$$GI = \frac{S^+ - S^-}{S^+ + S^-} \quad (14)$$

The number s^+ represents the number of times a distance between two points which belong to the same cluster is strictly smaller than the distance between two points not belonging to the same cluster. The number s^- represents the number of times a distance between two points lying in the same cluster is strictly greater than a distance between two points not belonging to the same cluster. Values of this index is in the interval [-1, 1]. Large values of GI indicate a good clustering.

viii) The Silhouette Index (SI)

The Silhouette validation technique [19] calculates the silhouette width for each sample, average silhouette width for each cluster and overall average silhouette width for a total data set. The average silhouette width could be applied for evaluation of clustering validity and also decides how

good the number of selected clusters is. To construct the silhouettes SI(i) the following formula is used:

$$SI(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (15)$$

where $a(i)$ = average distance between point i and to all other points in other clusters

b(i)= minimum of the average distance of i to points in any other cluster

SI(i) can obtain values in interval [−1, 1].

The value of SI is 1 indicates that x_i is close to points in its assigned cluster and far from other clusters, 0 indicates that x_i is close to boundary and -1 indicates is close to another cluster than its own cluster.

V. METHODOLOGY

The benchmark text corpora were collected to experiment the proposed framework Feature Extraction K-Means Evaluation Measures (FE-KMEANS-EM) depicted in Figure1. Pre-processing was conducted by tokenization, removing numbers, punctuation and stopwords. The text corpora represented in vector space model whose elements are TF-IDF. This is converted into Document-Term-Matrix (DTM) where the rows are corpus documents and columns are features. Document frequency feature Selection (DFFS) performed on DTM. The corpus was clustered applying K-Means (DK) algorithm. Feature extraction is performed using PCA and KPCA. K-Means algorithm clusters text corpora on extracted features. The various kernel functions used in KPCA are Radial basis function (DKK-RBF), polynomial (DKK-poly), Laplace (DKK-Laplace), Bessel (DKK-Bessel) and Sigmoid Tangent (DKK-Tanh). The quality of the obtained clustered results are evaluated and compared with several indices mentioned. The internal validation measures are based upon maximal and minimal return values. The relative validity measure (RVM) is determined by best accuracy recorded for the clustering algorithms.

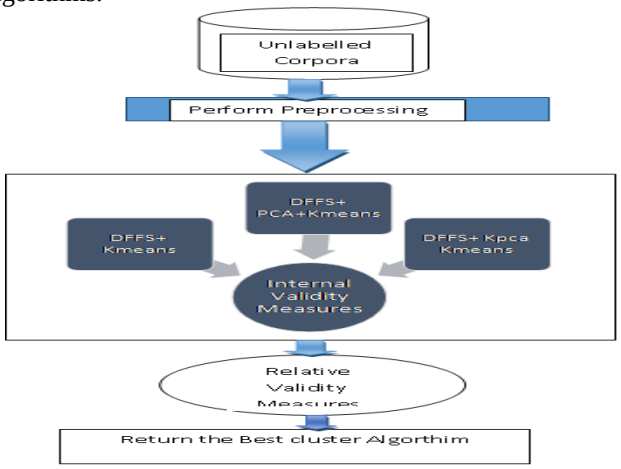

Figure 1: Proposed FE-KMEANS-EM Framework

Procedure: FE-KMEANS-EM
Input : Unlabelled Dataset for clustering
Output : Clustering accuracy with RVM
Step 1: Collect text corpora.
Step 2: Perform pre-processing
Step 3: On each corpora perform three phases of clustering
i. DFFS+K-Means (DK)
ii. DFFS+PCA+ K-Means (DPK)
iii. DFFS+KPCA+ K-Means (DKK)
Step 4: Calculate the clustering accuracy.
Step 5: Compute the Evaluation Measure (EM).
Step 5: Record the best EM.
Step 6: Return the cluster that matches clustering accuracy with best EM.

