Machine Learning

Homework2

By Mina Erfan

Dataset Info

- The dataset was created in 2014 by the University of Nottinghan, Ningbo, China
- The dataset was built from a collection of 1059 tracks covering 33 countries/area.
- The geographical location of origin was manually collected the information from the CD sleeve notes. The country of origin was determined by the artist's or artists' main country/area of residence.
- The position of each country's capital city (or the province of the area) have been taken by latitude and longitude as the absolute point of origin.
- The program MARSYAS[1] was used to extract audio features from the wave files.

Dataset

Data Set Characteristics:	Multivariate	**Number of Instances:**	1059	**Area:**	N/A
Attribute Characteristics:	Real	**Number of Attributes:**	68	**Date Donated**	2014-10-18
Associated Tasks:	Classification, Regression	**Missing Values?**	N/A	**Number of Web Hits:**	106939

Exp1: Dimension Reduction

Reducing Dimensions: Put aside samples' labels. Use PCA to reduce the dimensionality of the features based on the POV value. Repeat this process for 80%, 85%, 90%, and 95% rates respectively. Report dimensions for each POV value and train a model. Compare those results with results from Exp1 HomeWork1.

Dimension Reduction

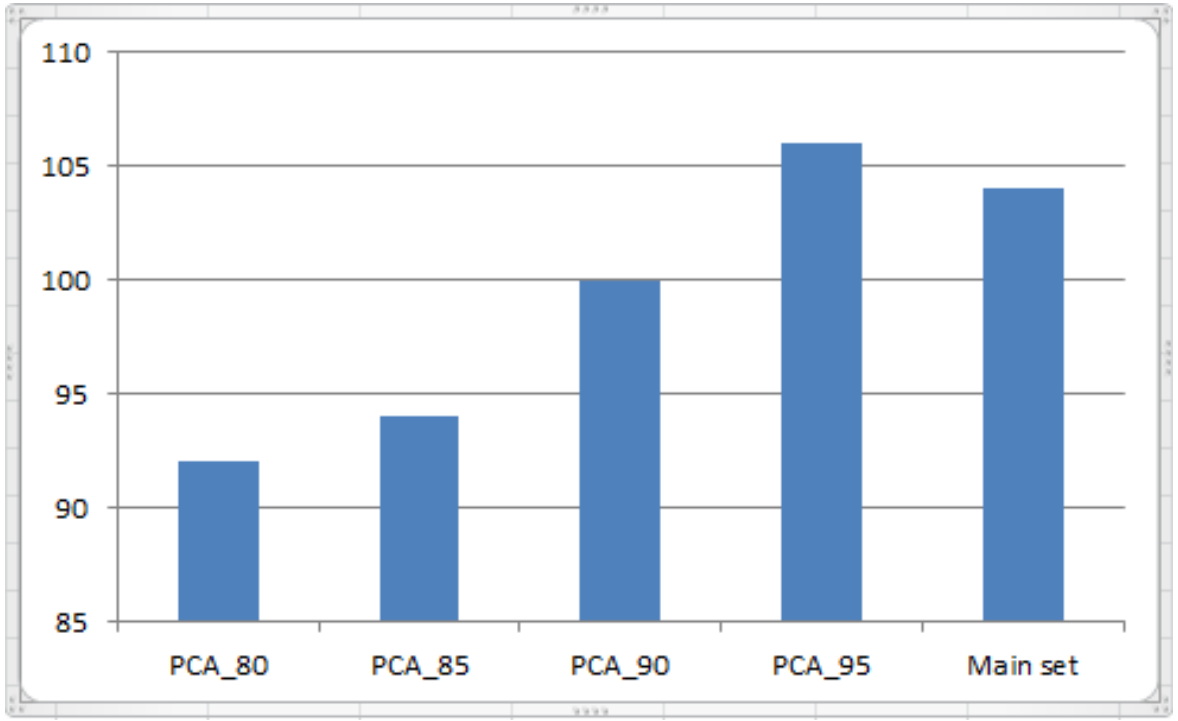
- PCA with Weka is used for dimension reduction

