Spectral Clustering Toolbox

Deepak Verma
deepak@cs.washington.edu

Marina Meila
mmp@stat.washington.edu

July 8, 2014

1 Introduction

This toolbox contains the code written to perform various spectral clustering algorithms. The details related to the code and some experiments is available in [VM03]. This document is very short and the reader is encourage to look at the directories to see other code.

2 Using the toolbox

2.1 Quick Start

To get up and cranking :

1. Set SPECTRAL_HOME to be the directory where you unpacked the library.
2. Start matlab.
3. Call `init_spectral`. (sets up the path and global options).
4. `assignment=cluster_algo(similarity,number_of_clusters)` : Gives you the desired clustering.
5. Remember that all the vectors that you see would be column vectors.

2.2 Data Input/Output

Reading a data file (see `data` directory for some examples). :
`[similarity,cluster_assignments,points]=read_from_data_file(filePrefix,directory)`
Reads the data file `directory/filePrefix` (default dir=`data`) and assigns the `similarity`, the points and true `cluster_assignments`. If either of the the above is not defined empty matrix is returned.

2.3 Spectral Algorithms

The algorithms are present in the `algos` and `algos/allalgos` directory. The latter just contains files which act convenient shortcut names to popular algorithms. Algorithm `njw` is described in [NJW02] and `mcut` is described in [MS00]. For the details and comparison of all the algorithms see [VM03].

3 Experimental Framework

3.1 Running Experiments

To run a bunch of experiments together use:
`run_single_experiment(dataFile,cluster_algo_list,k_range,sigma,iterations,outdir,plot_points)`
This runs the experiments on `dataFile` for the algorithms `cluster_algo_list` ,varying the input number of clusters in the list `k_range`. The `iterations` is the *list* of iterations indices and are useful when there is a random

element in the algorithm. `sigma` is the σ used for affinity matrix ([NJW02] in case the points (see section 2.2 are present in `dataFile`. The results of each algorithm is written a file in the `outdir` (with a default value used). If `plot_points` is 1 then the results are displayed after each iteration for 2D points. (default 0).

3.2 Plotting graphs

To the plot the graphs on the experiments ran using `run_single_experiment` use:

```
plot_metric_save(dataFile,cluster_algo_list,k_range,iterations,metric,plot_stdev,outdir)
```

The arguments mean the same as above. `metric` is used specify the metric to be used to compare clustering produced w.r.t. true clustering. The metrics available are

- `vi` : Variation of Information ([Mei02]).
- `ce` : Clustering Error (see [VM03] for details).
- `wi` : One sided Wallace Index ([Wal83], also see [VM03]).

4 Datasets

4.1 Artificial

Some artificial datasets are provided in the `data` directory. All of them are 2D points which offers various levels of difficulty to the spectral algorithms. They are modelled after [NJW02]. To see these (or any other 2D) plots use `plot2Dpoints_with_clusters`.

An interesting dataset (not in 2D) called `block-stochastic` ([MS00]) is also provided. It is a similarity matrix designed ([VM03]) to illustrate the case when spectral methods work and linkage based methods fail.

4.2 Real Datasets

Coming soon....

References

[Mei02] Marina Meila. Comparing clusterings. Technical Report 418, UW Statistics Department, 2002.

[MS00] Marina Meila and Jianbo Shi. Learning segmentation by random walks. In *NIPS*, pages 873–879, 2000.

[NJW02] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 849–856, Cambridge, MA, 2002. MIT Press.

[VM03] Deepak Verma and Marina Meila. A comparison of spectral methods. Technical Report UW-CSE-03-05-01, Dept. of Computer Science and Engineering, University of Washington, 2003.

[Wal83] David L. Wallace. Comment. *J. Amer. Statist. Assoc.*, pages 269 – 576, 1983.