International Journal of Engineering & Technology, *7 (3.29) (2018) 533-537*

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A contemplate report on clustering evaluation and nonlinear clustering in high-dimensional data

P. Penchala Prasad [1] *, Dr. F. Sagayaraj Francis [2], Dr. S. Zahoor-Ul-Huq [3]

[1] Asst. professor, G. Pulla Reddy Engineering College, Kurnool
[2] Professor, Pondicherry Engineering College, Puduchchery
[3] Professor, G. Pulla Reddy Engineering College, Kurnool
**Corresponding author E-mail: m.neelakanta@gmail.com*

Abstract

Every day people use large volumes of data, for future purpose data can be classified into different categories such as clusters. The main intension of the cluster is to divide unlabeled finite dataset in to different set of structures. Distribution of clusters classified into linearly independent clustering and non- linearly independent clustering. Non-linear independent clustering means at least one group with rounded boundaries or of arbitary figures. Many clustering algorithms don't calculate approximately interior clusters. Several indexes used and planned for different Scenarios. There is no combining procedure for cluster assessment. We reconsider the existing clustering quality process and measure is difficult context designed for high-dimensional clustering. Dimensionality affect dissimilar clustering value indexes in dissimilar modes; few are preferred, to establish clustering quality in several ways. We are discuss in this paper, clustering evaluation, internal criteria, cluster quality indices, comparison of various clustering algorithms, problems in analyzing high dimensional data, clustering techniques for high dimensional data and perspectives and future directions.

Keywords: *Linear Clustering; Non- Linear Clustering; High-Dimensional Data; Hubness; Data Clustering; Cluster Indexes; Internal Indices; External Indices; Distance Concentration.*

1. Introduction

Data collecting is important in a mixture of fields like computer science, medical science, social science, and economics etc, data clustering problem is divided in to linearly independent clustering and nonlinearly independent clustering.

2. Clustering evaluation

The evaluation of unsupervised learning is difficult, there is no goal model to compare with, the true result is unknown, it may depend on the context and the task to perform.

Evaluation of the cluster means to avoid finding patterns in noise. Clustering evolution can be done by the comparison of clustering algorithms and the comparison of different models/ parameters. Before clustering a data set we can test if there are actually clusters, we have to assess the proposition of the subsistence of patterns in the data versus a dataset uniformly distributed.

Cluster Quality criteria

We can use different methodologies, to evaluate the quality of cluster, they are external criteria, internal criteria and relative criteria.

Comparison with a model partition/ labelled data is called external criteria, quality measures based on the examples/ quality of the partition is called internal criteria and comparison with other clustering's is called relative criteria. Data clustering is very difficult to reach exact cluster configuration for some values or not exact to zero for some data sets. Can we reach the goal to solve the comparison of different clustering algorithms; It may use the different indexes. Each index quantifies different solutions. Finding the quality of data we can use the internal and external indexes.

3. Internal criteria

The indices are based on the model of the groups, we can use indices based on the attributes values measuring the properties of a good clustering are distance distribution and values distribution indices.

Indices:

Some of the indices correspond directly to the objective function optimized Quadratic error/ Distorsion (k-means), Log likelihood (Mixuture of Gaussians/EM), Calinski-Harabasz index, Davies-Bouldin criteria, Silhouette index(maximum class spread/ variance) etc.,

Clustering Quality Indexes:

The following indices are used for the finding compactness and separation between clusters.

1) Silhouette index:

Silhouette index evaluates the point wise quality estimations [1]. Each point is calculate approximately for a point derived from two quantities, $X_{i.p}$= average distance of same cluster and $Y_{i.p}$= average distance of different cluster.

$$SIL(z_q) = \frac{X_{i.p}-Y_{i.p}}{max_{X_{i.p},Y_{i.p}}}$$

$$SIL = \frac{1}{N}\sum_{i=1}^{N} SIL(z_q)$$

Silhouette index inflexible to range huge datasets.

2) Simplified Silhouette index:

Simplified Silhouette index is the estimation of Silhouette coefficient. It calculates both inter and intra cluster Centroid distances [2].