Weka.Filter.Principal components

Dimension Reduction

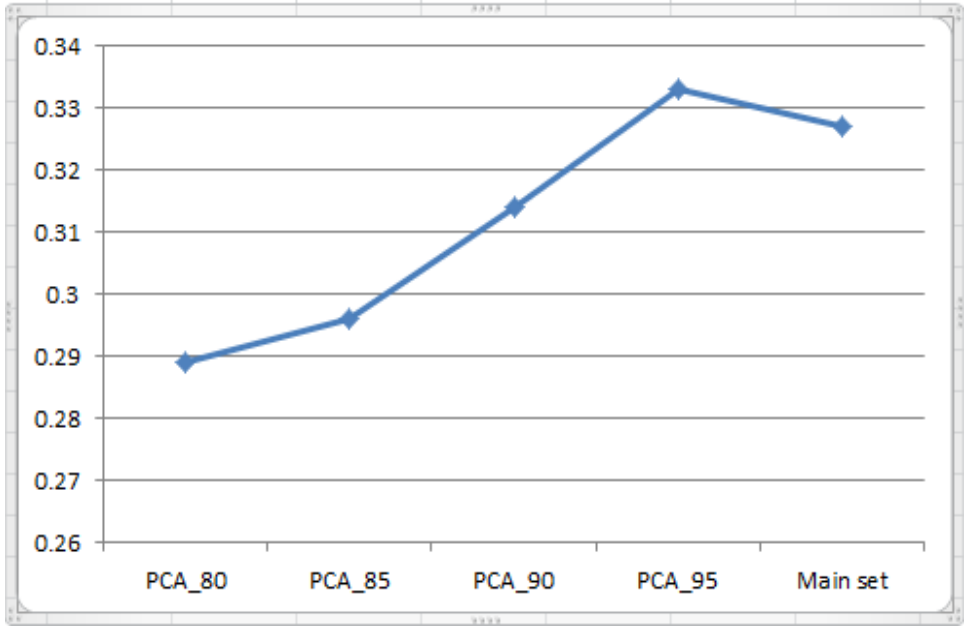
	Features	Correctly classified instances	TP Rate	FP Rate	precision	Recall	F-Measure
Main Set	68	104	0.327	0.024	0.369	0.327	0.322
PCA_80	19	92	0.289	0.026	0.3	0.289	0.276
PCA_85	24	94	0.296	0.026	0.329	0.296	0.286
PCA_90	31	100	0.314	0.026	0.347	0.314	0.308
PCA_95	40	106	0.333	0.025	0.381	0.333	0.329

