# Clustering datasets

Speech and Image Processing Unit
School of Computing
University of Eastern Finland
P.O.Box 111
FIN-80101 Joensuu
Finland

## Image data

**Bridge**
(256x256)

4096 vectors,
16-d

4x4 pixel blocks ts txt
4x4 binarized pixel blocks ts txt
4x4 pixel blocks: 25% randomly sampled (for training) ts txt
4x4 pixel blocks: 75% randomly sampled (for testing) ts txt

**House**
(256x256)

34112 vectors,
3-d

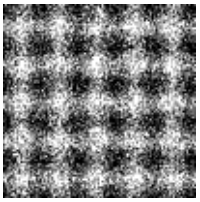
RGB-values, quantized to 5 bits per color ts txt
RGB-values, 8 bits per color ts txt

**Miss America**
(360x288)

6480 vectors,
16-d

4x4 pixel blocks from the difference image of frame 1 and 2 ts txt
4x4 pixel blocks from the difference image of frame 2 and 3 ts txt

## Birch-sets

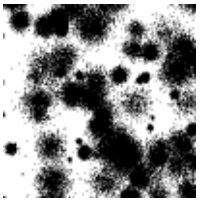


**Birch1**

**Birch2**

Synthetic 2-d data with 100 000 vectors and 100 clusters.

Zhang et al., "BIRCH: A new data clustering algorithm and its applications", *Data Mining and Knowledge Discovery*, 1 (2), 141-182, 1997.



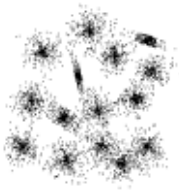
**Birch3**

**Birch1:** Clusters in regular grid structure ts txt
**Birch2:** Clusters at a sine curve ts txt
**Birch3:** Random sized clusters in random locations ts txt

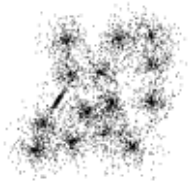
## S-sets

Synthetic 2-d data with 5000 vectors and 15 Gaussian clusters with different degree of cluster overlapping.

**S1**



**S2**



**S3**

**S4**

P. Fränti and O. Virmajoki, "Iterative shrinking method for clustering problems", *Pattern Recognition*, 39 (5), 761-765, May 2006.

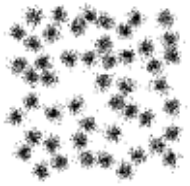
**S1:** ts txt
**S2:** ts txt
**S3:** ts txt
**S4:** ts txt

**Source and labels:** zip

# A-sets

**A1**
3000 vectors,
20 clusters

**A2**
5250 vectors,
35 clusters



**A3**
7500 vectors,
50 clusters

Synthetic 2-d data with varying number of clusters and vectors.

**A1:** ts txt
**A2:** ts txt
**A3:** ts txt

# Dim-sets

**Dim2**

Synthetic data with Gaussian clusters in multi-dimensional space.
1351-10126 vectors, 2-d - 15-d

ts txt

# DIM-sets (other)

**DIM032**
1024 vectors,
16 clusters
32 dimensions

**DIM064**
1024 vectors,
16 clusters
64 dimensions

Dim-sets.

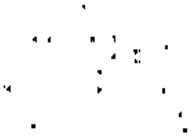
**DIM032:** ts txt
**DIM064:** ts txt
**DIM128:** ts txt
**DIM256:** ts txt
**DIM512:** ts txt
**DIM1024:** ts txt

Ground truths in cb and txt format.

**DIM128**
1024 vectors,
16 clusters
128 dimensions

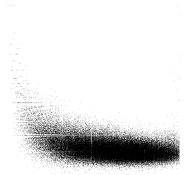
**DIM256**
1024 vectors,
16 clusters
256
dimensions

**DIM512**
1024 vectors,
16 clusters
512 dimensions



**DIM1024**
1024 vectors,
16 clusters
1024
dimensions

# KDDCUP04Bio set



**KDDCUP04Bio**
145751 vectors,
2000 clusters
74 dimensions

KDDCUP04Bio biology dataset.

**KDDCUP04Bio:** ts txt

# UCI datasets

**Thyroid**
215 vectors,
2 clusters
5 dimensions

Thyroid dataset.

**Thyroid:** ts txt

**Wine**
178 vectors,
3 clusters
13 dimensions

Wine dataset.

**Wine:** ts txt

**Yeast**
1484 vectors,
10 clusters
8 dimensions

Yeast dataset.

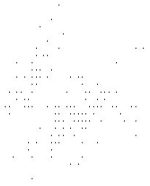
**Yeast:** txt
**Yeast_times100:** ts txt

**Breast**
699 vectors,
2 clusters
9 dimensions

Breast-cancer-Wisconsin dataset.

**Breast:** ts txt

info



**Iris**
150 vectors,
4 dimensions
3 clusters

Iris dataset.

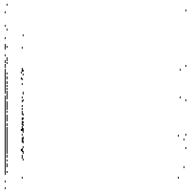
**Iris:** ts
txt without labels
txt with labels



**Glass**
214 vectors,
9 dimensions
7 clusters

Glass dataset.

