Roll No.	Experiment No. 09	Marks :
BATCH -		Sign :

Aim: Demonstrate Clustering algorithm on data sets using data mining tools (WEKA, R tool, XL Miner, Orange etc.)

Apparatus: WEKA

Theory:

What are Clustering Algorithms in Data Mining?

Clustering algorithms in data mining are an unsupervised Machine Learning Algorithm that comprises a set of data points in clusters so that the objects belong to precisely the same group. Clustering algorithms in data mining will help to split data into several subsets. Each subset has data like one another. Those subsets are called clusters; given that our client database is split into clusters, we can make a decision about just who we believe is most suitable for that item.

Clustering algorithms in Data Mining helps in the identification of aspects. That is similar to land use on an earth observation site. It can additionally aid in title groups of houses of a city. It is based on home geography, worth, and style location. Clustering algorithms in Data Mining likewise help classify documents on the internet for data discovery. Also, we use Data clustering within outlier detection apps. For example, detection of card frauds. As a data mining feature, cluster analysis serves as a tool. That is gaining insight into the distribution of data.

Clustering algorithms in data mining is a method helpful for exploring data. It's constructive when there are many causes and no clear all-natural groupings. At this point, clustering data mining algorithms can be used to locate whatever organic collections might exist.

Types of Clustering Algorithms in Data Mining

1. **Centroid-based Clustering Algorithms in Data Mining:**

   Centroid-based clustering algorithms in data mining organize the data into non-hierarchical clusters, in contrast to hierarchical clustering algorithms defined below. K means regarded as the widely used centroid-based clustering algorithm. Centroid-based algorithms are practical but delicate to first factors & outliers. This program concentrates on k means since it's a practical, effective, and straightforward clustering algorithm.

2. **Density-based Clustering Algorithms in Data Mining:**

   Density-based clustering algorithms in data mining link areas of high illustration density into clusters. It allows for shaped distributions so long as dense regions can be attached . These algorithms have difficulties with data of different thicknesses and dimensions. Additionally, by design, these algorithms don't assign outliers to clusters.

3. **Distribution-based Clustering Algorithms in Data Mining:**

   Distribution-based algorithms in data mining strategy, this particular clustering algorithms assume the data of distributions, like Gaussian distributions. The distribution-based algorithm clusters data into three Gaussian distributions. As the distance from the distribution's middle advances, the probability that an area belongs to the division decreases. The bands show which reduction in prospect. If you don't understand the distribution type in your data, you need to use a unique algorithm.

4. **Hierarchical Clustering Algorithms in Data Mining:**

   Hierarchical clustering algorithms in data mining produce a tree of clusters. Hierarchical clustering, not surprisingly, is ideally suited to hierarchical details, like taxonomies. See Comparison of sixty-one Sequenced Escherichia coli Genomes by Oksana Lukjancenko, Trudy Wassenaar & Dave Ussery for a good example. An additional advantage is that just about any cluster could be picked by cutting the tree to the proper degree.

DEMONSTRATION (Hierarchical Clustering)

Generated Dataset:

Viewer

Relation: weka.datagenerators.clusterers.BIRCHCluster-S_1_-a_10_-k_4_-N_1..50_-R_0.1..1.41_-O