VI. CORPUS USED

The unlabelled text corpus used ranges from few kilobytes to several megabytes in size. They are the BBC Sports, BBC Abstract, Newsgroup20 divided into four categories, Reuters, C50, Enron ranging from Enron1 to Enron6, Lingspam and Ohshamud a medical abstract dataset. Total of about 16 text corpora were taken for analysis [20].

VII. USED ENVIRONMENT AND LIBRARIES

R programming environment [21] is taken for implementing our proposed work and validating the cluster with respect to internal relative criteria. The library "clustercrit" is used for measuring the various internal / external validation measures [22].

VIII. EXPERIMENTAL RESULTS

Collect unlabelled documents needed for study. This is in unstructured format. To make it good for cluster analysis convert into structured format. Use vector space model to represent this in Document-Term matrix. This Document Term matrix is large to perform clustering hence perform document frequency feature selection (DFFS) where the frequency selected ranges from 40-60. After this process of DFFS execute the K-Means (DK). Using DFFS, feature extraction using PCA with K-Means (DPK) and KPCA with K-Means (DKK) is performed and the clustering accuracy results recorded. The internal and relative validity measures are used.

The internal cluster indices like C_INDEX (CI), DAVIES_BOULDIN (DB), SD_SCAT (SS) and S_DBW (SD) returns the minimum value and indices like CALINSKI_HARABASZ (CH), DUNN (DU), and GAMMA (GI) and SILHOUETTE (SI) returns the maximum value. These indices are used to measure the quality of a clustering result by comparing all these three algorithms using relative indices. Hyphen in the table indicates that there

was no implementation of ICV's or that the calculation failed producing NaN, Inf, -Inf. The values are recorded for internal validity measures. The values for various clustering algorithm with different internal indices are as shown in Table 1. Table 1.1-Table 1.16 gives the tabulation of the ICV values obtained for various datasets.

Majority of the indices like CI, DB, CH, GI and SI proved to be the best indices for almost all the clustering algorithms implemented. Index like SD was satisfied for the DKK-Bessel, DKK-Tanh and DKK-Rbf whereas the DU index supports K-Means (DK) and PCA (DPK) type of clustering. Table 2 tabulates the best results of clustering techniques with CVI for all datasets and figure 2 graphically shows the best clustering accuracy.

In general the clustering algorithms that satisfied for the datasets are: DKK-Poly clustering satisfying for BBC Abstract, Ng20group1 to Ng20group4, Reuters, Enron5, Enron6, Lingspam and Ohshamud datasets, DKK-Tanh clustering satisfying for Enron1, Enron2 datasets, DKK-Bessel clustering satisfying for Enron3 and Enron4 datasets, DKK-Laplace clustering satisfying for C50 dataset and DKK-Rbf clustering satisfying for BBC sports dataset. Table 2 tabulates the best results of clustering techniques with CVI for all datasets and figure 2 graphically shows the best clustering accuracy. For all 16 datasets the overall results proved that the Kernel type of clustering (DKK) is the best.

Table 1: Cluster Validation index for different unlabelled corpora

Table 1.1. CVI for BBC Sports

BBC Sports	CI	DB	SS	SD	CH	DU	GI	SI
DK	0.0103	0.3933	0.9499	1.3225	1702.6630	0.0257	0.9955	0.6315
DPK	-	-	-	-	--	-	-	-
DKK-Rbf	0.0036	**0.2461**	**0.0854**	**0.0854**	**4488.4520**	0.0379	0.9981	-
DKK-Poly	0.0015	0.3980	0.9197	1.3651	1939.3260	0.0357	0.9991	0.6137
DKK-Tanh	0.0152	0.4374	0.5911	1.1016	1316.9810	0.0343	0.9919	0.6414
DKK-Bessel	**0.0014**	0.3857	0.6839	0.9128	1960.6190	0.0424	**0.9992**	0.6434
DKK-Laplace	0.0091	0.3592	0.1752	0.7106	3128.8100	**0.1024**	0.9958	**0.7093**

