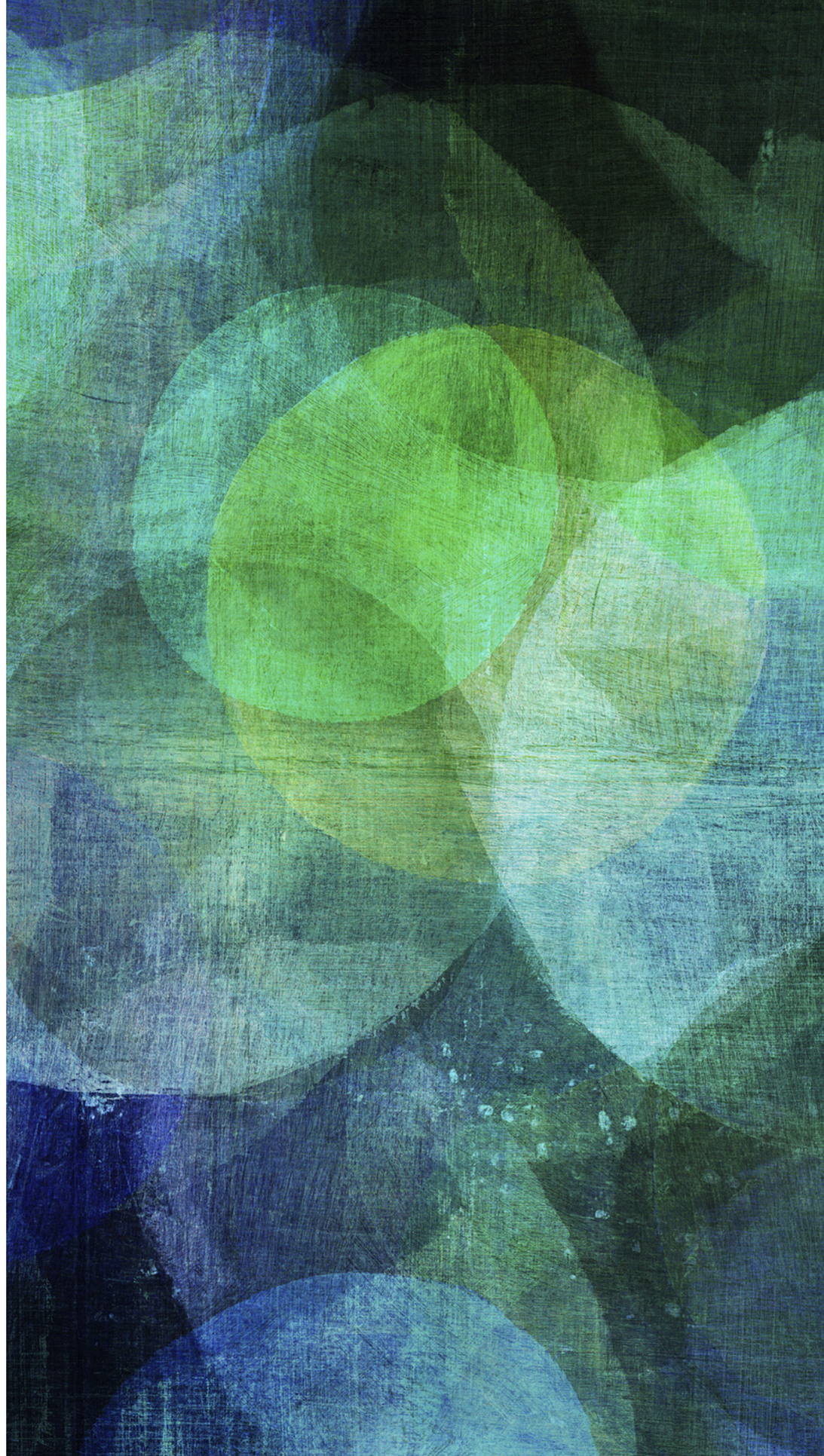
PES UNIVERSITY

A MODULAR SPATIAL CLUSTERING ALGORITHM WITH NOISE SPECIFICATION

Presented By - Akhil K



CONTRIBUTIONS

- Bacteria-Farm algorithm - a novel modular algorithm which grows outward from the centre of measure.
- Noise specification - a method to specify noise to be discarded from the data, by the algorithm.