No.	1: X0 Numeric	2: X1 Numeric	3: X2 Numeric	4: X3 Numeric	5: X4 Numeric	6: X5 Numeric	7: X6 Numeric	8: X7 Numeric	9: X8 Numeric	10: **X9** Numeric
1	-0.171...	1.0378...	2.9991...	-0.273...	3.8551...	3.7998...	3.8950...	4.0822...	2.7569...	0.0862...
2	0.5345...	0.5324...	3.6527...	-0.221...	4.6790...	4.6934...	3.9745...	4.3601...	1.8819...	0.9547...
3	0.5046...	1.7008...	4.0605...	0.2344...	3.3844...	3.9514...	3.6765...	3.8944...	2.0200...	1.8553...
4	0.8336...	1.4759...	3.5188...	0.2383...	3.9035...	3.7028...	4.8577...	3.5825...	1.2203...	1.4978...
5	0.6551...	1.3649...	4.2998...	-0.499...	3.9990...	4.2444...	3.0261...	4.2628...	1.4195...	1.6963...
6	1.5341...	1.1098...	3.5235...	-0.615...	4.0044...	4.0112...	3.5238...	3.7237...	1.2913...	1.4486...
7	1.1845...	2.1825...	4.0266...	-0.072...	3.3990...	4.1342...	3.0585...	4.2313...	1.2319...	0.4371...
8	0.7649...	0.7426...	4.3073...	0.5585...	3.9064...	3.8504...	4.2847...	4.5428...	1.4273...	1.5038...
9	0.6685...	1.6500...	3.9576...	0.3093...	3.5215...	4.0392...	3.5468...	4.3196...	1.7374...	1.1175...
10	0.2731...	1.8389...	3.2376...	-0.831...	4.4581...	3.4260...	4.2921...	3.4098...	2.3885...	0.0490...
11	0.5902...	2.1606...	4.1989...	0.4405...	3.5339...	4.1855...	3.0894...	3.6013...	1.3645...	1.2291...
12	1.6757...	1.7218...	4.0323...	-0.695...	4.0307...	3.8531...	4.1371...	3.4242...	1.0138...	1.2922...
13	1.3551...	1.4929...	4.3141...	-0.607...	3.5141...	3.3644...	3.1316...	3.3190...	1.7557...	1.5336...
14	0.5185...	0.6719...	3.8110...	0.0462...	3.5410...	3.8399...	4.2226...	3.2851...	1.2733...	1.5470...
15	0.5293...	0.7879...	3.3443...	0.1744...	4.4766...	3.7894...	3.7191...	3.5423...	2.0121...	2.1139...
16	1.3585...	0.5102...	4.0238...	0.0883...	3.1840...	3.6245...	4.2887...	3.1691...	1.7471...	1.2172...
17	0.6067...	0.4637...	5.0956...	0.5730...	3.3689...	4.2753...	3.7326...	4.3454...	1.6942...	1.3563...
18	0.8779...	1.2669...	4.1996...	0.2798...	3.7374...	3.3762...	4.0847...	3.4243...	2.1856...	1.7196...
19	0.6197...	1.5779...	3.2796...	0.3647...	4.1412...	4.1587...	3.7416...	3.2838...	1.6219...	2.1544...
20	0.5323...	1.3711...	3.7355...	0.4168...	3.3090...	4.0456...	3.6330...	4.0382...	1.7704...	2.2174...
21	0.3494...	1.0990...	4.2898...	0.3180...	3.7844...	4.3886...	3.4088...	3.7297...	0.3235...	0.8789...
22	1.3412...	1.0743...	3.7841...	-0.447...	4.7843...	3.8564...	3.8373...	4.1206...	1.4030...	0.8818...
23	0.3314...	1.3014...	4.3188...	-0.080...	4.0754...	4.7630...	3.4946...	3.5281...	1.9541...	0.7048...
24	0.7927...	1.3556...	3.1251...	0.5712...	3.2207...	4.3885...	4.1476...	3.8953...	1.6754...	1.1283...

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Generate... | Undo | Edit... | Save...

Filter
Choose None — Apply | Stop

Current relation
Relation: weka.datagenerators.clusterers.BIRCHCluster-S_1_-a_10_-k_4_-N_1..50_-R_0.1..1.41_-O
Instances: 136 — Attributes: 10 — Sum of weights: 136

Attributes
All | None | Invert | Pattern

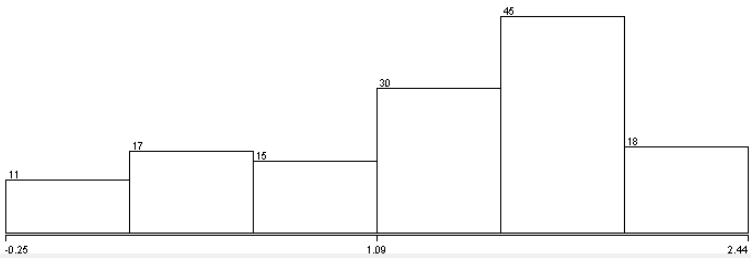
No.	Name
1	X0
2	X1
3	X2
4	X3
5	X4
6	X5
7	X6
8	X7
9	X8
10	X9

Remove

Selected attribute
Name: X0 — Type: Numeric
Missing: 0 (0%) — Distinct: 136 — Unique: 136 (100%)

Statistic	Value
Minimum	-0.255
Maximum	2.436
Mean	1.316
StdDev	0.678

Class: X9 (Num) — Visualize All



Status
OK — Log — x 0

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Clusterer
Choose HierarchicalClusterer -N 2 -L SINGLE -P -A "weka.core.EuclideanDistance -R first-last"