Table 1.2 . CVI for BBC Abstract

BBC Abstract	CI	DB	SS	SD	CH	DU	GI	SI
DK	0.4358	2.4699	3.5641	-	23.7107	0.0923	0.3306	0.1790
DPK	0.4717	2.0862	2.3917	-	17.3759	**0.1122**	0.3591	0.2457
DKK-Rbf	0.1183	6.6932	3.5217	-	21.0990	0.0596	0.7996	0.1122
DKK-Poly	**0.0011**	**0.4735**	3.1170	3.7332	**3376.8890**	0.0328	**0.9998**	**0.6313**
DKK-Tanh	0.0684	0.5366	**0.7563**	**1.3791**	1938.0420	0.0023	0.8669	0.5332
DKK-Bessel	0.0412	0.9018	2.3369	4.4561	1237.1030	0.0012	0.8662	0.3560
DKK-Laplace	0.0443	0.7636	0.2968	1.5510	1851.1050	0.0024	0.8506	0.4332

Table 1.3. CVI for NG20-group1

Ng20-group1	CI	DB	SS	SD	CH	DU	GI	SI
DK	0.5226	4.7017	5.4852	-	11.6900	**0.1446**	0.1284	0.0786
DPK	0.4893	2.5135	3.4294	-	14.8907	0.0942	0.2321	0.2779
DKK-Rbf	0.2155	12.4517	5.1399	-	13.7809	0.0791	0.6444	0.0603
DKK-Poly	0.0016	**0.3296**	2.5207	2.5267	**16107.3900**	0.0519	**0.9997**	**0.7598**
DKK-Tanh	**0.0664**	0.6790	**0.7116**	**2.4605**	3507.1850	0.0013	0.7936	0.4584
DKK-Bessel	0.0703	0.6511	1.5501	3.4343	3907.5060	0.0006	0.8054	0.5244
DKK-Laplace	0.1463	0.5895	2.3309	3.3753	5838.9380	0.0004	0.5518	0.4948

Table 1.4. CVI for NG20-group2

Ng20-group2	CI	DB	SS	SD	CH	DU	GI	SI
DK	0.4620	2.4002	4.7522	-	12.1708	0.1504	0.2038	0.1707
DPK	0.4174	2.2460	4.1180	-	20.1204	**0.2028**	0.2785	0.3069
DKK-Rbf	0.2479	15.5478	3.9258	-	18.4082	0.0662	0.5760	0.1644
DKK-Poly	0.1034	**0.3002**	1.9912	**3.0653**	**13243.4800**	0.0043	**0.9747**	**0.7403**
DKK-Tanh	0.0793	0.6340	**1.2788**	3.0821	4252.1460	0.0019	0.7838	0.4649
DKK-Bessel	**0.0483**	0.5491	4.0963	7.5576	4007.8920	0.0001	0.8405	0.5328
DKK-Laplace	0.0788	0.5276	2.1016	5.4812	5263.7680	0.0003	0.8014	0.5780

Table 1.5. CVI for NG20-group3

Ng20-group3	CI	DB	SS	SD	CH	DU	GI	SI
DK	0.4913	3.0998	3.8770	-	12.8549	**0.1603**	0.2023	0.1215
DPK	0.4764	4.1664	5.0049	-	16.6411	0.1364	0.2045	0.1555
DKK-Rbf	0.2328	19.0644	2.6582	-	12.5147	0.1052	0.5023	0.0918
DKK-Poly	**0.0007**	**0.3063**	4.2834	4.2868	**15277.4400**	0.1092	**0.9998**	**0.7362**
DKK-Tanh	0.0417	0.6732	**0.8653**	**1.5736**	3676.1980	0.0026	0.9347	0.4586
DKK-Bessel	0.0505	0.4502	1.0414	2.7283	12544.7700	0.0004	0.8399	0.6701
DKK-Laplace	0.0762	0.5426	2.2910	5.6062	5919.0530	0.0002	0.7616	0.5353

