CENG574 Final Report

Ardan, AY, Yılmaz

Department of Computer Engineering, METU, yilmazardan@gmail.com

Baran, BG, Gülmez

Department of Computer Engineering, METU, baran.gulmez07@gmail.com

CCS CONCEPTS • Data Projection• Clustering •

Additional Keywords and Phrases: DB-SCAN, K-Means Clustering, Hierarchical Clustering, PCA, IsoMAP

1 Dataset Description

Electrical power plants need a fault detection system in the minimum possible time for the sake of both equipment protection and stability. For this, a power system is designed and tested for both normal and faulty conditions using MATLAB whose results have been recorded to a dataset.

The following figure shows the simulated system, where three current and voltage values have been measured, namely, Ia, Ib, Ic, Va, Vb, Vc.

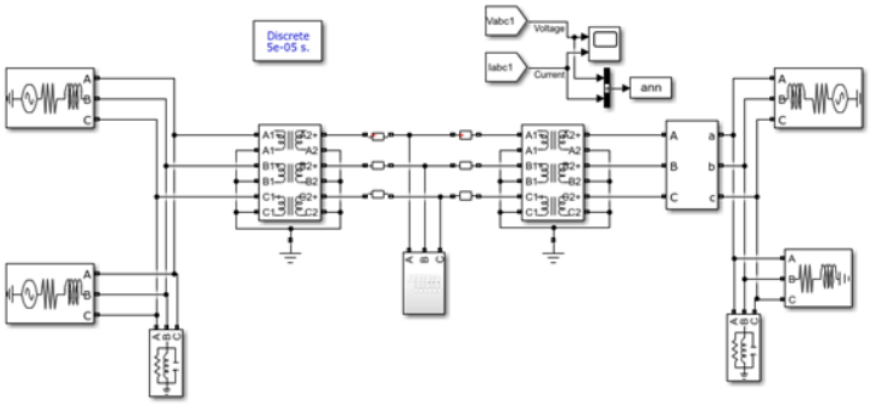

Figure 1: Electrical Fault detection and classification, 2022 [1]

1.1 Problem Definition

The following table shows the simulation instances where voltage and current values for sources A, B, C and the output which indicates the faultiness of the system. As can be inferred, this is a binary classification/clustering problem, that is, the system is either faulty (indicated by output = 0) or not.

	Output..S.	Ia	Ib	Ic	Va	Vb	Vc
1	0	-170.47220	9.2196135	161.25258	0.0544900	-0.6599209	0.6054309
2	0	-122.23575	6.1686674	116.06709	0.1020000	-0.6286115	0.5262016
3	0	-90.16147	3.8136322	86.34784	0.1410255	-0.6052769	0.4642513
4	0	-79.90492	2.3988035	77.50611	0.1562725	-0.6022353	0.4459629
5	0	-63.88525	0.5906674	63.29459	0.1804515	-0.5915014	0.4110499
6	0	-55.95468	-1.0018817	56.95656	0.1934141	-0.5906954	0.3972813

Figure 2: The input and outputs of the system

2 Preliminary Analysis

For preliminary analysis, first, the class imbalance is checked. And the number of faulty instances is 6505, whereas that of non-faulty ones is 5496; hence this dataset can safely be considered as balanced.

The following plots show the effect of change in the current and voltage measures on the output.