3) Dunn index:

It find the diameter of large cluster and small cluster distance ratio. [3].

$$DI = \min_{i,j\in\{1...k\},i\neq j}\left(\frac{\min_{x_p\in C_i}\min_{x_q\in C_j}\|x_p - x_q\|}{\min_{l\in\{1...k\}}\max_{x_p,x_q\in C_l}\|x_p - x_q\|}\right)$$

4) Davies Bouldin Index:

This used based on intra and inter cluster ratio distances. [4]

$$DBI = \frac{1}{K}\sum_{i=1}^{K}\underset{j\neq i}{max}\left(\frac{\frac{1}{|C_i|}\sum x_p\in C_i\|x_p - \overline{x_i}\| + \frac{1}{|C_j|}\sum x_q\in C_j\|x_q - \overline{x_j}\|}{\|\overline{x_i} - \overline{x_j}\|}\right)$$

Minimal values are to get superior cluster configurations using Davies-Bouldin index.

5) Isolation index:

Isolation index denied an average measure of neighbors in the data of their cluster label [5].

$$\delta_{i,k} = local\ neighborhood\ disagreement\ ratio$$

$$IS = \frac{1}{N}\sum_{i=1}^{N}(1 - \delta_{i,k})$$

A weighted version proposed later [6].

6) C-index:

C-index used in missing uniqueness from the intra cluster and external values [7]. Let $IN_{m,n}$ is the function, m and n equal returns 1 , 0 otherwise. The factor $\theta = \sum_{m,n\epsilon\{1...i\},m\neq n} IN_{m,n}\|x_m - x_n\|$. The θ can takes into the large and small values of the data, that is taken by final C-index.

$$CIndex = \frac{\theta - min\theta}{max\theta - min\theta}$$

7) $\sqrt[C]{K}$ Index:

It may take individual characteristics into corresponding assistances to distances of both inter and intra clusters. [8]. Assume $SS_z = \sum_{p=1}^{N}\|{x_p}^z - x^{-z}\|$ be the contribution of mean $\overline{x}$. The n cluster distance

$$SSS_z = SS_z - \sum_{i=1}^{K}\sum_{x_p\in C_i}({x_p}^z - x^{-z})^2.\ \sqrt[C]{K}\,Index = \frac{1}{d.\sqrt{K}}\sqrt{\frac{SSS_z}{SS_z}}$$

Ratio increasing with K.

8) Calinski-Harabasz index:

It discovers variance diffusions of both inter and intra It discovers variance diffusions of both inter and intra clusters. [9].

Let X_B is inter cluster matrix and X_W be the intra cluster matrix.

$$X_B = \sum_{i=1}^{K}|C_i|(\overline{x_\iota} - \overline{x})(\overline{x_\iota} - \overline{x})^T$$

$$X_W = \sum_{i=1}^{K}\sum_{x_p\in C_i}(\overline{x_\iota} - \overline{x})(\overline{x_\iota} - \overline{x})^T$$

$$CHI = \frac{trace(X_B)}{trace(X_V)}.\frac{N-K}{K-1}$$

It leads the high values, high variance expected to dense and well separated cluster configurations .

9) Fowlkes-mallows index:

It defines for the given data. $Precision = \frac{TP}{TP+FP}$ and $Recall = \frac{TP}{TP+FN}$ then $FM = \sqrt{Precision.Recall}$. This index gives better counterpart between compared partitions of the data. Hierarchical clustering uses the compared clustering outputs.

10) Goodman-Kruskal's index:

It is derivative distances of concordant and discordant pairs [10]. Goodman-Kruskal's index evaluates the distance of object to same cluster and objects belonging to dissimilar clusters. It represent D_+ (no. of concordant pairs) and D_- (no. of discordant pairs).

$$GKI = \frac{D_+ - D_-}{D_+ + D_-}$$

11) G_+ index:

Concordance and discordance data pairs are used this index. It is easy to derive, it takes discordant counts only. Let $n = \frac{N(N-1)}{2}$ be number of points. The G_+ index discordant distance total number distance comparisions, namely $G_+ = \frac{2S_-}{n(n-1)}$. The higher values to get lower clustering quality. $\overline{G_+} = 1 - G_+$ in experiments as an alternative.

12) McClain-Rao index:

It symbolizes the mean of intra cluster and inter cluster distances quotient. $Z_{p,q}$ is indicator function . In this function p, q returns 1 they are in same cluster, otherwise 0.