Table 1.6. CVI for NG20-group4

Ng20-group4	CI	DB	SS	SD	CH	DU	GI	SI
DK	0.4850	2.5319	4.2595	-	10.8776	**0.1257**	0.2380	0.2027
DPK	0.4489	2.9572	3.5916	-	16.7063	0.1147	0.2931	0.2746
DKK-Rbf	0.2171	12.9983	3.3765	-	13.7046	0.0587	0.6397	0.0511
DKK-Poly	**0.0183**	0.7082	1.8715	-	**15406.8000**	0.0011	**0.9929**	**0.6430**
DKK-Tanh	0.0582	**0.6023**	0.5970	**1.5352**	3967.2070	0.0025	0.8429	0.4977
DKK-Bessel	0.0486	0.6127	3.5952	9.9227	5376.7230	0.0002	0.8279	0.5117
DKK-Laplace	0.0461	0.9001	**0.2171**	2.2626	4359.0760	0.0016	0.8459	0.3858

Table 1.7. CVI for Reuters

Reuters	CI	DB	SS	SD	CH	DU	GI	SI
DK	0.3716	2.0634	3.3798	-	18.7002	**0.1602**	0.1933	0.3179
DPK	0.3086	4.0800	3.8293	-	24.1614	0.1086	0.3608	0.2076
DKK-Rbf	0.4611	11.8102	2.3940	-	16.2773	0.0918	0.1066	0.0638
DKK-Poly	**0.0014**	**0.5206**	**0.0588**	-	**111376.8000**	0.00001	0.8967	**0.5951**
DKK-Tanh	0.0067	0.6340	0.1031	-	10590.3500	0.0005	0.8870	0.4878
DKK-Bessel	0.0074	0.6983	0.0875	-	20083.3800	0.0007	0.9454	0.4721
DKK-Laplace	0.0062	0.6116	0.0453	-	32624.4100	0.0004	**0.9523**	0.5414

Table 1.8. CVI for C50

C50	CI	DB	SS	SD	CH	DU	GI	SI
DK	0.1211	1.5585	3.2381	-	36.3107	**0.0193**	0.7080	-
DPK	0.2651	5.9922	3.1720	-	8.7153	0.0035	0.3413	-
DKK-Rbf	0.2708	7.3044	2.2833	-	5.7629	0.0027	0.3162	-

DKK-Poly	**0.0009**	0.7631	0.1621	-	1363.7330	0.00002	0.9093	0.3715
DKK-Tanh	0.0015	0.7671	**0.0247**	-	**6017.7560**	0.0002	0.9309	0.3821
DKK-Bessel	0.0046	0.7979	0.0458	-	2562.5600	0.0004	0.8766	0.3582
DKK-Laplace	0.0029	**0.7527**	0.0298	-	2790.3270	0.0008	**0.9693**	**0.3905**

Table 1.9. CVI for Enron1

Enron1	CI	DB	SS	SD	CH	DU	GI	SI
DK	0.4199	1.5425	1.1550	1.1550	13.3734	0.2822	0.0740	0.2679
DPK	0.2615	0.7023	0.5530	0.5530	91.7470	**0.4083**	0.5116	0.5859
DKK-Rbf	0.3542	21.0070	1.2988	2.2744	6.1262	0.0328	0.2927	0.0711
DKK-Poly	0.2996	**0.0527**	0.5244	0.5252	1522.8790	0.0462	0.9725	**0.9662**
DKK-Tanh	**0.0077**	0.0759	**0.4791**	**0.4795**	**3310.4590**	0.1608	**0.9966**	0.9492
DKK-Bessel	0.2259	0.4559	0.5301	0.9278	1427.6090	0.0017	0.8759	0.6922
DKK-Laplace	0.2205	0.8941	1.0141	2.0162	986.5751	0.0016	0.6778	0.4859

Table 1.10. CVI for Enron2

Enron2	**CI**	**DB**	**SS**	**SD**	**CH**	**DU**	**GI**	**SI**
DK	0.4351	1.5899	1.6883	1.6883	13.3227	0.0627	0.0768	0.2576
DPK	0.3084	2.1919	5.7172	5.7172	39.0421	**0.3378**	0.4542	0.2209
DKK-Rbf	0.4077	29.2192	1.1666	-	3.9985	0.0287	0.1978	0.0575
DKK-Poly	0.6774	0.5442	1.9747	1.9794	265.3494	0.0459	0.9504	0.6479
DKK-Tanh	0.2381	**0.2865**	0.8895	**0.8940**	**2815.9640**	0.0488	**0.9543**	**0.8027**
DKK-Bessel	0.2363	0.4844	**0.6413**	1.2899	919.5026	0.0005	0.8931	0.6702
DKK-Laplace	**0.2286**	0.8006	0.9520	2.0647	706.9899	0.0005	0.7362	0.5273