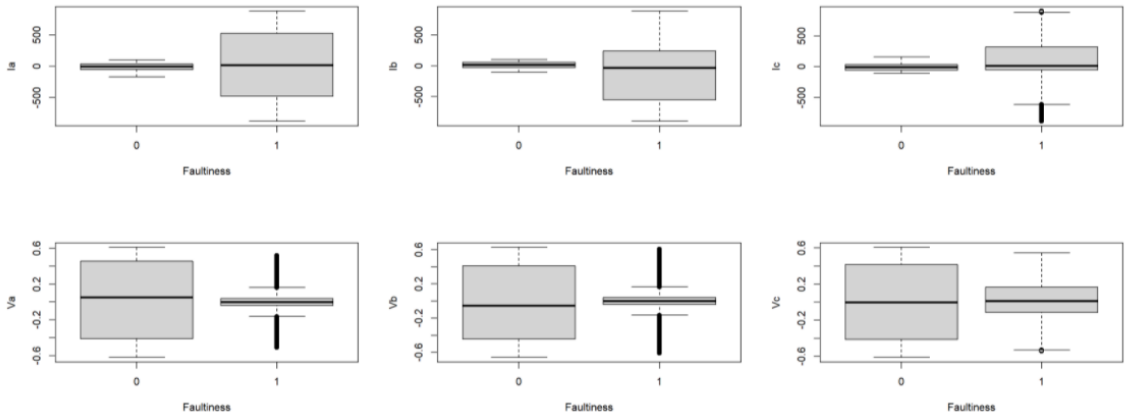

Figure 3: Effect of input change on the output.

As can be inferred from the preceding graphs, the range of current values not to produce fault in the system has a narrower range than that of those to cause fault. That's why one can deduce that the current values outside this range will potentially cause fault. Further, the variance on the voltage measures to produce fault is greater, hence their contribution to the principal components are expected to be greater.

2.1 Principal Component Analysis (PCA)

Using a linear combination of inputs, PCA finds principal components to carry much of the information so that the data can be represented in fewer dimensions.

The following figure shows the proportion of variance each principal component carries.

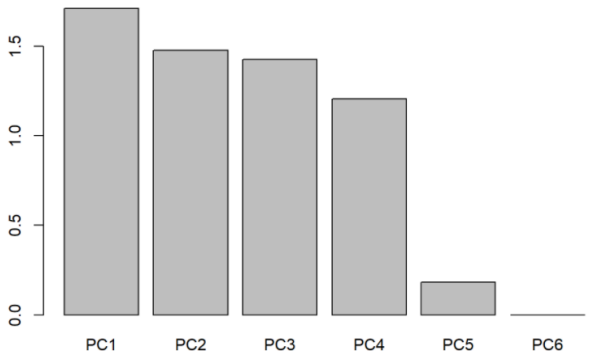

Figure 4: Proportion of variance each principal component carries

As can be seen from the graph, the first four principal components are the ones to carry much of the information, hence appropriate to represent the data. However, for visualization, only the projection of the first two and three principal components are given below.

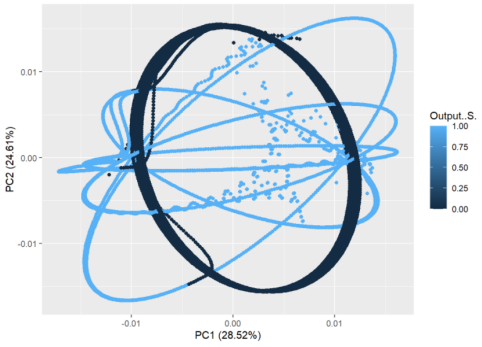

Figure 5: PCA using the first 2 PCs. Dark points are non-faulty.

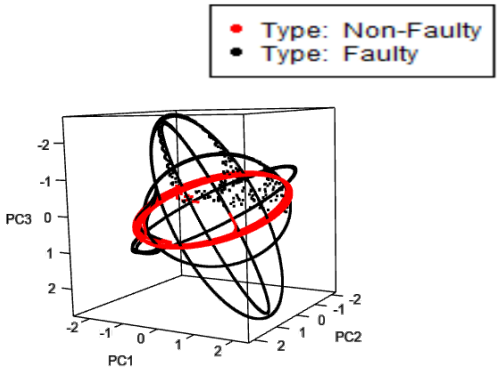

Figure 6: PCA using the first 3 PCs

The resulting graphs are consistent, ie, the non faulty data are clustered much more densely in the middle, which is also coherent with the preliminary analysis comments.

