Clustering by Nonnegative Matrix Factorization Using Graph Random Walk (Suppemental Document)

Anonymous Author(s)
Affiliation
Address
`email`

1 Datasets

Brief description of the datasets

- STRIKE: the Pajek *Strike* dataset, a small social network during a strike.
- KOREA: the Pejek *Korea* dataset, a communication network within a small enterprise.
- AMLALL: the AML/ALL leukemia gene expression data, with 5000 genes.
- DUKE: the LIBSVM *duke breast-cancer* dataset, for predicting the clinical status of human breast cancer by using gene expression profiles, originally with 7129 features.
- HIGHSCHOOL: the Pejek *Highschool* dataset, a friend network in a high school; we used a subset of the five largest classes.
- KHAN: the *khan* dataset from the textbook "The Elements of Statistical Learning", gene expression data, 2318 genes.
- POLBOOKS: the *Books about US politics* dataset from Newman's collection, a network of books about US politics published around the time of the 2004 presidential election and sold by the online bookseller Amazon.com.
- FOOTBALL: the Pejek *Football* dataset, network data, games between 22 soccer teams which participated in the World Championship in Paris, 1998.
- IRIS: the UCI *Iris* dataset.
- CANCER: the *14cancer* dataset from the textbook "The Elements of Statistical Learning", gene expression data of 16063 genes.
- SPECT: the UCI *Low Resolution Spectrometer* dataset, spectra derived from the IRAS-LRS (Infra-Red Astronomy Satellite-Low Resolution Observation) database, originally with 102 attributes.
- ROSETTA: microarray data set from Rosetta Inpharmatics, Inc, originally with 12634 dimensions.
- ECOLI: the UCI *Ecoli* dataset, containing protein localization sites, originally with 8 attributes.
- IONOSPHERE: the UCI *ionosphere* dataset, for classification of radar returns from the ionosphere, originally with 34 attributes.
- ORL: the AT&T ORL database of face images, each image of size 92×112
- UMIST: the *Sheffield (previously UMIST) Face Database*, each face image of size 92×112.
- WDBC: the the LIBSVM *breast-cancer* dataset, originally named *Wisconsin Breast Cancer* in UCI, with 10 features.

Table 1: Dataset statistics

Dataset	#samples	#classes	Domain	Source
STRIKE	24	3	social	PAJEK
KOREA	35	2	social	PAJEK
AMLALL	38	3	gene	AMLALL
DUKE	44	2	medical	LIBSVM
HIGHSCHOOL	60	5	social	PAJEK
KHAN	83	4	gene	ELML
POLBOOKS	105	3	social	NEWMAN
FOOTBALL	115	12	social	PAJEK
IRIS	150	3	biology	UCI
CANCER	198	14	medical	ELML
SPECT	267	3	astronomy	UCI
ROSETTA	300	5	gene	ROSETT
ECOLI	327	5	protein	UCI
IONOSPHERE	351	2	radar	UCI
ORL	400	40	image	ORL
UMIST	575	20	image	UMIST
WDBC	683	2	medical	LIBSVM
DIABETES	768	2	medical	LIBSVM
VOWEL	990	11	audio	LIBSVM
MED	1033	31	text	LSI
PIE	1166	53	image	PIE
YALEB	1292	38	image	YALEB
TERROR	1293	6	social	LINQS
ALPHADIGS	1404	36	image	ROWEIS
COIL-20	1440	20	image	COIL
YEAST	1484	10	biology	UCI
SEMEION	1593	10	image	UCI
FAULTS	1941	7	steel	UCI
SEG	2310	7	image	UCI
ADS	2359	2	network	CHEN
CORA	2708	7	text	LINQS
MIREX	3090	10	music	CHEN
CITESEER	3312	6	texts	LINQS
WEBKB4	4196	4	texts	CMUTE
7SECTORS	4556	7	texts	CMUTE
SPAM	4601	2	email	ELML
CURETGREY	5612	61	image	CURET
OPTDIGITS	5620	10	image	UCI
GISETTE	7000	2	image	LIBSVM
REUTERS	8293	65	texts	UCI
RCV1	9625	4	texts	RCV1
PENDIGITS	10992	10	image	UCI
PROTEIN	17766	3	protein	LIBSVM
20NEWS	19938	20	texts	CMUTE
MNIST	70000	10	image	MNIST
SEISMIC	98528	3	sensor	LIBSVM

- DIABETES: the LIBSVM *diabetes* dataset, originally from UCI, with 8 features.
- VOWEL: the LIBSVM *vowel* dataset, originally from UCI, with 10 features.
- MED: the *MED* abstract text collection, with 5831 words.
- PIE: the *PIE* face image dataset, each of size 32×32
- YALEB: the *Yale-B* face image dataset, each of size 168×192.
- TERROR: the *Terrorist Attacks* from LINQS, relationships of terrorism attack entities.