Table 1.11. CVI for Enron3

Enron3	**CI**	**DB**	**SS**	**SD**	**CH**	**DU**	**GI**	**SI**
DK	0.4032	2.0538	2.9798	2.9831	14.6872	0.0463	0.1431	0.1943
DPK	0.2947	2.0416	5.7791	5.7791	38.4209	0.1595	0.4515	0.2314
DKK-Rbf	0.4232	22.6275	1.1719	2.1618	5.0898	0.0399	0.1419	0.0981
DKK-Poly	0.0025	1.0304	23.1360	23.1439	2629.8460	0.0562	0.9999	0.5668
DKK-Tanh	0.0187	0.0969	0.6132	0.6136	2799.9180	0.0917	0.9977	0.9284
DKK-Bessel	**0.0001**	**0.0655**	**0.2706**	**0.2712**	**9299.2300**	**0.2364**	**1.0000**	**0.9571**
DKK-Laplace	0.2731	1.0245	1.0214	2.1563	366.3097	0.0003	0.6940	0.4747

Table 1.12. CVI for Enron4

Enron4	**CI**	**DB**	**SS**	**SD**	**CH**	**DU**	**GI**	**SI**
DK	0.4137	2.9086	3.2802	3.2802	13.6206	0.2751	0.1401	0.1054
DPK	0.2851	1.0679	1.4704	1.4704	39.1343	0.4045	0. 4461	0.4346
DKK-Rbf	0.3646	2.3103	4.1983	4.1983	40.7026	0.3588	0.2942	0.2117
DKK-Poly	0.0020	1.0029	21.1023	21.1110	2696.3990	0.4388	0.9999	0.6314
DKK-Tanh	0.3381	0.4841	1.3841	1.4193	1347.6980	0.0051	0.9036	0.6816
DKK-Bessel	**0.00002**	**0.1184**	**0.4885**	**0.4891**	**8692.1450**	**0.4612**	**1.0000**	**0.9149**

DKK-Laplace	0.2420	0.9642	1.0758	2.2668	511.6021	0.0012	0.7298	0.4972

Table 1.13. CVI for Enron5

Enron5	CI	DB	SS	SD	CH	DU	GI	SI
DK	0.3043	1.0414	1.3565	1.3565	64.5377	0.3659	0.4357	0.4565
DPK	0.2707	2.0133	5.120	5.1202	63.5300	0.3659	0.5371	0.2724
DKK-Rbf	0.3934	16.5429	1.2918	2.2918	10.8016	0.0247	0.2189	0.0651
DKK-Poly	**0.0026**	0.9891	14.0199	14.0231	**3136.6900**	**0.4503**	**0.9997**	0.6406
DKK-Tanh	0.1184	**0.1192**	**0.4163**	**0.4166**	3042.8550	0.0842	0.9486	**0.9158**
DKK-Bessel	0.2477	0.5530	0.6488	1.1990	1059.0350	0.0014	0.8745	0.6538
DKK-Laplace	0.2524	0.8353	0.8740	2.0321	793.3212	0.0005	0.7121	0.5133

Table 1.14. CVI for Enron6

Enron6	CI	DB	SS	SD	CH	DU	GI	SI
DK	0.4783	1.7997	2.5831	2.5831	13.6270	**0.2780**	0.0389	0.2728
DPK	0.3223	2.1470	5.8562	5.8562	51.6193	0.1561	0.4769	0.2133
DKK-Rbf	0.3531	2.6434	5.3318	5.3545	50.0178	0.2468	0.3844	0.1645
DKK-Poly	**0.0023**	1.0058	21.2329	21.2397	**3181.2950**	0.1517	**0.9999**	0.5998
DKK-Tanh	0.2342	**0.3179**	1.6560	**1.6631**	2793.7450	0.0193	0.9514	**0.7664**
DKK-Bessel	0.2624	0.7906	**0.7794**	1.6913	923.3860	0.0004	0.7597	0.5496
DKK-Laplace	0.2435	0.9543	1.1748	2.2573	632.1692	0.0002	0.6433	0.4652