BACTERIA-FARM ALGORITHM

- Bacteria-Farm algorithm is a spatial-data clustering algorithm which groups together points which are closely related to each other.
- The algorithm starts at a *measure of centre* and the clusters grow outward from that point.
- On reaching a threshold (maximum number of points the cluster can contain), the algorithm terminates.

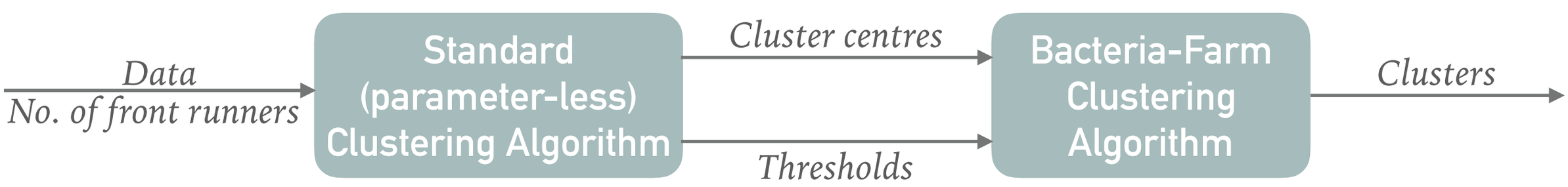


Figure 1: Architecture of Bacteria-Farm model

ILLUSTRATION OF WORKING OF THE BACTERIA-FARM ALGORITHM

The parameters passed to the algorithm are :

- Threshold - maximum number of points a cluster can contain.
- Centroid - a measure of centre
- Number of Front Runners - Front Runners are defined as the points on the periphery of the the cluster. Hence, “front runners”.

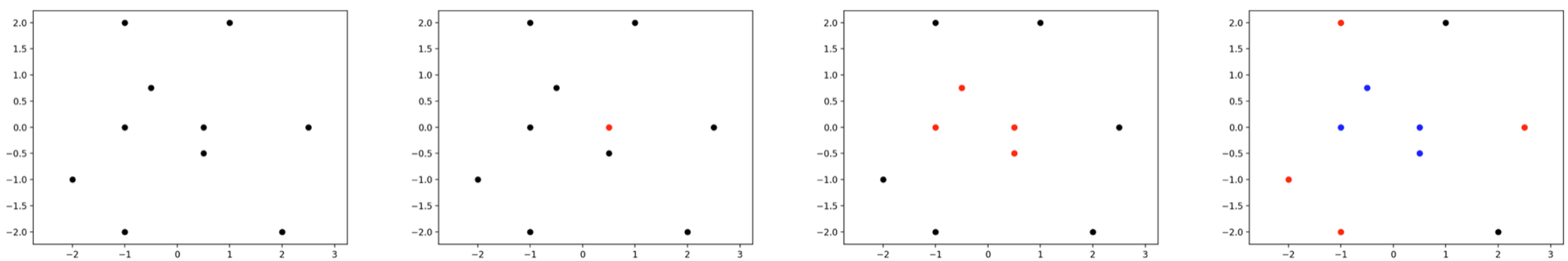


Figure 2: Illustration of working of the Bacteria-Farm algorithm

NOISE SPECIFICATION

- Since we can alter the threshold fed to the Bacteria-Farm algorithm, an interesting property is observed. As the clusters grow outward, varying the threshold essentially varies the noise specified to the algorithm.
- With this property, we are free to choose the amount of noise to be discarded from the data by the algorithm.