Dimension Reduction

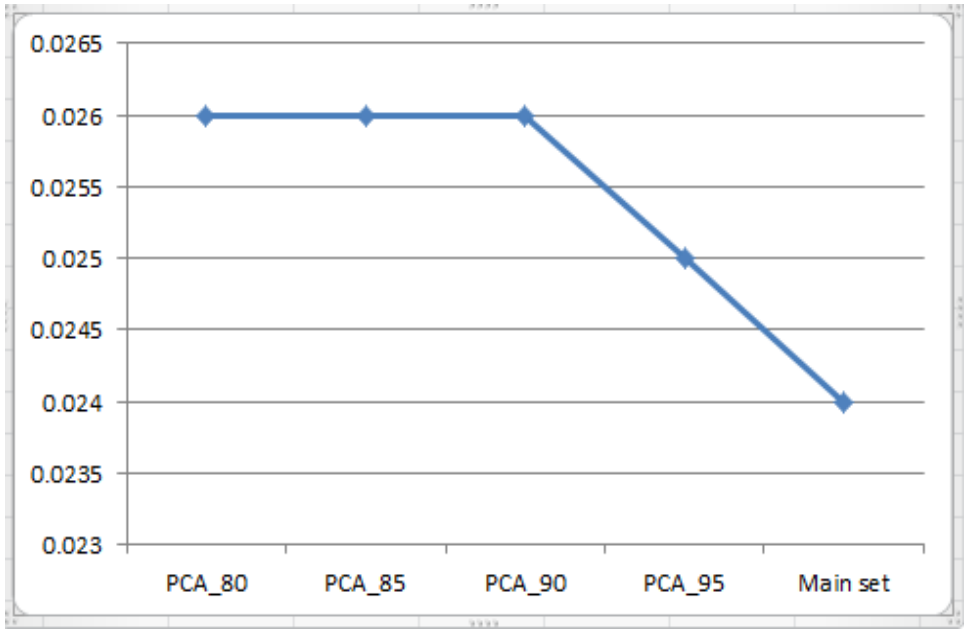


Corectly classified instances

Dimension Reduction

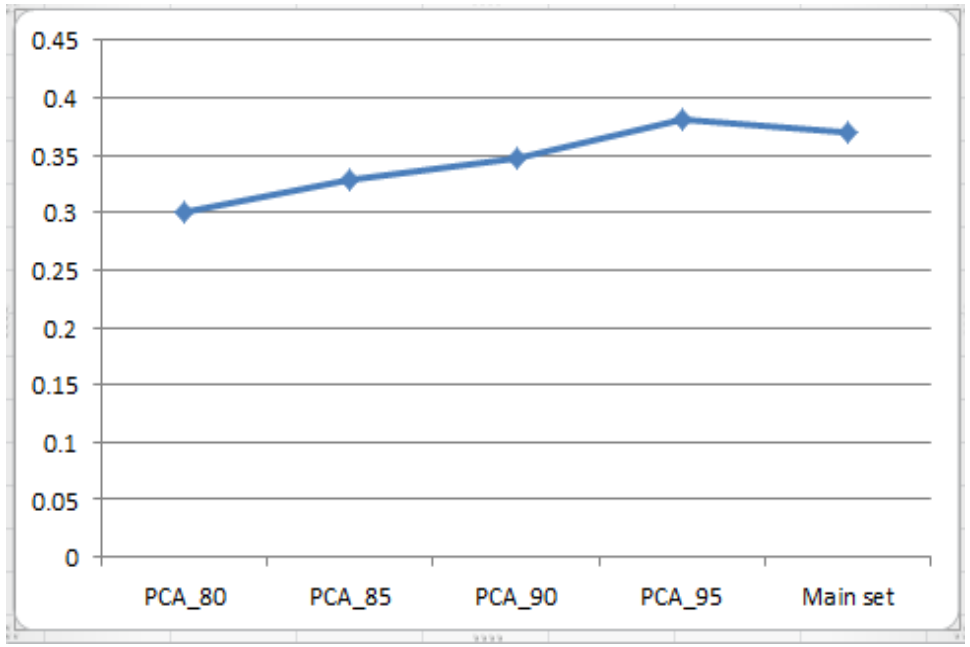


TP Rate

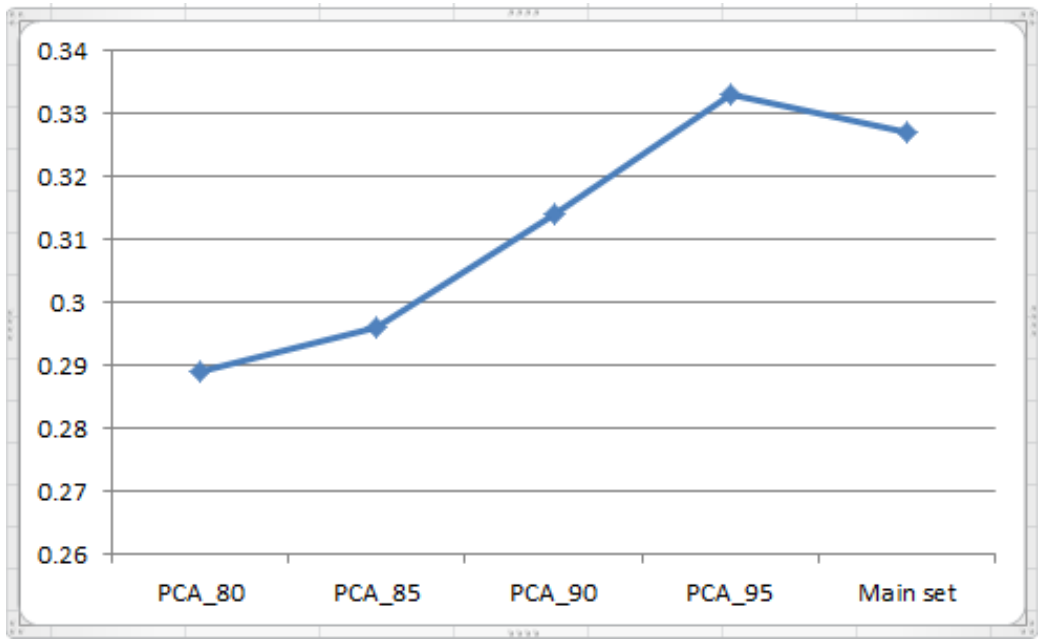


FP Rate

Dimension Reduction

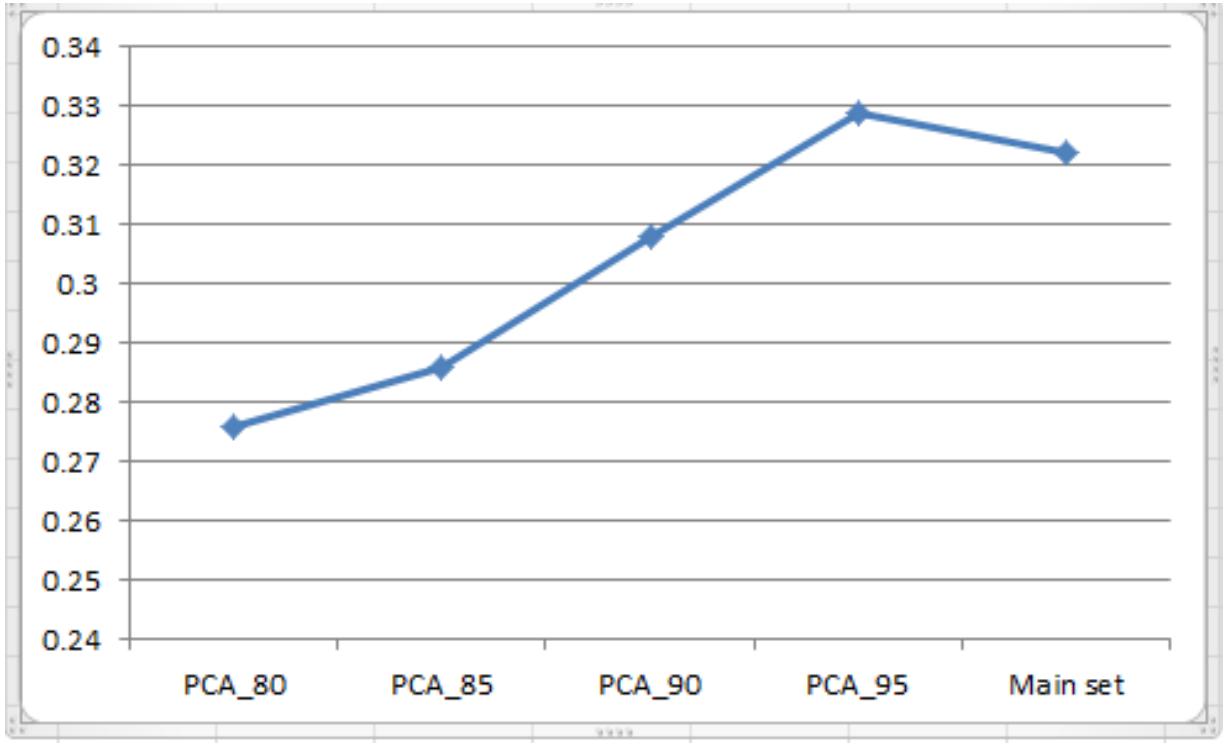


Precision



Recall

Dimension Reduction



F-Measure

Exp2: Clustering

Put aside samples' labels. Apply K-means, C-means, Fuzzy, and EM algorithms. Compare new labels with original labels and report variations for each class separately. Then, consider clusters to be the same as the number of classes. Investigate how precise samples' labels are assigned.

K-means

- Weka is used for K-means Clustering

Weka.Clusters.SimpleKmeans

: K-means

```
Clustered Instances

0       207 ( 20%)
1       155 ( 15%)
2       120 ( 11%)
3       352 ( 33%)
4       225 ( 21%)


Class attribute: continent
Classes to Clusters:

   0   1   2   3   4  <-- assigned to cluster
  22   0   1  19  27 | America
  93  15  23  83  91 | Africa
  62 107  67 131  71 | Asia
  22  33  28 118  32 | Europe
   8   0   1   1   4 | Australia

Cluster 0 <-- Africa
Cluster 1 <-- Asia
Cluster 2 <-- Australia
Cluster 3 <-- Europe
Cluster 4 <-- America

Incorrectly clustered instances :       713.0    67.3277 %
```

#samples in each class	Classes
305	Africa
405	Asia
14	Australia
266	Europe
69	America

Correctly Clustered	Incorrectly Clustered	Clustering	Class	Cluster
713= %67.32	346= %32.68	207=%20	Africa	Cluster 0
		155=%15	Asia	Cluster 1
		120=%11	Australia	Cluster 2
		352=%33	Europe	Cluster3
		225=%21	America	Cluster 4