Table 1.15. CVI for Lingspam

Lingspam	CI	DB	SS	SD	CH	DU	GI	SI
DK	0.2688	1.7423	11.6454	-	16.6315	**0.1225**	0.6460	0.1967
DPK	0.2506	5.3556	3.2926	-	16.3514	0.0333	0.5081	0.1972
DKK-Rbf	0.1158	9.1791	3.6523	-	15.5582	0.0243	0.7462	0.0721
DKK-Poly	0.0166	**0.4579**	0.3362	-	**15349.8000**	0.0006	0.9067	**0.6385**
DKK-Tanh	0.0299	0.6053	0.4928	**2.5265**	3962.6490	0.0010	0.8158	0.4867
DKK-Bessel	**0.0123**	0.5707	0.2528	-	10510.6800	0.0004	0.8838	0.5176
DKK-Laplace	0.0187	0.5289	**0.1377**	4.0697	2542.6150	0.0013	**0.9305**	0.5687

Table 1.16. CVI for Ohshamud

Ohshamud	CI	DB	SS	SD	CH	DU	GI	SI
DK	0.4758	1.9428	3.2438	-	14.5514	0.1275	0.3114	0.2587
DPK	0.5423	2.0035	1.6164	-	13.4007	**0.2202**	0.1551	0.2611
DKK-Rbf	0.1728	13.3503	2.7752	-	12.8126	0.0676	0.6548	0.0634
DKK-Poly	**0.0008**	**0.4266**	3.5428	4.0202	**3454.0930**	0.0170	**0.9998**	**0.6398**
DKK-Tanh	0.0358	0.5924	**0.6166**	**1.4353**	2884.3900	0.0032	0.8662	0.4969
DKK-Bessel	0.0544	0.6593	1.2721	3.4462	1691.2430	0.0006	0.8108	0.4510
DKK-Laplace	0.0644	0.5981	1.3708	3.5528	1809.3540	0.0007	0.7689	0.4832

Table 2: Best result of clustering techniques with CVI

Cluster	Accuracy for various dataset															
	BBC Sports	BBC Abstract	Ng20-group 1	Ng20-group 2	Ng20-group 3	Ng20-group 4	Reuters	C50	Enron 1	Enron 2	Enron 3	Enron 4	Enron 5	Enron 6	Lingspam	Ohshamud
DK	92.6	5.4	1.4	1.4	1.4	1.2	7.1	49.4	0.5	0.4	0.5	0.4	2.0	0.4	6.9	3.6
DPK	94.2	88.2	90.8	87.2	85.7	89.3	98.6	98.5	49.5	45.7	44.7	46.1	48.2	45.6	98.3	90.6
DKK-Rbf	**97.0**	74.2	81.1	86.4	85.5	49.5	98.8	98.2	9.5	6.0	6.0	42.0	10.7	41.0	93.7	82.9
DKK-Poly	93.4	**89.0**	**94.8**	**94**	**94.5**	**94.5**	**99.8**	97.3	34.2	42.2	45.1	47.1	**49.2**	**46.0**	**98.5**	**91.8**
DKK-Tanh	90.6	82.3	79.5	83.4	80.5	81.6	97.7	98.8	**53.1**	**46.7**	46.7	30.8	**48.6**	42.8	94.6	88.2
DKK-Bessel	93.5	74.8	76.9	82.6	93.4	85.7	98.8	98.6	32.8	19.9	**74.4**	**74.1**	24.8	19.8	97.9	81.4
DKK-Laplace	95.8	81.7	86.8	86.1	86.9	83.0	99.3	**99.4**	25.2	16.0	10.3	14.5	19.7	14.5	91.8	82.4