Table 2: Data sources

PAJEK	`http://vlado.fmf.uni-lj.si/pub/networks/data/`
NEWMAN	`http://www-personal.umich.edu/~mejn/netdata/`
LIBSVM	`http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`
ELML	`http://www-stat-class.stanford.edu/~tibs/ElemStatLearn/data.html`
UCI	`http://archive.ics.uci.edu/ml/`
LINQS	`http://www.cs.umd.edu/projects/linqs/projects/lbc/index.html`
ROWEIS	`http://www.cs.nyu.edu/~roweis/data.html`
CHEN	`http://idl.ee.washington.edu/SimilarityLearning/`
ORL	`http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html`
UMIST	`http://www.sheffield.ac.uk/eee/research/iel/research/face`
PIE	`http://vasc.ri.cmu.edu/idb/html/face/`
YALEB	`http://vision.ucsd.edu/~leekc/ExtYaleDatabase/ExtYaleB.html`
COIL	`http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php`
MNIST	`http://yann.lecun.com/exdb/mnist/`
20NEWS	`http://people.csail.mit.edu/jrennie/20Newsgroups/`
WEBKB	`http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/`
CMUTE	`http://www.cs.cmu.edu/~TextLearning/datasets.html`
RCV1	`http://www.ai.mit.edu/projects/jmlr/papers/volume5/lewis04a/`
CURET	`http://www.robots.ox.ac.uk/~vgg/research/texclass/index.html`
LSI	`http://web.eecs.utk.edu/research/lsi/`
AMLALL	[1]
ROSETT	[2]

- ALPHADIGS: the *Binary Alphadigits* dataset from Sam Roweis' collection, each sample is a binary image of a letter or digit, of size 20×16
- COIL-20: the *COIL-20* dataset from Columbia University Image Library, toy images of different angles, each image of size 128×128.
- YEAST: the UCI *Yeast* dataset, for predicting the cellular localization sites of proteins, originally with 8 features.
- SEMEION: the UCI *Semeion Handwritten Digit* dataset; each sample is an image of size 16×16
- FAULTS: the UCI *Steel Plates Faults* dataset, steel plate faults, classified into 7 different types, originally with 27 dimensions.
- SEG: the UCI *Image Segmentation* dataset, image patches from 7 outdoor images, originally with 19 high-level features.
- ADS: the *Internet_Ads* similarity dataset from Chen's collection.
- CORA: the LINQS *Cora* dataset, text documents, with 1433 words
- MIREX: the *Mirex07* similarity dataset from Chen's collection.
- CITESEER: the LINQS *CiteSeer* dataset, text documents, with 3703 words
- WEBKB4: the *WebKB4* dataset from CMU Text Learning group, text documents; 10,000 words with maximum information gain are preserved.
- 7SECTORS: the *4 Universities* dataset from CMU Text Learning group, text documents classified to 7 sectors; 10,000 words with maximum information gain are preserved.
- SPAM: the *spam* dataset from the textbook "The Elements of Statistical Learning", originally with 57 dimensions.
- CURETGREY: texture image data, each image of size 200×200. We down-sampled images to 100×100 before extracting the scattering features.
- OPTDIGITS: the UCI *optical recognition of handwritten digits*, originally with 64 dimensions.
- GISETTE: the LIBSVM *gisette* dataset, handwritten digits, subset of MNIST, originally with 5000 dimensions.

- REUTERS: the UCI *Reuters-21578* dataset, text documents, with 18933 words.
- RCV1: text documents from four classes, with 29992 words.
- PENDIGITS: the UCI *pen-based recognition of handwritten digits* dataset, originally with 16 dimensions.
- PROTEIN: the LIBSVM *protein* dataset, originally with 357 dimensions.
- 20NEWS: text documents from 20 newsgroups; 10,000 words with maximum information gain are preserved.
- MNIST: handwritten digit images, each of size 28×28.
- SEISMIC: the LIBSVM *SensIT Vehicle (seismic)* dataset, distributed sensor network data for vehicle classification, originally with 50 dimensions.

2 External algorithm sources

- *1-Spectral Clustering*[1]
- *Interaction Component Model*[2]

References

[1] Jean-Philippe Brunet, Pablo Tamayo, Todd R. Golub, and Jill P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences*, 101(12):4164–4169, 2004.

[2] P. Kim and B. Tidor. Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome Research*, 13(7):1706–1718, 2003.

[1] `http://www.ml.uni-saarland.de/code/oneSpectralClustering/oneSpectralClustering.html`

[2] `http://netpro.r-forge.r-project.org/`