Notes

Parameterless clustering by dynamic tree-cutting

Simon Lehmann Knudsen
simkn15@student.sdu.dk
ECTS: 10
04/09-2017 - 31/01-2018
5th semester in Computer Science

December 16, 2017

Contents

1 Testing

1.1 Generating randomized data sets: TODO explain multidimensional scaling and what it does

Skip talking about first approach ????? First approach test output files located in folder randomSimV1

```
The first approach was to take the upper matrix and randomize the similarities
according to original similarities.  Such that the random similarities
was taken from the set of similarities from the original data, while
the same number could be taken multiple times.  However this approach
turned out as misleading, as the costs from tclust peaked at 1.000.000,
and the actual cost is peaking at around 60.000.
```

Next approach required looking into multidimensional scaling, as this should make a better scaling of the randomized similarities. MDS made it possible to scale the data into higher dimensions, making a "better match" of similarities. Already at five dimensions the costs is peaking at around 220.000. A big reduction in the costs from the first approach.

1.2 Evaluation of cost plots

With respect to the fold change we do not see any interesting pattern with the small data set. However, the big data set shows somewhat of a pattern with a foldchange just above 4. The lowest fold change is 3.6639 at 20 dimensions, where the highest is 6.5309 at 10 dimensions. The highest looks like an outlier, since it is much higher than the other values. Looking at the second highest at 100 dimensions with 4.9422 we get much closer to the mean of 4.2535. As all the plottings show, the threshold which has the biggest gap in cost between random and original data is extremely close to the highest cost peak in the random data. In 19 out of 20, the threshold with the biggest gap in cost is the highest peak of cost in the random data set.

1.2.1 Small Data Set

1.2.2 Big Data Set

1.3 Small Data Set

1.3.1 Original

1.3.2 Random

1.4 Big Data Set

1.4.1 Original

1.4.2 Random

1.5 Hierarchical Clustering

1.5.1 Binary Search of Threshold

Dimensions	Threshold max gap	fold change
5	122	227.3405
10	151	39.0570
15	153	35.0167
20	153	34.4680
25	148	27.9837
30	160	30.6777
35	128	101.9536
40	140	36.7628
45	139	39.2452
50	147	24.3221
55	125	147.9133
60	123	165.325
65	125	149.7192
70	118	352.4266
75	125	144.0012
80	134	57.8595
85	124	154.1030
90	130	73.9996
95	124	145.2467
100	136	43.5962

Figure 1: Small Data Set

2 Theory

2.1 TransClust / Transitivity Clustering / Weighted Graph Projection problem

partitioningBiologicalDataWithTransitivityClusteringSUPPLEMENT.pdf
ComprehensiveClusterAnalysisWithTransitivityClustering.pdf
largeScaleClusteringOfProteinSequencesWithFORCE-aLayoutBasedHeuristicForWeightedClus

2.2 Hierarchical Clustering / Dendrogram

2.3 Cluster Validation

2.4 Multidimensional Scaling

Dimensions	Max cost Threshold	Threshold max gap	fold change ($\mu = 4.25353$)
5	78	78	4.9265
10	107	107	6.5309
15	92	92	3.9525
20	84	84	3.6639
25	90	90	3.6681
30	96	96	4.2508
35	98	97	4.6519
40	87	87	3.8082
45	93	93	4.1851
50	89	89	3.7935
55	90	90	3.8843
60	95	95	4.3252
65	91	91	3.8819
70	91	91	3.9391
75	94	94	4.1228
80	101	101	4.2029
85	91	91	4.0458
90	100	100	4.3406
95	93	93	3.9544
100	106	106	4.9422

Figure 2: Big Data Set

3 Cluster Analysis - Book

3.1 Chapter 1: An introduction to classification and clustering

3.2 Chapter 2: Detecting clusters graphically

3.3 Chapter 3: Measurement of proximity

3.4 Chapter 4: Hierarchical clustering

There are two forms of hierarchical Clustering, agglomerative and divisive. Agglomerative is starting with n clusters and joins them until one cluster is left. Divisive is the opposite, going from one to n clusters. With hierarchical methods, joins or split of clusters are irrevocable. Thus these operations cannot be undone. Agglomerative are the most common procedure, since divisive is very expensive procedure. One of the important issues with hierarchical clustering is deciding on the correct number of clusters. Typically the result of a hierarchical clustering is represented as a dendrogram, a two-dimensional diagram.

3.4.1 Agglomerative

Most widely used. Produces a series of partitions of the data. First consists of `n` cluster, all of size one. Last one consists of one cluster of size `n`. Some of the basic operations to fuse the closest clusters are `single linkage` and `Centroid linkage`. The differences arises in terms of which distance measure is preferred.

- Table 4.1
- Single linkage: Distance is defined as the closest pair of individuals between two clusters. The pair consists of one individual from each cluster.
- Centroid linkage: Calculation the euclidean distance. Requires access to the original data. Operates on a proximity matrix.
- Complete linkage: Furthest neighbour, distance of most distant pair of individuals.
- Average linkage: Average distance from all pairs of individuals that are made with one individual from each cluster.
- Median linkage: .
- Weighted average linkage: .
- Centroid clustering: Uses median linkage.
- Error sum of squares, page 77.

3.4.2 Divisive

This procedure is very expensive, thus heuristics are needed.

3.5 Chapter 5: Optimization clustering techniques

3.6 Chapter 6: Finite mixture densities as models for cluster analysis

3.7 Chapter 7: Model-based cluster analysis for structured data

3.8 Chapter 8: Miscellaneous clustering methods

3.9 Chapter 9: Some final comments and guidelines

4 Articles

4.1 Articles Sources

- A gold standard set of mechanistically diverse enzyme superfamilies.
  Shoshana D Brown*, John A Gerlt†, Jennifer L Seffernick‡ and Patricia C Babbitt§.
  Article downloaded from SDU library and located in articles folder.