2.2 Projection of Data

2.2.1 Linear Projection

Three different methods of MDS are used, namely, ratio, interval, and monotone split MDS, which are all to minimize the Kruskal's stress function transforming the dissimilarity matrices in a different manner. [2]

Results of the MDS methods are almost the same, hence only one is provided below, which is the result of the Ratio MDS.

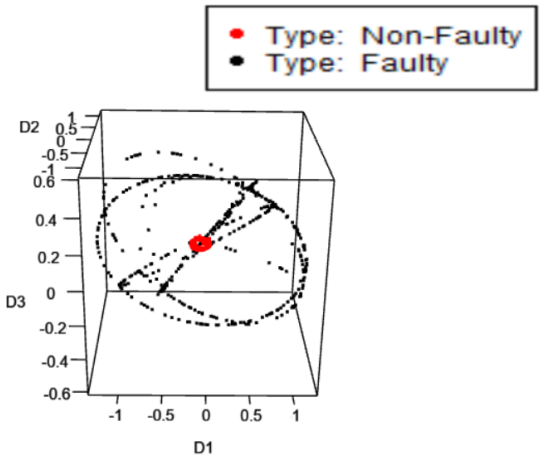

Figure 7: Resulting scaling of Ratio MDS.

The above graph is also coherent with the PCA results and preliminary analysis in that the non-faulty cluster is in the middle.

2.2.2 Non-Linear Projection

For nonlinear mapping, ISOMAP is used, as it is known to perform well on data that is distributed on a spherical manifold as seen with the previous graphs.

2.2.2.1 Hyperparameter Optimization

ISOMAP preserves the distances based on the neighborhood of a data point, which is either defined by the number of neighbors or a distance metric. For this, the number of neighbors is chosen, as it is a hyperparameter that can be optimized more easily. To optimize the k, k-NN algorithm is applied and the resulting accuracy values produced on different k, number of clusters, are plotted on the following graph. As

can be seen from the following graph, small k values yield almost perfect accuracy, however, really small k values are known to be more susceptible to outliers. That's why k=10,25,50 has been tried, but the resulting projection did not change much, so k=10 is chosen for the sake of computational complexity.

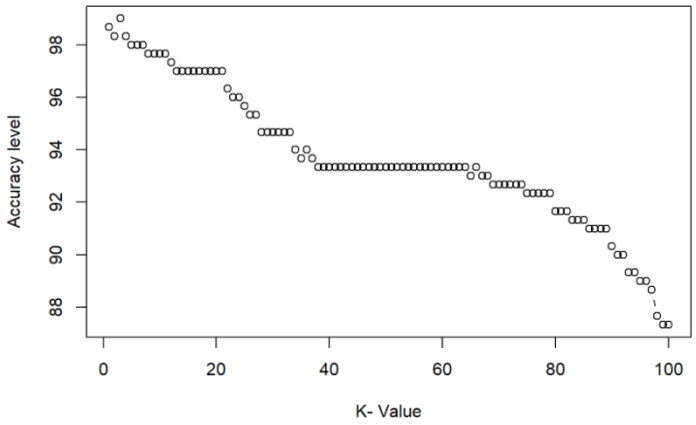

Figure 8: Yielded accuracy on different k values for k-NN

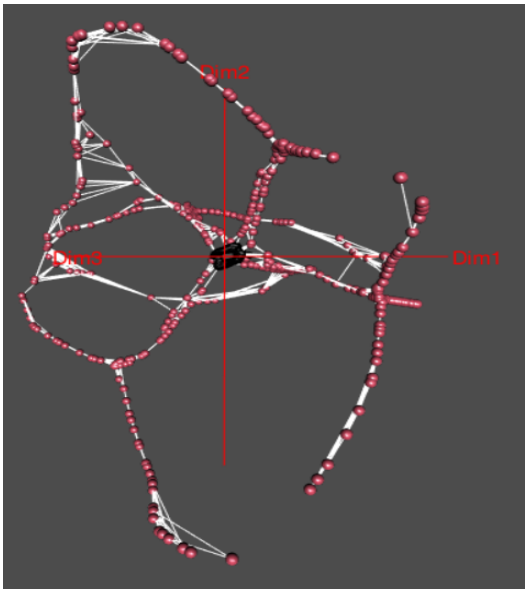

Figure 9: Result of ISOMAP with k=10

3 Clustering of Data

3.1 Hierarchical

3.1.1 Agglomerative Hierarchical Clustering (HAC)

Four different methods of HAC have been applied, namely, max-link, min-link, mean-link, and ward-link method.
The level to cut the dendrogram from has been chosen to be two, as it is a binary classification problem. One can see the true class labels at the bottom of each dendrogram, which should give an insight about the cluster validity.