Let$C_d = \sum_{i=1}^{K}|A_i|(N - |A_i|)$

And

$$W_d = \sum_{i=1}^{K}\frac{|A_i|(|A_i| - 1)}{2}$$

$$MCRI = \frac{\sum_{p,q\in\{1...N\},p\neq q} Z_{p,q}\|x_p - x_q\|}{\sum_{p,q\in\{1...N\},p\neq q}(1 - Z_{p,q)})\|x_p - x_q\|}.\frac{C_d}{W_d}$$

13) PBM index:

PBM is calculated depending on cluster Centroids.

$$PBMI = \frac{\sum_{l=1}^{N}\|x_l - \overline{x}\|}{\sum_{x_m\in C_j}\|x_m - \overline{x_m}\|}.\frac{max_{i,j\in\{1...K\},i\neq j}\|\overline{x_i} - \overline{x_j}\|}{K}$$

$$PBMI = \left(\frac{\sum_{l=1}^{N}\|x_l - \overline{x}\|}{\sum_{x_m\in C_j}\|x_m - \overline{x_j}\|}.\frac{max_{i,j\in\{1...K\},i\neq j}\|\overline{x_i} - \overline{x_j}\|}{K}\right)^2$$

14) Point Biserial Index:

It is based on inter and intra cluster maximization condition [11]. Number of pairs of data $= \frac{N(N-1)}{2}$. Let C_d and W_d represents inter cluster and intra cluster pairs. Let intra cluster average distance represents d_W and inter cluster average distance represents d_b . Z_d the standard deviation.

$$PBSI = \frac{(d_b - d_w).\frac{\sqrt{w_d.c_d}}{n}}{Z_d}$$

15) Rand index:

Rand index compare with single pair label and substitute cluster configurations. Let the same cluster with the same label pairs denotes x , same cluster with different labels pairs be y, different clusters with different lables pairs be z and w.

$$RAND_I = \frac{p+s}{p+q+r+s}$$

An enhanced adaptation of the Rand Index anticipated by [12], considered as successful cluster validation indices by Adjusted Rand index [13].

$$AdjustedRandomIndex = \frac{\binom{N}{2}(p+s) - [(p+q)(p+s) + (r+s)(q+s)]}{\binom{N}{2}^2 - [(p+q)(p+r) + (r+s)(q+s)]}$$

16) SD index

It is a scatter and separation combination. Scatter denoted by W_S and separation denoted by W_D respectively. $SDI = \alpha W_S + W_D$.

It is evaluated under global vectors and cluster variance with more features. Variance vectors denote V_t and V_{Ci} for d-dimensional .for i→{1...N}. $W_S = \frac{\sum_{i=1}^{N} \left\| V_{C_i} \right\|}{K.\|V_t\|}$. on the previous dispense , centroid distance separation is defined ,

$$W_D = \frac{max_{i\neq j}\left\|\overline{x_i}-\overline{x_j}\right\|}{min_{i\neq j}\left\|\overline{x_i}-\overline{x_j}\right\|} \cdot \sum_{i=1}^{N} \frac{1}{\sum_{j=1,j\neq i}^{N}\left\|\overline{x_i}-\overline{x_j}\right\|} \cdot$$

17) τ index:

It corresponds to correlation between binary matrix and distance matrix, it identifies data points to be same cluster or not. Let $n = \frac{N(N-1)}{2}$ defined by N data points. Similarly let C_d represent the inter cluster and W_d represents the intra cluster pairs. Let$t_{bw} = \binom{C_d}{2} + \binom{W_d}{2}$ indicates number pairs.

$$\tau = \frac{S_+ - S_-}{\frac{T(T-1)}{2} - t_{bw} * \frac{T(T-1)}{2}}$$

$$\tau = \frac{S_+ - S_-}{\left(\frac{T(T-1)}{2} - t_{bw}\right)\frac{T(T-1)}{2}}$$

The study concludes that Silhouette, Davies-Bouldin and Calinski Harbasz index fine in a wide range of situations.

4. External criteria

External Criteria indices measure if a clustering is similar to a model partition P.

It is equivalent to have a labelled dataset, if we do not have a model these criteria are used to evaluate the results of using dissimilar parameters of different algorithms.

The indices main advantage is autonomous of cluster explanation: Any clustering algorithm can be used these means.