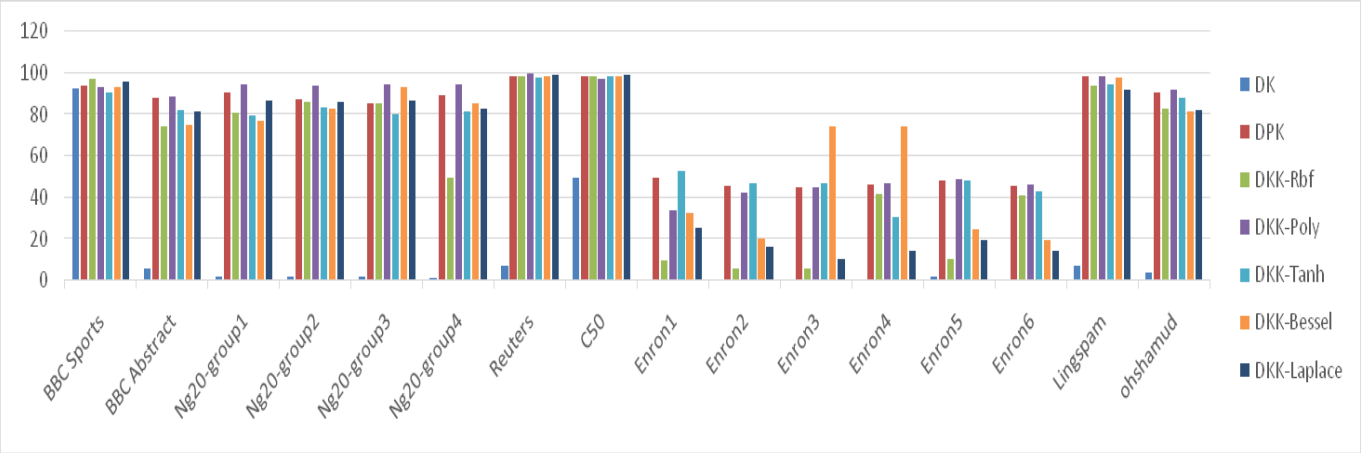

Figure 2: Clustering Accuracy

IX. CONCLUSION AND FUTURE SCOPE

In this paper several cluster validity indices have been evaluated and tested for the unlabelled corpora and tried to compare the accuracy of these validity indices. Majority of relative index value when comparing with the clustering algorithms the proposed feature extraction methods proved to be the best algorithm for clustering. In future, this study can be carried out for other clustering algorithms and results verified.

ACKNOWLEDGEMENT

We the authors are indebted to the R Core Team for providing the open source software with programming environment and necessary packages to implement this research work in a successful way.

REFERENCES

[1] K.P. Agrawal, S.Garg, P. Patel, "Performance Measures for Densed and Arbitrary Shaped Cluster", International Journal of Computer Science & Communication, vol 6, no.2, pp.338-350, 2015.

[2] Y. Liu, Z. Li, H. Xiong, X. Gao, J. Wu,"Understanding of Internal Clustering Validation Measure", 2010 IEEE InternationalConference on Data Mining Australia, pp.911-916, 2010.

[3] S. Saitta, B. Raphael, I.F.C. Smith, "A Bounded Index for Cluster Validity", Machine Learning and Data Mining in Pattern Recognition, Springer, Heidelberg, LNAI.4571, no.1, pp.174-187, 2007

[4] Mustakim, "Centroid K-Means clustering Optimization using Eigen vector principal component analysis", Journal of Theoretical and Applied InformationTechnology , vol.95, no.15, pp.3534-3542, 2017