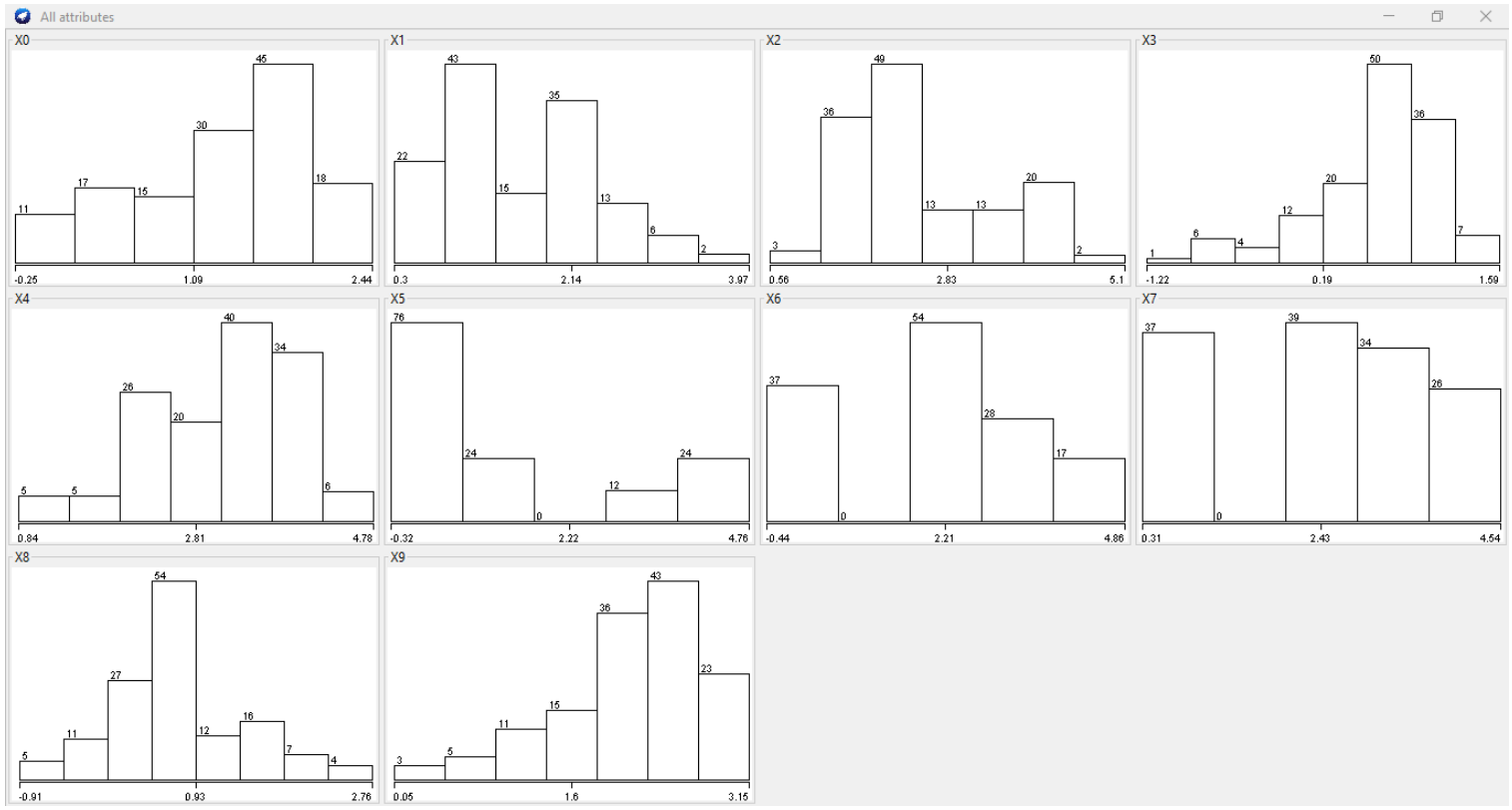
Cluster mode
(•) Use training set
( ) Supplied test set — Set...
( ) Percentage split — % 66
( ) Classes to clusters evaluation
(Num) X9
[✓] Store clusters for visualization