EM

- Weka is used for EM clustering

weka.clusterer.EM

: EM

```
Clustered Instances

0       179 ( 17%)
1       208 ( 20%)
2       189 ( 18%)
3       221 ( 21%)
4       262 ( 25%)


Log likelihood: -79.49149


Class attribute: continent
Classes to Clusters:

   0   1   2   3   4  <-- assigned to cluster
  19   2   1  32  15 | America
  66  42  27 108  62 | Africa
  64 120 116  54  84 | Asia
  19  42  45  26 101 | Europe
  11   2   0   1   0 | Australia

Cluster 0 <-- America
Cluster 1 <-- Asia
Cluster 2 <-- No class
Cluster 3 <-- Africa
Cluster 4 <-- Europe

Incorrectly clustered instances :       711.0     67.1388 %
```

#samples in each class	Classes
305	Africa
405	Asia
14	Australia
266	Europe
69	America

Correctly Clustered	Incorrectly Clustered	Clustering	Class	Cluster
711= %67.13	348= %33.86	179=%17	America	Cluster 0
		208=%20	Asia	Cluster 1
		189=%18	No class	Cluster 2
		221=%21	Africa	Cluster3
		262=%25	Europe	Cluster 4

:Fuzzy C-mean in Matlab

[centers,U,objFunc] = fcm(data,Nc)

- performs fuzzy c-means clustering on the given data and returns Nc cluster centers.
- also returns the objective function values at each optimization iteration for all of the previous syntaxes.

: Fuzzy C-mean

Result :

Cluster0	Cluster1	Cluster2	Cluster3	Cluster4	
0	28	0	0	41	America
0	108	1	0	157	Europe
0	121	1	0	183	Africa
0	198	2	0	205	Asia
0	6	0	0	8	Australia

#samples in each class	Classes
305	Africa
405	Asia
14	Australia
266	Europe
69	America

Correctly Clustered	Incorrectly Clustered	Clustering	Class	Cluster
732= %69.12	327= %30.88	0=%0	No class	Cluster 0
		461=%44	Africa	Cluster 1
		4=%0	Europe	Cluster 2
		0=%0	No class	Cluster3
		594=%56	sia	Cluster 4

Exp3: agglomerative clustering

Use agglomerative clustering (single-link and complete-link separately) and depict samples' Dendrogram. Then again, consider clusters to be the same as the number of classes, and compare new labels with original labels and report conflicts. Also report and rank the best achieved number of clusters based on all clustering methods used in B and C.