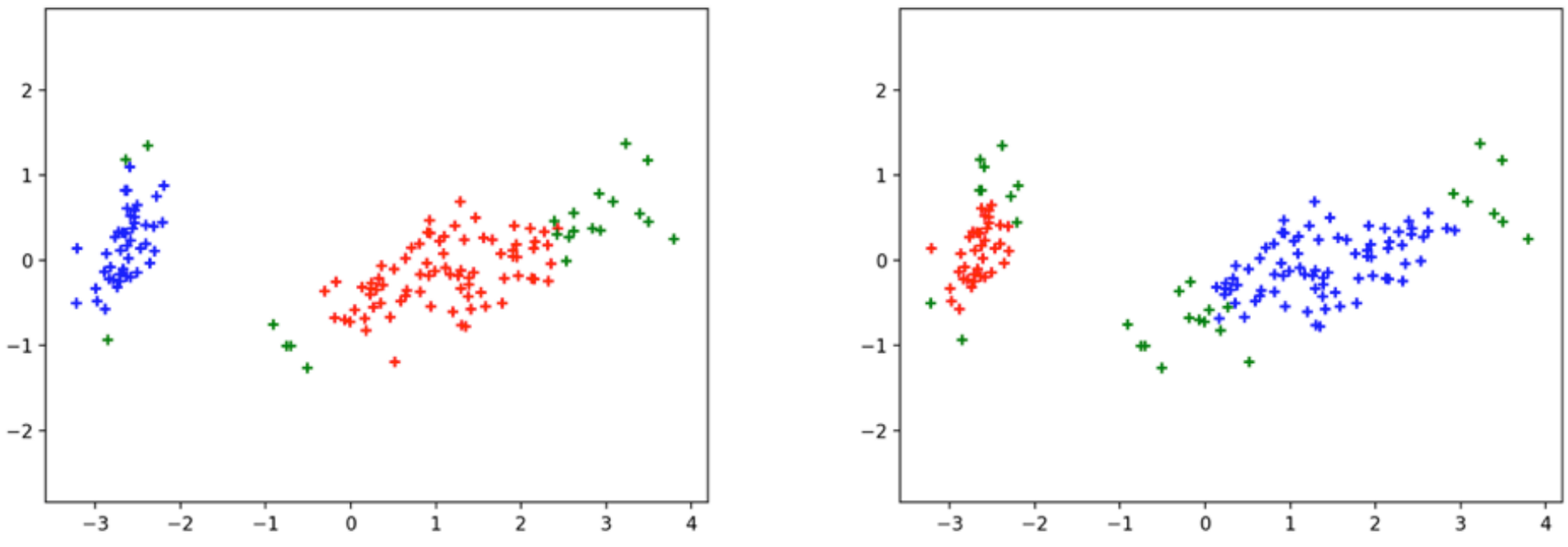


Figure 3: Varying noise specifications (15 percent and 20 percent) on the popular Iris dataset

ROBUSTNESS IN THE MODEL

- Many current clustering algorithms' outputs vary a lot when the input parameters are changed slightly i.e., they aren't robust.
- Our model is designed to be robust as most of its parameters are generated by the first part of the model. The diagram below illustrates the robustness of the model's only input parameter - Front Runners.

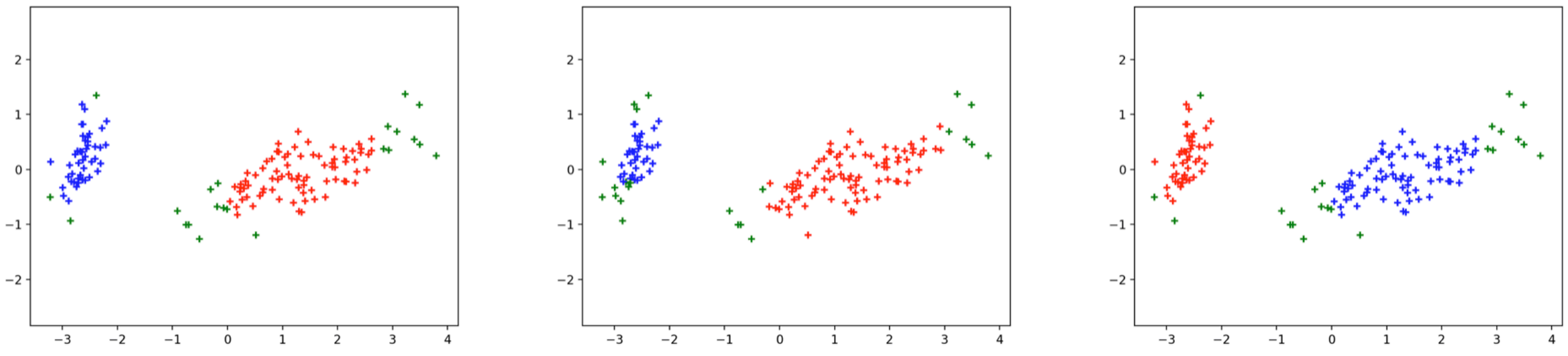


Figure 4: Clustering on Iris data with varying numbers of front runners (3,5,7 respectively)