The following four figures show the resulting dendrograms with the above-mentioned methods.

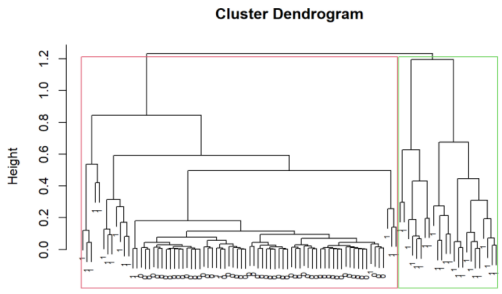

Figure 10: Cluster dendrogram with Max-Link Method

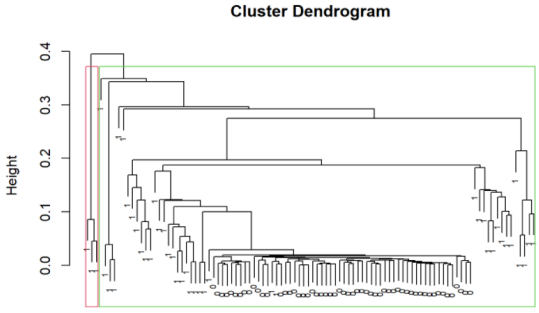

Figure 11: Cluster dendrogram with Min-Link Method

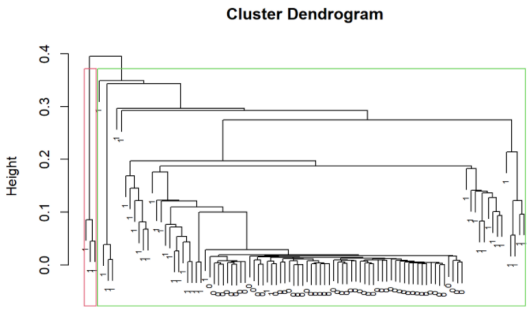

Figure 12: Cluster dendrogram with Mean-Link Method

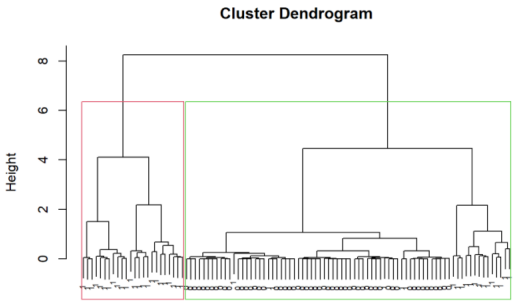

Figure 13: Cluster dendrogram with Ward Method

One can gain insight into the validity of the formed clusters by analyzing the above graphs, ie, examining how mixed the labels are within a formed cluster. As a result, Max-Link and Ward methods are the ones with the acceptable results whose formed clusters are presented below.

Figure 14: Resulting clusters with Max-Link HAC

Figure 15: Resulting clusters with Ward's Link HAC

3.2 K-Means

This well-known method forms clusters minimizing an objective function, sum of squared distances between each data point to the center of the cluster to which it is assigned.

3.2.1 Elbow Method

For k-means clustering, first the hyperparameter k, the number of clusters, needs to be chosen. For this, the elbow method is applied, whose graph, representing the number of clusters vs the objective function, is given below.