Agglomerative Clustering

- Weka is used for single link clustering

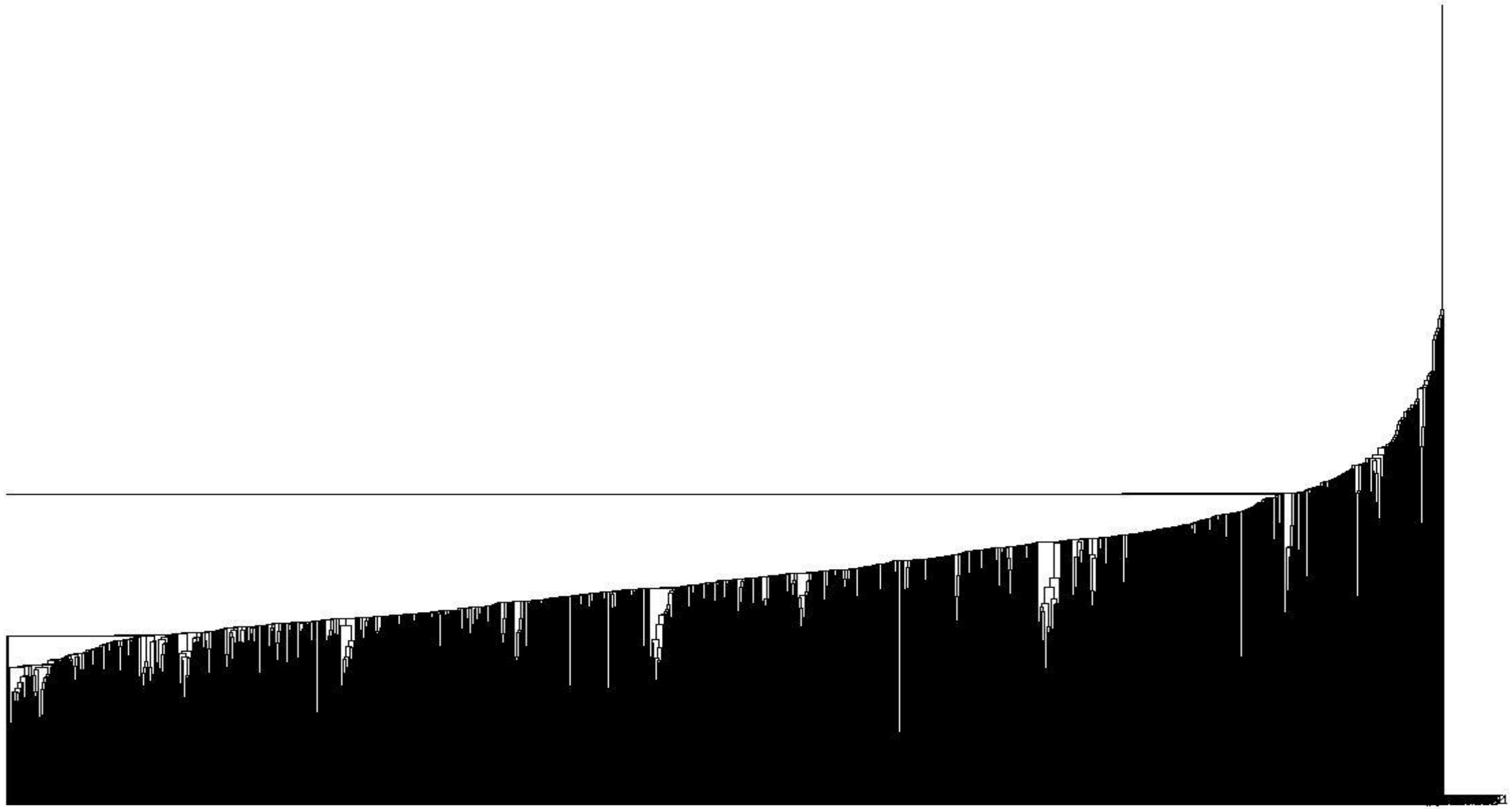
weka.clusterers.HierarchicalClusterer

: Single Hierarchical cluster

```
Clustered Instances

0      1055 (100%)
1         1 (  0%)
2         1 (  0%)
3         1 (  0%)
4         1 (  0%)


Class attribute: continent
Classes to Clusters:

   0   1   2   3   4  <-- assigned to cluster
  69   0   0   0   0 | America
 305   0   0   0   0 | Africa
 434   1   1   1   1 | Asia
 233   0   0   0   0 | Europe
  14   0   0   0   0 | Australia

Cluster 0 <-- Asia
Cluster 1 <-- No class
Cluster 2 <-- No class
Cluster 3 <-- No class
Cluster 4 <-- No class

Incorrectly clustered instances :       625.0    59.0179 %
```

#samples in each class	Classes
305	Africa
405	Asia
14	Australia
266	Europe
69	America

Correctly Clustered	Incorrectly Clustered	Clustering	Class	Cluster
625= %59.02	434= %40.98	1055=%100	Asia	Cluster 0
		1=%0	No class	Cluster 1
		1=%0	No class	Cluster 2
		1=%0	No class	Cluster3
		1=%0	No class	Cluster 4