**Glass:** ts
txt without labels
txt with labels



**Wdbc**
569 vectors,
32 dimensions
2 clusters

Wdbc dataset.

**Wdbc:** ts
txt numeric, 31 dim.
txt

# g2 sets

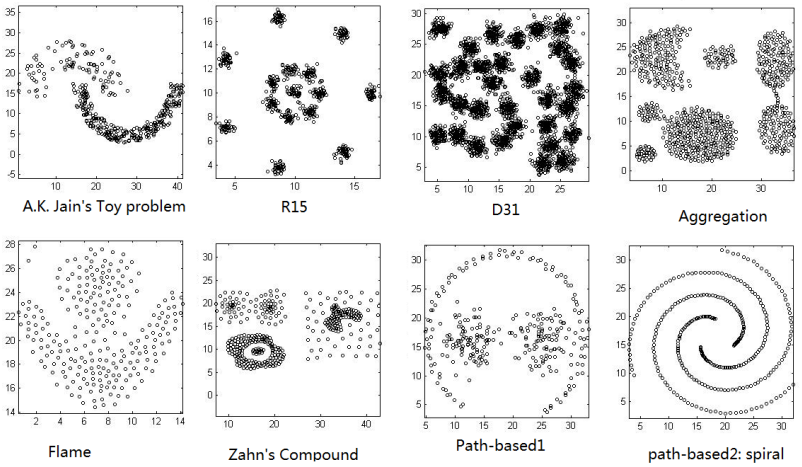
**g2-2-30**

Gaussian clusters dataset.

**g2:** ts's in zip file (53MB)

1024 vectors per cluster,
2 clusters
1-1024 dimensions
variance 10-100

# Shape sets




Third column is the label.

**Aggregation**
788 vectors,
2 dimensions
7 clusters

**Aggregation:** txt
Gionis, A., H. Mannila, and P. Tsaparas, Clustering aggregation. ACM Transactions on Knowledge Discovery from Data (TKDD), 2007. 1(1): p. 1-30.

**Compound**
399 vectors,
2 dimensions
6 clusters

**Compound:** txt
Zahn, C.T., Graph-theoretical methods for detecting and describing gestalt clusters. Computers, IEEE Transactions on, 1971. 100(1): p. 68-86.

**Pathbased**
300 vectors,
2 dimensions
3 clusters

**Pathbased:** txt
Chang, H. and D.Y. Yeung, Robust path-based spectral clustering. Pattern Recognition, 2008. 41(1): p. 191-203.

**Spiral**
312 vectors,
2 dimensions
3 clusters

**Spiral:** txt
Chang, H. and D.Y. Yeung, Robust path-based spectral clustering. Pattern Recognition, 2008. 41(1): p. 191-203.

**D31**
3100 vectors,
2 dimensions
31 clusters

**D31:** txt
Veenman, C.J., M.J.T. Reinders, and E. Backer, A maximum variance cluster algorithm. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2002. 24(9): p. 1273-1280.

**R15**
600 vectors,
2 dimensions
15 clusters

**R15:** txt
Veenman, C.J., M.J.T. Reinders, and E. Backer, A maximum variance cluster algorithm. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2002. 24(9): p. 1273-1280.

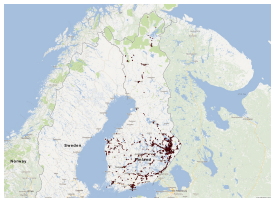
**Jain**
373 vectors,
2 dimensions
2 clusters

**Jain:** txt
Jain, A. and M. Law, Data clustering: A user's dilemma. Lecture Notes in Computer Science, 2005. 3776: p. 1-10.

**Flame**
240 vectors,
2 dimensions
2 clusters

**Flame:** txt
Fu, L. and E. Medico, FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data. BMC bioinformatics, 2007. 8(1): p. 3.

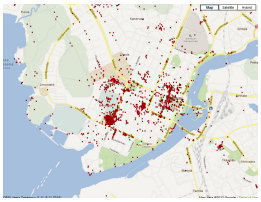
# Mopsi locations



**Users' locations**
13467 vectors,
2 dimensions

Mopsi locations Finland until 2012 dataset.

**Users' locations:** cb txt

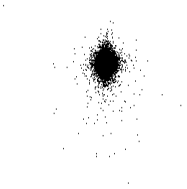


**Users' locations, Joensuu**
6014 vectors,
2 dimensions

Users' locations in Joensuu 2012 dataset.

**Users' locations Joensuu:** ts txt

# Europe



**Europe**
169308 vectors,
2 dimensions

Europe dataset.

**Europe:** ts txt

# Census

Census dataset.

census.zip

Includes files:

census1000.ts
census2000.ts
census4000.ts
census8000.ts
census16000.ts
census32000.ts
census64000.ts
census128000.ts
census256000.ts
census512000.ts

# Miscellaneous

| | |
|---|---|
| **ConfLongDemo_JSI_164860** | ConfLongDemo_JSI_164860.txt |
| **t4.8k** | t4.8k.txt |
| **MINST** | MINST.txt |

## Related links

- Programming interface (modu*.zip) to handle data sets (cb/ts-format)
- Windows software for visualizing data sets
- Software for converting data sets to text

- PPM/PNM/PBM image formats