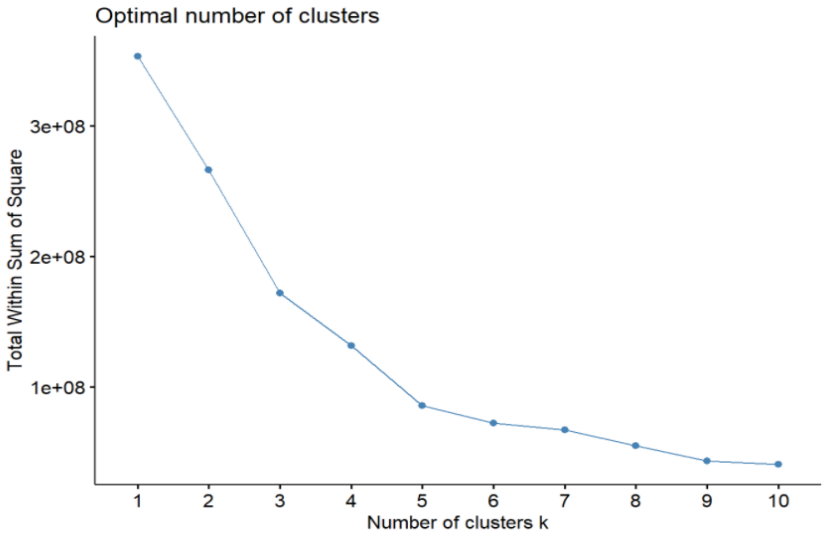

Figure 16: Change in the objective with different number clusters

k values are chosen to be 2, as this is already known to be a binary classification problem, and 3, as this is where the elbow on the graph appears to be.

3.2.2 Resulting Clusters

Figure 17: 2-means resulting clusters

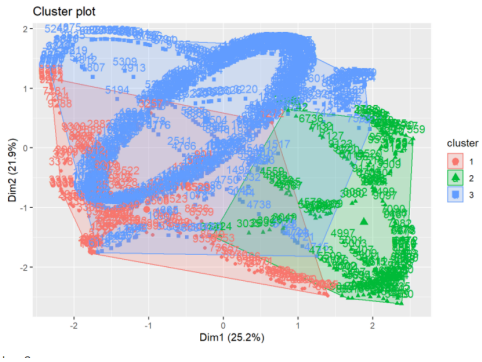

Figure 18: 3-means resulting clusters

3.3 DB-SCAN

Since the data contains ring like structures, using Spectral Clustering and DB-SCAN would be most suitable. However, DB-SCAN is much faster, thus it is preferred over Spectral Clustering. “dbscan” package is for DB-SCAN.

DB-SCAN assumes that clusters are dense regions separated from the other regions by density, thus it is a density based clustering method. DB-SCAN starts with a point and creates a circle of epsilon radius and classifies them as Core point, Border point and Noise point. If the circle around the chosen point contains at least a “minPoints” number of points, that point is classified as a Core point. If the number of points is less than “minPoints” in the circle, then it is classified as a Border point. If there are no data points in the circle, then it is classified as Noise point.

3.3.1 Hyperparameter Optimization

Parameter optimization is made on a subset with 1000 samples. R documentation[4] suggests using minPts as “equal to the dimensionality of the data plus one or higher”. There are 6 input features so 6-7 is suggested. Values between 5-10 with different “eps” values are tested and 7 is works best as suggested.

Again R documentation suggests using “kNNdistplot()” for choosing “eps” value. A sudden increase of the kNN distance suggests that the right of the sudden increase is most likely to be outliers. Inspecting the below graph the sudden increase is between 0 and 70. After searching the 0-70 range manually, the best value found for “eps” is 15.