Ignore attributes
Start | Stop

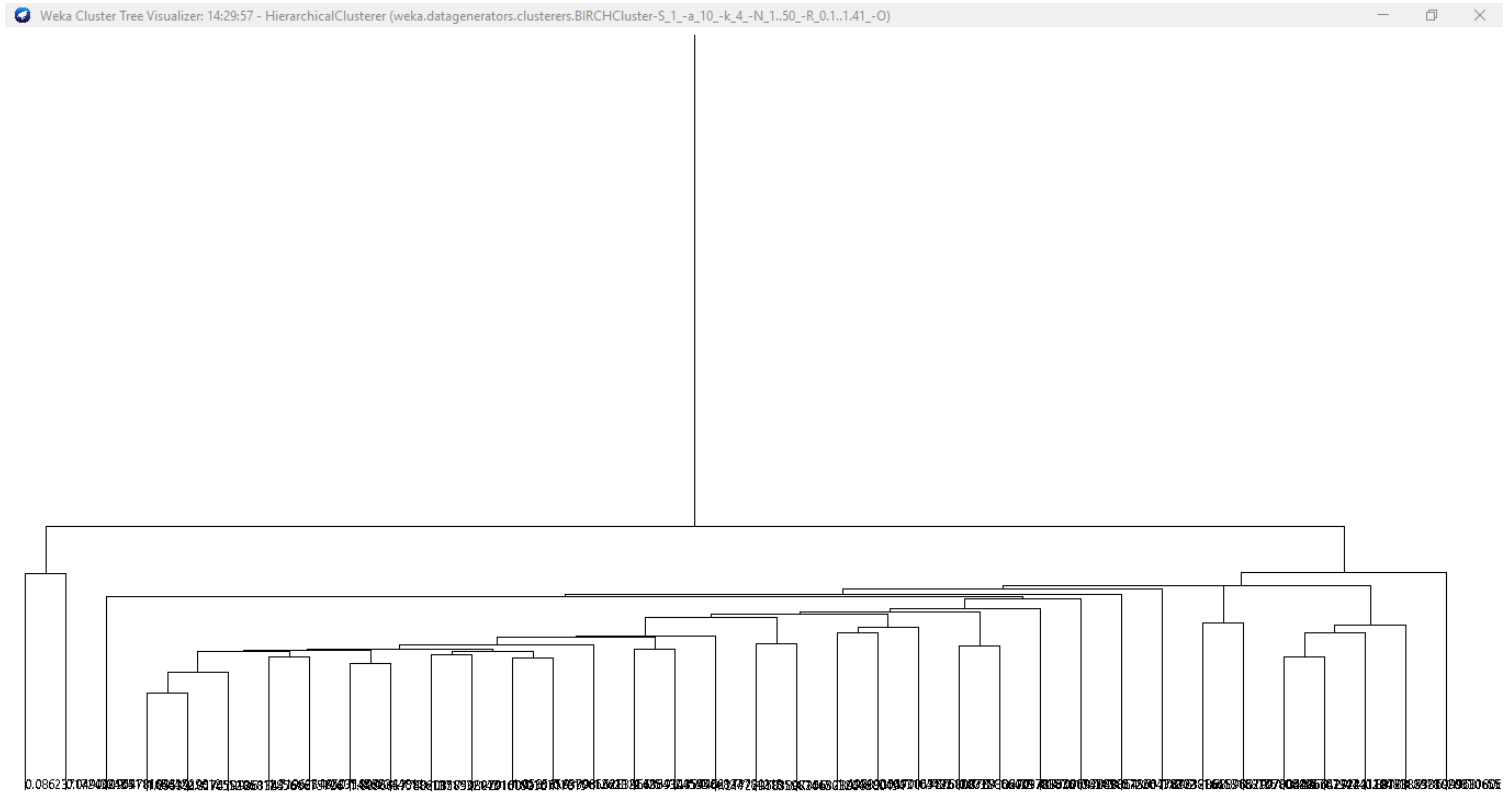
Result list (right-click for options)
14:29:57 - HierarchicalClusterer

Clusterer output

```
Attributes:   10
              X0
              X1
              X2
              X3
              X4
              X5
              X6
              X7
              X8
              X9
Test mode:    evaluate on training data


=== Clustering model (full training set) ===

Cluster 0
((0.086237172432434:0.43284,0.04900010511165642:0.43284):0.09551,((((((0.9547810443529012:0.38862,((((((((((((1.855391857655206:0.19438,2.21

Cluster 1
(((((((2.299609939167056:0.2918,2.041598242467054:0.2918):0.15113,(((1.6947640823555565:0.39476,2.5286127711263884:0.39476):0.01421,(((2.7107



Time taken to build model (full training data) : 0.1 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       36 ( 26%)
1      100 ( 74%)
```



Hierarchical Tree Visualizer:



Conclusion:

In conclusion, Weka stands as a powerful open-source data mining tool, offering a user-friendly interface, a rich set of algorithms including classification, association rule mining, and clustering. Its versatility caters to both beginners and experts, enabling effective preprocessing, experimentation, and data analysis across diverse domains. Weka's robust capabilities make it an invaluable asset for extracting insights from data, empowering users to uncover patterns, relationships, and natural groupings within their datasets.