Indices

Many indexes are based on coincidence of each pair of examples in the groups of two Clusterings. The computations are based on we have four values

i) The two examples belong to same class in both Partitions(a)
ii) The two examples belong to same class in X, but not in P(b)
iii) The two examples belong to same class Partitions, but not in X(x)
iv) The two examples belong to different classes in both Partitions(d)

Rand/Adjusted Rand statistic, Jaccard Coefficient, Folkes and Mallow index these are used in external criteria.

1) Rand/ Adjusted Rand Statistic:

$$R = \frac{(p+s)}{(p+q+r+s)}$$

2) Adjusted Rand:

$$ARand = \frac{p - \frac{(p+r)(p+s)}{p+q+r+s}}{\frac{(p+r)+(p+q)}{2} - \frac{(p+q)(p+r)}{p+q+r+s}}$$

3) Jaccard Coefficient: $J = \frac{p}{p+q+r}$

4) Folkes and Mallow index : $FM = \sqrt{\frac{p}{p+q} \cdot \frac{p}{p+r}}$

5. Number of clusters

A topic related to cluster validation is to decide if the no. of clusters obtained is correct one, number of clusters is important specially for the algorithms that need this value as a parameter, the usual procedure is to compare the characteristics of clustering's of different sizes, usually internal criteria indices are used in this comparison, a graphic of this indices different no of clusters can illustrate no of clusters is more feasible.

Some of the internal validity indices can be used for this purpose: Calinski Harabasz index, Silhouette index. Using the within class scatter matrix other criteria can be defined: Hartigan index, Krzanowski Lai index.

Hartigan Index= $H(I) = \left[\frac{S_M(I)}{S_M(I+1)} - 1\right](n - I - 1)$

Krzanowski Lai index =$KL(I) = \left|\frac{DIFFERENCE(I)}{DIFFERENCE(I+1)}\right|$being

$$DIFFERENCE(I) = (I-1)^{\frac{2}{p}} S_W(I-1) - I^{2/p} S_W(I)$$

Cluster Stability

The idea is that if the model chosen for clustering a dataset is correct, it should be stable for different samplings of the data, the procedure is to obtain different subsamples of the data and cluster them and test their stability

Using disjoint samples:

i) The dataset is divided in two disjoint samples that are clustered separately
ii) Indices can be defined to assess stability, for example using the distribution of the number of neighbours that belong to the complementary sample.Using non disjoint samples:
iii) The data set is divided in three disjoint samples(s1,s2,s3)ii)
iv) Two clustering's are obtained from s1Us3, s2Us3
v) Indices can be defined about the coincidence of the common examples in both partitions

6. Cluster visualization

The other method for measurement is to picture the data and seem for clusters

Dimensionality reduction:

i) Project the dataset to 2 or 3 dimensions
ii) The clusters in the new space could represent clusters in the original space
iii) The confidence depends on the reconstruction error of the transformed data and that the transformation maintains the relations in the original space

Distance matrix visualization:

i) The distance matrix represents the examples relationships
ii) Can be rearranged so the closer examples appear in adjacent columns
iii) Patterns in the rearranged matrix can show cluster tendency
iv) Both methodologies are computationally expensive

Distance matrix

There are several methods, the simplest one is use hierarchical cluster algorithm and rearrange matrix using in-order traversal of the tree. Results will depend on the algorithm used and the distance/ similarity function, Can be applied to quantitative and qualitative data, patterns in the distance matrix is not always guarantee of clusters in the data.

Separation and density of cluster can calculate using the internal indexes and portioning of data calculation to be used for external indices. We can get best result with single clustering method with increasing the parameters use relative quality indexes. Clustering quality assessment talk about inference of built-in data for their collection and applicability.

7. Literature survey

Many Clustering algorithms are quantifies type of data and shapes, outliers and input parameters. We are compared with the partitional algorithms, hierarchical algorithms, density based algorithms and grid based algorithms characteristics.

Here 'n' is the no. of data points and k the no. of clusters.

Characteristics of hierarchical algorithms.

Here ‘n’ is the no.of points in the dataset.
Characteristics of Density based algorithms.
Category

* Here ‘n’ is points in the dataset.
Characteristics of grid based clustering algorithm.