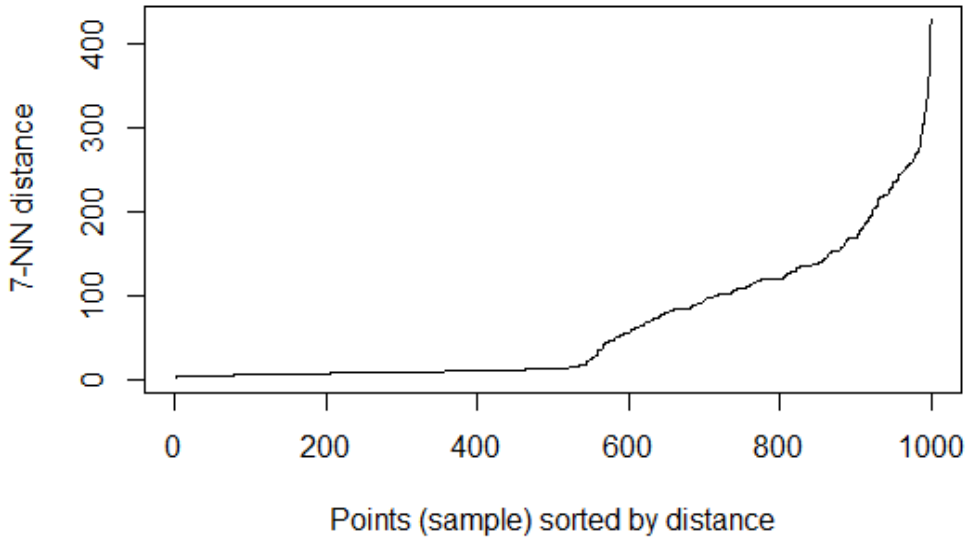

Figure 19: kNNDistPlot() for 7-NN of the data

Again after choosing the “eps” value, DB-SCAN is run on a different subset. For this, the random seed is changed. This way DB-SCAN started with different initialization each time. It is seen that it robustly gives the same consistent results on different initializations. Figure 20 illustrates DB-SCAN results are quite similar to the ground truth. This result is used for external and internal evaluations.

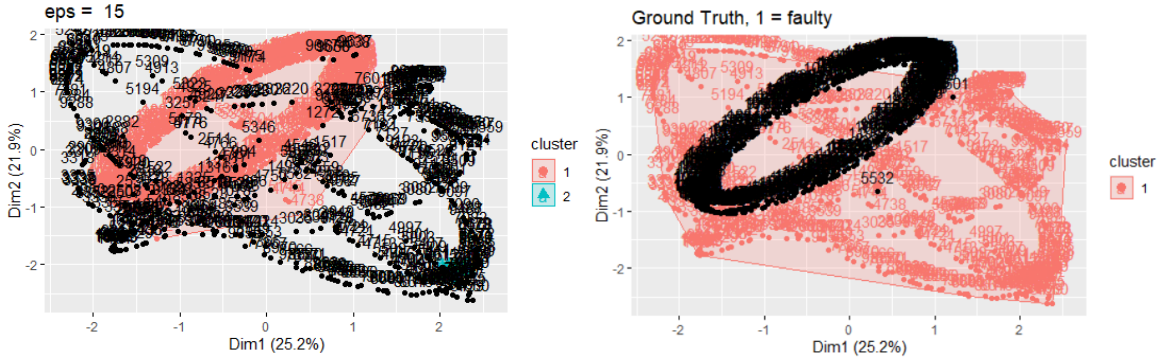

Figure 20: DB-SCAN result (left), ground truth labels(right)

However, it is worth mentioning that DB-SCAN is sensitive to parameter change. Slightly varying both “eps” or “minPts” dramatically changes results.