Agglomerative Clustering

- Weka is used to do clustering using complete link

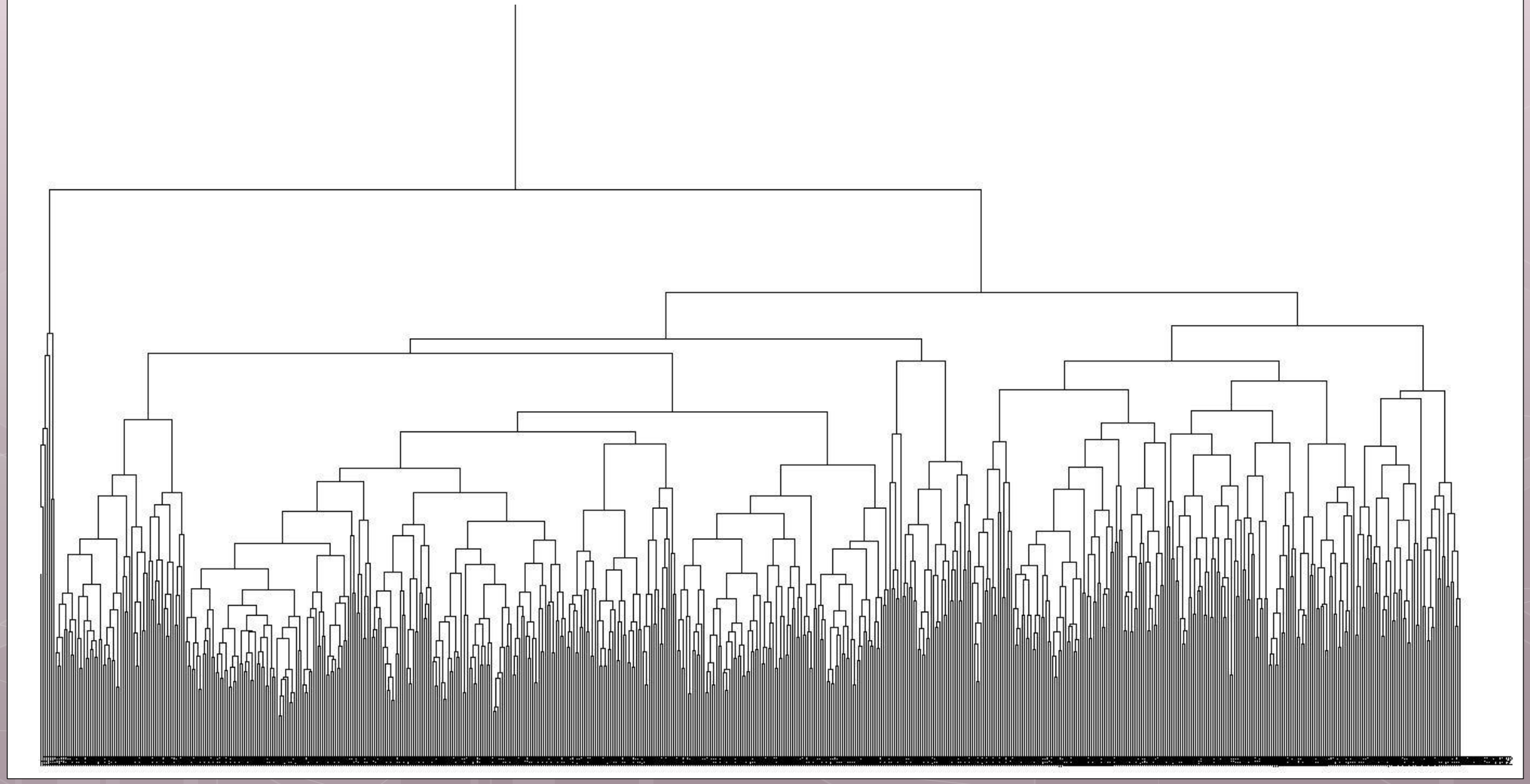
weka.clusterers.HierarchicalClusterer

: Complete Hierarchicalcluster

```
Clustered Instances

0       657 ( 62%)
1       328 ( 31%)
2        71 (  7%)
3         2 (  0%)
4         1 (  0%)


Class attribute: continent
Classes to Clusters:

   0   1   2   3   4  <-- assigned to cluster
  58  11   0   0   0 | America
 231  60  14   0   0 | Africa
 211 181  43   2   1 | Asia
 147  72  14   0   0 | Europe
  10   4   0   0   0 | Australia

Cluster 0 <-- Africa
Cluster 1 <-- Asia
Cluster 2 <-- Europe
Cluster 3 <-- No class
Cluster 4 <-- No class

Incorrectly clustered instances :       633.0     59.7734 %
```

#samples in each class	Classes
305	Africa
405	Asia
14	Australia
266	Europe
69	America

Correctly Clustered	Incorrectly Clustered	Clustering	Class	Cluster
633= %59.77	426= %40.23	657=%62	Africa	Cluster 0
		328=%31	Asia	Cluster 1
		71=%7	Europe	Cluster 2
		2=%0	No class	Cluster3
		1=%0	No class	Cluster 4



Comparison

	Correctly classified instances	Incorrectly classified instances
Fuzzy C-mean	30.88 %	69.12 %
K-means	32.68 %	67.32 %
EM	33.86 %	67.13 %
complete hierarchical	40.23 %	59.77 %
Single hierarchical	40.98%	59.02 %

Thank You