Category Partitional

Name of the algorithm	Data set type	Time Complexity	Shapes	Outliers, noise	Input parameters(variables)	Final Out comes
K-Mean	Numerical	$O(n)$	Non-convex	No	No.of clusters	Centres of clusters
K-mode	Catogorical	$O(n)$	Non-convex	No	No.of clusters	Modes
PAM	Numerical	$O(k(n-k)^2)$	Non-convex	No	No.of clusters	Medoids
CLARA	Numerical	$O(k(40+k)^2+k(n-k)$	Non-convex	No	No.of clusters	Medoids
CLARANS	Numerical	$O(kn^2)$	Non-convex	No	No. of clusters, no of neighbours	Medoids
FCM	Numerical	$O(n)$	Non-convex	No	No.of clusters	Cluster Centres

Category hierarchical

Name of the algorithm	Data set type	Time Complexity	Shapes	Outliers, noise	Input parameters(variables)	Final Out comes
BIRCH	Numerical	$O(n)$	Non-convex	Yes	Cluster radius ,branching factor (BF)	CF= LS(Linear sum of the points in the cluster)+ SS (the square of N data points)
CURE	Numerical	$O(n^2 \log n)$	Arbitary	Yes	Number of clusters, number of cluster representatives	data values assignment to clusters
ROCK	Catogorical	$O(n^2+nm_m m_a+n^2 logn)$,$O(n^2,nm_m m_a)$	Arbitary	Yes	No. of clusters	data values assignment to clusters

Category Density-based

Name of the algorithm	Data set type	Time Complexity	Shapes	Outliers, Noise	Input parameters(variables)	Final Out comes
DBSCAN	Numerical	$O(nlogn)$	Arbitary	Yes	Radius of cluster , no. of objects	data values assignment to clusters
DENCLUE	Numerical	$O(nlogn)$	Arbitary	Yes	Radius of cluster , no. of objects	data values assignment to clusters

Category Grid-based

Name of the algorithm	Data set type	Time Complexity	Shapes	Outliers, noise	Input parameters(variables)	Final Out comes
Wave Cluster	Special data	$O(nlogn)$	Arbitary	Yes	Wavelets, the number of grid cells, no. of wavelet transforms	Clustered objects
STING	Special data	$O(K)$	Arbitary	Yes	objects in a cell	Clustered objects

* Here ‘n’ is points in the dataset.
Non-linear clustering algorithms were support the CURE, ROCK, DBSCAN, DENCLUE, Wave Cluster and STING. These are containing convex shapes.

8. Problems in analyzing high-dimensional data

Sparsity can leads and increase the dimensionality of the data it containing the volume [14]. Handling of data is sparse handling is difficult, estimating based on depending density .The big datasets to get the noise to fairly accurate the number of dimensions. Large data sets do not overcome these problems entirely.

Data to be in high dimensional there is a challenge for learning of various pattern-recognition methods including comparison methods of search [15], linear and non-linear classification mechanisms [16], kernel methods [17], privacy-preserving data pre-processing methods [18], artificial neural networks and clustering methods [19]. In another view of data to be in time-series the domain data is high-dimensional, so inserted data not to involve a high inherent dimensionality to the correspondence of high-dimensions. Data to be in low dimensional machine learning approaches can lead, in high dimensional projections improve the performance, various methods can be used. The inherited dimensionality of data can be estimating in different ways [20], [21]. Distance Concentration and hubness these two methods evaluate the high-dimensional data.

Distance Concentration:

In distance concentration variance is constant, up to expected value for distance should increase for grouping the data. K-nearest neighbour method is used for distance concentration. Nearest and farthest neighbours difference is can come out to disappear in inherent data in high-dimensional , impression of adjacent neighbours had questioned for many dimensions [22] , [23].They Elaborating the precise decisions for steadiness between distance utility of data is in high-dimensional to be inconsequential [24]. Redesigning Metrics can improve scrutiny of high-dimensional data. The secondary distances results in local scaling, global scaling, practical applications and shared neighbour distances [25].

Hubness:

In clustering hubness plays a vital role. K-nearest neighbour graph in hubs increasing the dimensionality, the distribution of occurrences becomes skewed. Different domains of text, images, time series and audio signals shows the data in Hubness practically. In data analysis standard types of machine learning methods can interfere [26].