3.3.2 Appyling Final Parameters to All Data

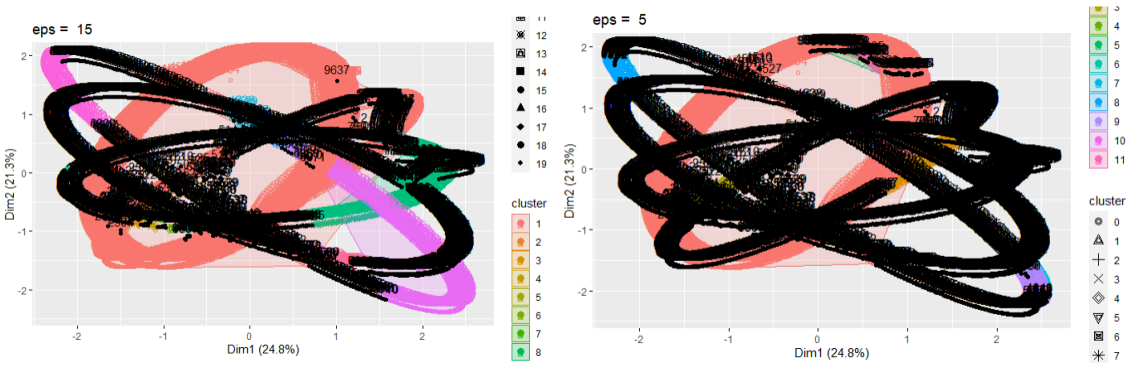

Figure 21: DB-SCAN with eps = 15, DB-SCAN with eps = 5

Applying DB-SCAN on all data eps=15 still gives quite well results, however eps=5 works better. Here, we see one of the disadvantages of DB-SCAN: the number of clusters is not predictable. However, as mentioned earlier, when there are more than 2 clusters, the most crowded cluster is taken as one cluster and all others are combined to make binary clustering. Because the ground truth to be compared has only 2 classes.

4 Evaluation of Clusters

For the evaluation of clusters, each method's best working parameter configuration is taken. External indices compare the clustering with ground truth while internal indices compare the inter and intra clusters distances.

4.1 External

"ClusterR" package is used for external evaluation. R documentation[5] explains how ARI values should be interpreted: ARI >= 0.90 excellent recovery; 0.80 =< ARI < 0.90 good recovery; 0.65 =< ARI < 0.80 moderate recovery; ARI < 0.65 poor recovery. According to this interpretation DB-SCAN has excellent recovery and the other two have poor recoveries. Ward's Link and 2-Means have excellent sensitivity, which means they find all True Positives. This is because they over pick samples as positives. This also explains very low specificity which means picking True Negatives. Inspecting these metrics, one can say that Ward's Link and 2-Means methods tag all positives and half of the negatives as positive. Matthews' Correlation Coefficient takes into account all possible metrics, namely, True Positives, True Negatives, False Positives and False Negatives. Thus, overall this metric gives a general idea by itself. As in other methods, DB-SCAN performs much better regarding MCC.

Table 1: Best Methods' External Metrics For Comparison

	Ward's Link	2-Means	DB-SCAN
Rand	0.29	0.24	0.92
Accuracy	0.77	0.74	0.98
Sensitivity	1.00	1.00	0.98
Specificity	0.500	0.44	0.97
Matthews Correlation Coefficient	0.59	0.55	0.96

External Metrics Table

4.2 Internal

"clValid" package is for internal evaluation metrics. It is worth mentioning that the "clValid()" function does not take the DB-SCAN method as one of the input methods. Thus only the "Silhouette" index is calculated for DB-SCAN separately.

Lower values of Connectivity Index indicate better clustering and it takes values between 0 and ∞. Connectivity measures how distant samples are placed in the same cluster as their nearest neighbors in the data space. Thus, connectivity is a measure of compactness of clusters.

Higher values of Dunn index indicate better clustering and it takes values between 0 and ∞.Dunn Index is the ratio of the smallest distance between samples from different clusters to the largest intra-cluster distance. Thus, the Dunn Index is a measure of separation of clusters.

Higher values of the Silhouette index indicate better clustering and it takes values between 0 and 1. While calculating a sample, other samples from both the same cluster and the different clusters are used. Thus, Silhouette Index combines both inter and intra cluster distance.

Table 2: Best Methods' Internal Metrics For Comparison

		Number of Clusters = 2	Number of Clusters = 6
hierarchical	Connectivity	13.79	29.20
	Dunn	0.14	0.18
	Silhouette	0.48	0.61
k-means	Connectivity	13.40	30.96
	Dunn	0.07	0.18
	Silhouette	0.51	0.62
DB-SCAN	Silhouette	0.88	NA

Internal Metrics Table

When Silhouette values are compared, DB-SCAN again outperforms Ward's Link and K-means clusterings. Comparisons are made for 2 clusters and 6 clusters since the dataset has 1 faulty and 5 non-faulty classes in ground truth.