CLUSTERING DATA OF ARBITRARY SHAPE

- Since we don't have elements such as 'radius' in the model, clustering isn't restricted to data with a standard shape.
- In Bacteria-Farm, the algorithm is only driven by Euclidean distance between points and this enables clustering data of any shape.

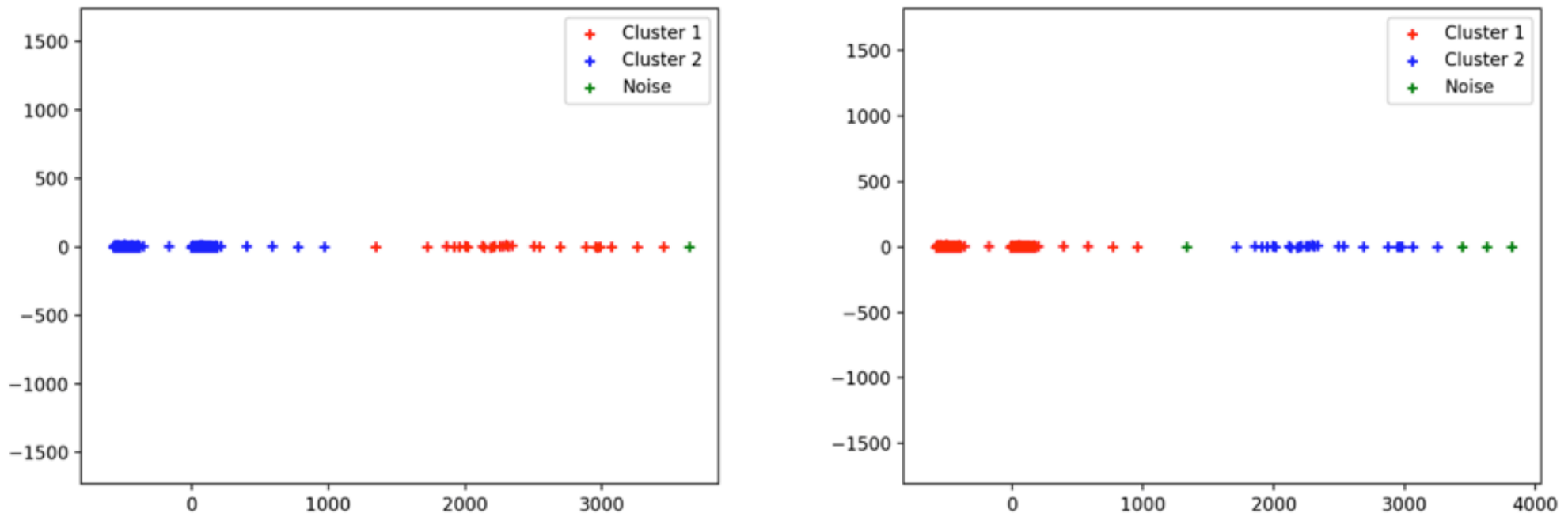


Figure 5: Visual comparison of performance between DBSCAN (left) and Bacteria-Farm (right) for a dataset of arbitrary shape

RESULTS

To evaluate the results, we have compared Bacteria-Farm model with DBSCAN and K-Means with respect to performance metrics - *Silhouette Coefficient* and *Calinski-Harabasz Index*.

Algorithm	**Silhouette Coefficient**	**Calinski-Harabasz Index**
K-Means	0.5842	14852.1314
DBSCAN	0.6103	16657.9614
Bacteria-Farm	0.6167	1483.1049

Table 1: Comparison of mean values of performance metrics over 91 real datasets, for the three clustering algorithms.