Data reduction, classification, ranking, representation learning [27], metric learning [28], outlier detection, and clustering [29] methods were proposed for hubness.Points in explicit states of records space is supplementary hubs than others are familiar with distance measures it transmit to distance concentration. The consciousness causes points to lie approximately on hyper-spheres something like cluster means.

Some points are closer to the average of other points based on the variance, so the variance should be non-negligible.

Medoids are highly informative for cluster hubs in high dimensions. Medoids and hubs are interacting with the same points. In high dimensional clustering hubness plays major advantage. Many Tools and clustering indexes analyze the behaviour of data.

High-Dimensional Clustering Techniques

High-dimensional clustering proposed various techniques.

In sub-space clustering low-variance used for growing the weights is frequent mechanism. Many limitations for subspace clustering of images, data streams and text documents. Various mixture approaches be used in sub space clustering, they are k-means and its annexes, decision-trees and density-based techniques. Alternatives for sub space clustering include spectral clustering, relevant set correlation, cluster ensembles, shared neighbor methods and expectation maximization (EM). Hubness-based clustering, currently proposed in high- dimensional problems. Document clustering has been applied for high-dimensional clustering. Integral part of clustering methods is data pre- processing and feature selection.

Clustering Quality Indexes:

Evaluation of cluster is tricky job and numerous approaches projected over the years. Compactness and separation between clusters are proposed.

Quality Clustering Indexes: Existing Surveys

Some revises states many clustering quality indexes. Dissimilar revises focusing on diverse characteristics of the clustering difficulty and dissimilar properties of good cluster quality process are accepted to enclose. Use the statistics plays dissimilar approaches for moving the problem solving.

Clustering Evaluation in Many Dimensions

Relative quality indexes are used to optimize or to compare different configured clusters in representation of same data set, it was chooses by the distance measure. Metric Selections are well, to perform significant comparisons across different representations of the data sets. Sometimes clustering indexes are non-trivial. A different combination of indexes are required for each data set, it is vastly impractical. If possible, use some indexes for cross-dimensional assessment and evaluations.

Stability in quality assessment

Dimensionality increases performance of clustering indexes of their constancy and variance are connected. In similar manner regular index and dissimilar indexes presentation is prejudiced in numerous methods. For instance overlapping of clusters, small amount of clusters dimensionality increased variance is also augmented for dissimilar constructions. We want meaningful results clustering evaluation stability is important, especially number of data sample are low.

Quantifying the Influence of Hubs

Clustering indexes can be partly confined and measured data to be in Hubness. Hub points are compulsory in order to get better clustering quality. Silhouette index hubness, hubs provide additional quality to the concluding Silhouette index approximation [29]. A quantity of the indexes explains need of hub points. Some of them are demonstrate in overturn.

9. Perspectives and future directions

For a given task strong and efficient clustering plays a major task in choosing grouping methods and understand the behaviour of quality measures in high data dimensionality. The mean and indexes affects the dimensionality of the data and constancy of quality assessment. Indexes behave differently the presence or absence of overlapping clusters. Comparison of different clustering algorithms can give better results. Quality indexes are must for high-dimensional clustering they are exact one to zero.

References

[1] Rousseeuw, P.J., Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J.Comput.Appl.Math. 20.53-65(1987).

[2] Vendramin, L., Campello, R.J.G.B., Hruschka. Relative clustering validity criteria: a comparative overview. Stat.Anal.DatMin.3(4),209-235(2010).

[3] Dunn, J.C.: Well-seperated clusters and optimal fuzzy operations. J.Cybern. $(1), 95-104(1974).

[4] Davies, D.L., Bouldin, D.W.: A cluster separation measure, IEEE Trans. Patteren Anal, Mach.Intell. 1(2), 224-227(1979).

[5] Pauwels,E.J.,Frederex, G.: Cluster-based segmentation of natural scenes, In: Proceedings of the 7th IEEE International Conference on Computer Vision(ICCV), Vol.2, pp.997-1002(1999)Hubert, L.J.,Levin,J.R.: A general statistical framework for assessing categorical clustering in free recall.Psycol.Bull.83(6),1072(1976).

[6] Fowlkes, E.B., Mallows, C.L.: A method for comparing two hierarchical clusterings. J.Am/Sta.Data Eng.19 (7), 553-569 (1983).