For k-Means and Ward's link, 2 clusters give better compactness and worse separation of clusters than 6 clusters. Worse separation is because as the number of clusters increases their distance also increases.

5 Comparison With Other Works

Other works made on this dataset can be found on Kaggle[1]. Firstly, all other works use python. For other works that use multiple methods, their best performing method is chosen for comparison. Mostly the best result is from Decision Tree method. Second work is the best and our results which do not have learnable parameters like Neural Networks or Decision trees perform similar to theirs.

Table 3: Comparison of our work with others' from Kaggle

	Our Work (DBSCAN)	Work1 Decision Tree	Work2 Random Forest	Work3 Decision Tree	Work4 NN	Work5 Random Forest	Work6 Polynomial Regression
Accuracy	0.98	0.98	0.99	0.85	0.82	0.99	0.98
Sensitivity	0.98	0.99	0.98	NA	NA	NA	NA
Specificity	0.97	0.97	0.99	NA	NA	NA	NA
Precision	0.97	0.98	0.99	0.85	NA	0.99	NA
Recall	0.98	0.99	0.98	0.85	NA	0.99	NA

Other Works Comparison Table

6 Conclusion

The dataset contains values for current and voltage measures, and whether the system is faulty or not on these. As a result of convenient results so far, it seems that data form clusters. That is, upon the arrival of new data instances, one can infer the conditions to potentially produce fault using these analyses, and take the necessary precautions, which was the whole purpose from the beginning.

Dimensionality reduction methods give simil.ar results to each other. They are best utilized in visualization of the data. Visualizing the ring-like structure gave much intuition and helped make them work better in later steps. Without visualizing the data it would be impossible to choose the most suitable method which is DB-SCAN.

Both Hierarchical and K-means clustering methods failed to cluster the data. DB-SCAN clustered it near perfectly. So it can be said that clustering quality of the data is very dependent on the clustering method. For nested structures like this, K-means is of no use. Hierarchical clustering is known to handle these structures better, however did not work on this dataset.

For the external evaluation metrics, accuracy, sensitivity and specificity do not give useful insight about the results. They are only useful if they all are inspected together. However Adjusted Rand Index and Matthew's Correlation Coefficient are quite explanatory by themselves. For the internal evaluation metrics, Silhouette coefficient is also enough by itself unlike Connectivity or Dunn indices. Hence it can be said that combined metrics are more useful.

REFERENCES

[1] Kaggle.com. 2022. *Electrical Fault detection and classification*. [online] Available at: <https://www.kaggle.com/esathyaprakash/electrical-fault-detection-and-classification?select=detect_dataset.csv> [Accessed 22 January 2022].

[2] Cda.psych.uiuc.edu. 2022. [online] Available at: <http://cda.psych.uiuc.edu/mds_509_2013/borg_groenen/chapter_nine.pdf> [Accessed 22 January 2022].

[3] Atul Adya, Paramvir Bahl, Jitendra Padhye, Alec Wolman, and Lidong Zhou. 2004. A multi-radio unification protocol for IEEE 802.11 wireless networks. In Proceedings of the IEEE 1st International Conference on Broadnets Networks (BroadNets'04) . IEEE, Los Alamitos, CA, 210–217. https://doi.org/10.1109/BROADNETS.2004.8

[4] rdocumentation.org. *R documentation for dbscan*, [online] Available at: <https://www.rdocumentation.org/packages/dbscan/versions/1.1-8/topics/dbscan> [Accessed 22 January 2022]

[5] 574rdrr.io *R documentation for ClusterR*, [online] Available at: <https://rdrr.io/cran/CrossClustering/man/ari.html#:~:text=from%20other%20methods.-,Details,%3B%20ARI%20%3C%200.65%20poor%20recovery> [Accessed 22 January 2022]