- Comprehensive cluster analysis with Transitivity Clustering.
  Tobias Wittkop, Dorothea Emig, Anke Truss, Mario Albrecht, Sebastian Böcker & Jan Baumbach
  Article downloaded from SDU library and located in articles folder.

- Comparing the performance of biomedical clustering methods.
  Christian Wiwie1, Jan Baumbach & Richard Röttger
  Article downloaded from sdu library and located in articles folder.

- Large scale clustering of protein sequences with FORCE -A layout based heuristic for weighted cluster editing.
  Tobias Wittkop, Jan Baumbach, Francisco P Lobo and Sven Rahmann.
  Article downloaded unsupervised learning and located in articles folder.

5 Meetings

5.1 27/09-17 - 14:15

Make subitems

1. TransClust is suggesting to parse the Gold Standard to find optimal parameters. How?
    - R package with F-measure
    - res = data.frame(protein = proteins, cluster = tclust_res$clusters[[1]])
    - best f-measure around 0.9 (threshold 23, 18?)
    - Plot with cost, threshold
    - Get test pipeline working.
    - Find the optimal clustering, with looping and increasing threshold. Measuring F-measure on each run.
    - rfactor to get unq string to numbers
    - Convert classes into numbers with rfactor and map into clusters from test result.
    - Random dataset: f-measure does not make sense here
    - test_dist = as.vector(as.dist(sim_matrix))
    - sample(x=test_dist, replace = TRUE, size = length(test_dist)))
    - Now you got your randomized similarity matrix from the initial dataset.
    - hierarchical clustering in R -> draw dendrograms
2. What is it that makes WTGP / FORCE so great ?
3. WTGP vs. FORCE ?
4. Weighted Transitivity Graph Projection vs. Weighted Graph Cluster Editing Problem ?

5.2 25/10-17 - 12:30

1. Is the F-measure calculations correct ?
    - If not, what do I need to change ?
    - When more tclust clusters point to the same gsCluster, should I find the one that gives the best score and swap ?
2. Last time we talked about plotting cost vs threshold on original data vs randomized:
    - What about the randomized similarity matrix testing ?
    - What about the plotting for cost/threshold, original data vs. randomized ?
3. hclust object ?

- merge: j is object j from data. -j is the singleton merged at this stage. Positive means it is a cluster, which was merged at an earlier stage.
- height: ?? What does this number represent At which height a merge/split was done ?
- order: Ordering for plotting, such that no branches are crossing
- labels: object name from file
- call: method call
- method: cluster method
- dist.method: distance method used

5.3 1/11-17 - 14:15

1. F-measure now finds the best candidate for a match to gsCluster, when multiple candidates. **output-1-320.txt**
2. F-measure sum all tp, fp, fn ? and then calculate F-measure(Would be for whole clustering)?
3. Status on hierarchical clustering
   - hc algorithm is "done".
   - Figuring out how to do the hc object/structure: hc$merge, hc$height, hc$order and dendrogram
   - Currently looking at ordering. Append to `order` with the proteins of the next split. If the split is a subsplit of an suborder, reorder the ordering of the subsplit.
   - Rearranging `order` will be **very** expensive
   - Currently finding all proteins for each cluster after tclust -> map into startIndex, endIndex in `order` for the given cluster.
   - Steps: order, merge, height
4. How should I handle a split at a certain threshold, when the split > 2. How should the dendrogram look like ? The dendrogam will have the split on same height -> will not look like a binary split.
5. Have not used time on randomized similarity
6. Will be done by the end of the week: hc, randomized similarity
   - What is next ? In case I have spare time before next meeting
   - Testing costs on randomSim ?
7. Do you want the code in mail ?

   F-measure (New)

   Current hc files
8. next meeting ? Monday ? Lectures 12:15-16:00

5.4 7/11-17 - 16:15

1. F-measure - corrected. Threshold 47 is the best with F-measure 0.9816, **outputFixedMeasure.txt**. Previously 40 with 0.9808. The web page has F-measure 0.9859 at 48.868 (47 is .0043 lower).

2. Random similarity

   Have made a few test outputs. Still getting costs above 1 mio.

   The costs of the randomized are very similar at a given threshold, **total.txt**.

3. > mean(as.vector(as.dist(simMatrix))) [1] 56.10247

4. Hierarchical clustering / Dendrogram

   Got a dendrogram showing, but the labels are very dense. How to make it better ?

   I have to make a normal hclust first, to get the object, and reassign variables.

5. Status on the bigger dataset ?

6. I will start looking at plotting with random similarity.

5.5 27/11-17 - 14:15

1. bigData results: Best threshold = 17.

2. Metric vs. non-metric multidimensional scaling

3. How should the points from multidimensional scaling be converted into a similarity matrix. Just calculate the distances ?

4. Binary threshold: Not quite done. In the debugging phase.

5. Bachelor thesis ?

   Meeting with Richard 29/11. Richard will come up with some suggestions.

5.6 16/12-17 - 14:00 : Skype

1. Last steps in project
   (a) Randomize every split ?(Every new cluster)
   We need our randomization distribution to work before going on
   (b) Why does the max difference tell us the optimal split?
   Forgot to ask
2. How much theory do I need to cover?
   All the theory that you have not invented yourself.
3. Should I expect answers if I write to you during the holiday?
   Just email Richard at any time
4. Report
   - Abstract
   - Introduction
   - Background (Related work)
   - Method
     - Each randomization strategy
     - Results section in the end of the whole section, for the reason of trying other strategies
   - Results
   - "Implementation"
   - Conclusion/Discussion