[7] Ratkowsky, D., Lance, G.: A criterion for determining the number of groups in a classification. Aust.Comt.J.10 (3), 115-117 (1978).

[8] Calinski, T., Harabasz, J.: Adendrite method for cluster analysis. Commun.Stat. Simul.Comput. 3(1), 1-27 (1974).

[9] Bougulia, N., Almakadmeh. Boutemedjet, S.: A finite mixture model for simultaneous high-dimensional clustering localized feature selection and outlier rejection. Expert Syst. Appl. 39(7) 6641-6656 (2012).

[10] Baker, F.B., Hubert, and L.J.: Measuring the power of hierarchical cluster analysis. J. Am.Stat.Assoc. 70(349), 31-38(1975).

[11] Milligan, G.W.: A Monte carlo study of thirty internal criterion measures for cluster analysis Psychometrika 46(2), 187-199(1981).

[12] Hubert, L., Arabie, and P.: Comparing patitions. J. Classif. 2(1), 193-218 (1985).

[13] Santos, J.M., Embrechts, and M.: On the use of the adjusted rand index as a metric for evaluating supervised classification. In: Proceedings of the 19th International Conference on Artificaial Neural Networks (ICANN), Part II.Lecture Notes in Computer Science,Vol.5769, pp.175-184. Springer, Berlin (2009).

[14] .Bellman, R.E: Adaptive Control Process-A Guided Tour. Princeton University Press, Priceton (1961).

[15] Chavez.E, Navarro G: Probabilistic Spell for Curse of dimensionality in metric spaces. Inf.Process Lett. 85 (1), 39-46(2003).

[16] Serpent, G., Pathical, S: Classification in high dimensional feature spaces: Random sub sample ensemble.In: Proceedings of the International Conference on machine Learning and Applications (ICMLA), pp.740-745(2009).

[17] Evangelista, P.F., Embrechts, M.J., Szymanski. Taming the curse of dimensionality in kernels and novelty detection.In: Applied Soft Computing Technologies: The challenge of complexity, pp.425-438, springer, Berlin (2006).

[18] Aggarwal, CC.: On randomization, public information and the curse of dimensionality. In. Proceedings of the 23rd IEEE International Conference on Data Engineering (ICDE) pp.136-145(2007).

[19] Randovanvic .M: Representations and Metrics in High Dimensional Data mining. Serbia (2011).

[20] Cater, raich, Hero: On local intrinsic dimension estimation and its applications. IEEE trans, Signal Process, 58 (2), 650-663 (2010).

[21] Gupta. M.D, Huang, T.S: Regularized maximum likelihood for intrinsic dimension estimation. Compt,Res.pep. (2012).

[22] Durrant, R, J, Kaban When is ‘nearest neighbour’ meaningful: a converse thermo and implications. J. Complex. 25(4), 385-397(2009).

[23] Zimek, Schubert,E, Kriegel, H.P: A survey on supervised outlier detection in high dimensional numerical data. Stat, Anal, Data Mining.5 (5), 363-387(2012).

[24] Kabana, Non-parametric detection of meaningless distances in high dimensional data. Stat.comput. 22(2), 375-385(2012).

[25] Yin, J., Fan, X., Chen, Y., Ren, and J Highdimensional shared neares neighbor clustering algorithm. In: Fuzzy Systems and Knowledge Discivery, Lecture Notes in Computer Science, vol.3614, pp. 484-484,Springer, Berlin(2005).

[26] Randovanovic,M: Nanopoulos,A., Ivanovic, M.: Nearest neighbours in high dimensional data: The emergence and influence of hubs. In: Proceedings of the 26th International Conference on Machine Learning (ICML), pp.865-872(2009).

[27] Tomasev, N., Rupnik,J., Mladenic, D., The role of hubs in cross-lingual supervised document retrieval . In: Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), pp.185-196. Springer, Berlin (2013).

[28] Tomasev, N., Mladenic, D.: Hubness aware shared neighbor distances for high dimensional k-nearest neighbor classification, Knowl,Ing,Syst. 39(1), 89-122(2013).

[29] Tomasev, N., Randavonic, M., Mladenic, D., Ivanovic, M., The role pf hubness in clustering high dimensional data IEEE Trans, Know Data Eng. 26(3), 739-751(2014).