[5] C. Legany, S. Juhasz, A. Babos, "Cluster Validity Measurement Techniques", Proceedings of the 5th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases, Spain, pp.388-393, 2006.
[6] T. Karkkainen, S.Jauhiainen, "A Simple Cluster Validation Index with Maximal Coverage", ESANN 2017 proceedingsEuropean Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning , i6doc.com publ, Belgium, pp.293-298, 2017.
[7] L.J.Deborah, R.Baskaran, A.Kannan, "A Survey on Internal Validity Measure for ClusterValidation", International Journal of Computer Science & Engineering Survey (IJCSES), vol.1, no.2, pp.85-102, 2010
[8] S.Jauhiainen,J.Hamalainen, T.Karkkainen, "Comparison of Internal Clustering Validation Indices for Prototype-Based Clustering Algorithms", ArticleAlgorithms , vol.10, no.105, pp.1-14, 2017.
[9] M. Charrad,Y. Lechevallier, M.B. Ahmed, G. Saporta, "On the Number of Clusters in Block Clustering Algorithms", Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference (FLAIRS 2010), pp.392-397, Florida
[10]J.Baarsch, M. EmreCelebi, "Investigation of Internal Validity Measures for K-Means Clustering", Proceedings of the Intn. Multiconference of Engineers and computer scientist, Hongkong, vol 1, 2012.
[11]A.Thalamuthu, I.Mukhopadhyay, X. Zheng, G.C. Tseng, "Evaluation and comparison of gene clustering method in microarray analysis", Bioinformatics, vol.22, no.19, pp.2405-2412, 2006.
[12]J.Schultz, L.Hubert, "Quadratic assignment as a general data analysis strategy", British Journal of Mathematical and Statistical Psychology, vol.29, no.2, pp.190-241,1976.
[13] D.W.Bouldin, D. L. Davies, "A cluster separation measure", IEEE Transaction on Pattern Analysis and Machine Intelligence PAMI-1,vol.3, no.2, pp.224-227, 1979.
[14]M. Halkidi, Y.Batistakis, M.Vazirgiannis, "Quality Scheme Assessment in the Clustering Process", Proc. of the 4th European Conference on Principles of Data Mining and Knowledge Discovery, pp.265-276, 2000
[15]M.Vazirgiannis, M.Halkidi, "Clustering validity assessment:Finding the optimal partitoning of a data set", Proceedings IEEE International Conference on data Mining, USA, pp.187-194, 2001.
[16] T. Harabasz, J. Calinski , "A dendrite method for cluster analysis", Communications in Statistics, vol.3, no.1, pp.1-27, 1974
[17] J.Dunn, "Well separated clusters and optimal fuzzy partitions", Journal of Cybernetics, vol.4, no.1, pp.95-104, 1974
[18] F. B Baker, L. J.Hubert, "Measuring the power of hierarchical cluster analysis", Journal of the American Statistical Association, vol.70, no.349, pp.31-38, 1975
[19] P.J.Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis", Journal of Computaional and Applied mathematics, vol.20, pp.53-65, 1987
[20] T.SenthilSelvi, R.Parimala, "Improving Clustering Accuracy using Feature Extraction Method", International Journal of Scientific Research in Computer Science and Engineering (isroset) ,vol.6, no. 2, pp.15-19, 2018.
[21]R Core Team, "R: A language and environment for statistical computing", R Foundation for Statistical Computing, Vienna, Austria, pp.1-2673, 2018, https://www.R-project.org/.
[22] B. Desgraupes, " clusterCrit: Clustering Indices", R package, pp.1-34, 2018, https://CRAN.Rproject.org/package=clusterCrit

Authors Profile

Mrs. T. Senthil Selvi graduated Master of Science and M.Phil in Computer Science from Srimathi Indira Gandhi College, Trichy affiliated to Bharathidasan University and now is a Research scholar and currently working as Assistant Professor in Periyar E.V.R. College, Tiruchirappalli. She has a teaching experience of about 22 years. Her research interest is in the field of Web Mining, Artificial Intelligence and Information Retrieval.

Dr. R. Parimala graduated with M.Sc., Applied Science at the National Institute of Technology, (formerly Regional Engineering College) Tiruchirapalli in 1990. She received her M.Phil., Computer Science at Mother Teresa University, Kodaikanal in 1999. She started teaching in 1999 at National Institute of Technology and is currently working as Assistant Professor in Department of Computer Science, Periyar E.V.R.College (Autonomous), Tiruchirapalli. She completed her Ph.D. at National Institute of Technology, Tiruchirappalli. Her area of research interests include Neural Networks, Data mining and Optimization Techniques.

