Clustering techniques for cancer subtype prediction

Peixuan Zhang

December 2019

1 Introduction

1.1 Clustering and Cancer subtype prediction

Cluster analysis has been quite popular as one of the unsupervised learning methods since last century. It can be simply illustrated as the exploration of inter-relationships among a collection of objects, by grouping them into several homogeneous clusters.

The goal of clustering analysis in Microarray data [5] [4] is to find patterns in genes or samples. When applied in genes, clustering techniques are useful in three main aspects: identifying groups of possible co-regulated genes, identifying spatial patterns and reducing redundancy.

Identifying the exact subtype of a cancer type is a significant step in the treatment of the disease. Using gene expression data, clustering based (unsupervised) methods could be well applied to cancer subtype prediction and discovery.

1.2 Data depth

Some brilliant researchers heuristically brought up a concept of ”data depth functions”. A depth function [6] is defined as for a distribution P in R^d, any function $D(x, P)$ which measures the degree of centrality of a point with respect to the whole data set. Higher depth corresponds to the ”center” and lower depth indicate an ”outlier”.

Depth functions [6] are usually expected to hold some mathematical or statistical properties, such as affine invariance, maximality at the center, decreasing along rays, etc.

We here introduce the simplest type of depth functions derived from Euclidean distance function

$$D(x, P) = \frac{1}{1 + O(x, P)}. \tag{1}$$

which is known as Mahalanobis depth.

In this project, the first goal is to reproduce partial analysis in [3]. Then we will try to propose some new depth-based clustering algorithms. These algorithms are then applied to the same data sets which are from NCBI Gene Expression Omnibus (GEO). The comparisons between the previous results in [3] and the new results will be given in the final part.

2 Data and Methods

2.1 Datasets

The sample data are coming from NCBI Gene Expression Omnibus (GEO). They are GSE51082, GSE57162, GSE66354, GSE85217, GSE94601.

2.2 Clustering Methods

For each dataset, a set of eleven well-known clustering algorithms including K-Means++, Density K-Means++, K-Means, Partitioning Around Medoids (PAM), Neural Gas [2], Self-Organizing Map (SOM) [1], Spectral Clustering (SC) and four Hierarchical Agglomerative Clustering (single linkage (HC-single), complete linkage (HC-comp), average linkage (HC-average) and centroid linkage (HC-centroid)).

The clustering results of the proposed depth-based algorithm were compared with those of eleven algorithms mentioned above.

Adjusted Rand Index (ARI) was employed as an external cluster validity index for assessing the quality of the clustering results.

2.3 R packages

R packages used in this project are GEOquery, Biobase, cluster, mclust, flexclust, kohonen, kernlab, pracma vegan, LICORS.

3 Statistical Analysis

3.1 Data Preprocessing

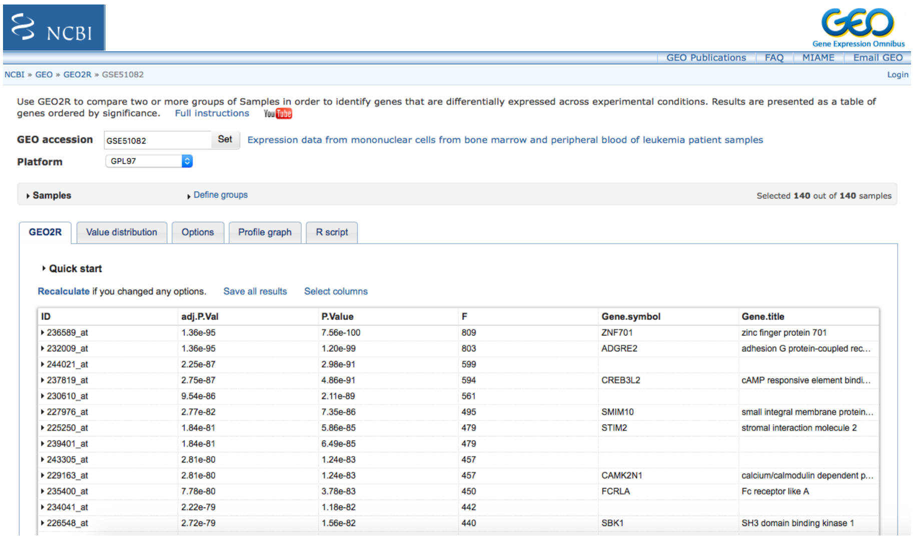
3.1.1 Data Cleaning

GSE66354 dataset has gene expression data corresponding to 149 brain tissue samples, out of which 13 are normal tissue samples taken from various parts of the brain. The rest 136 samples are used for analysis.

GSE94601 dataset contains gene expression data corresponding to 159 Lung Cancer tissue samples. The 22 samples have been found to have missing values even after gene or probe selection. We kept the rest 137 samples for our analysis.

3.1.2 Feature Selection

Feature selection was applied to all the datasets. The set of the most statistically significant 2000 genes or probes (genes or probes with the lowest 2000 adjusted p values) were considered for analysis. The GEO2R tool provided by NCBI GEO was used to compute the adjusted p values using Benjamini-Hochberg FDR method.



Use GEO2R to compare two or more groups of Samples in order to identify genes that are differentially expressed across experimental conditions. Results are presented as a table of genes ordered by significance.

GEO accession: GSE51082 — Expression data from mononuclear cells from bone marrow and peripheral blood of leukemia patient samples

Platform: GPL97

▸ Samples ▸ Define groups Selected **140** out of **140** samples

▸ Quick start

Recalculate if you changed any options. Save all results Select columns

ID	adj.P.Val	P.Value	F	Gene.symbol	Gene.title
▸ 236589_at	1.36e-95	7.56e-100	809	ZNF701	zinc finger protein 701
▸ 232009_at	1.36e-95	1.20e-99	803	ADGRE2	adhesion G protein-coupled rec...
▸ 244021_at	2.25e-87	2.98e-91	599		
▸ 237819_at	2.75e-87	4.86e-91	594	CREB3L2	cAMP responsive element bindi...
▸ 230610_at	9.54e-86	2.11e-89	561		
▸ 227976_at	2.77e-82	7.35e-86	495	SMIM10	small integral membrane protein...
▸ 225250_at	1.84e-81	5.86e-85	479	STIM2	stromal interaction molecule 2
▸ 239401_at	1.84e-81	6.49e-85	479		
▸ 243305_at	2.81e-80	1.24e-83	457		
▸ 229163_at	2.81e-80	1.24e-83	457	CAMK2N1	calcium/calmodulin dependent p...
▸ 235400_at	7.78e-80	3.78e-83	450	FCRLA	Fc receptor like A
▸ 234041_at	2.22e-79	1.18e-82	442		
▸ 226548_at	2.72e-79	1.56e-82	440	SBK1	SH3 domain binding kinase 1

Figure 1: GEO2R tool (GSE51082 as an example)

More information of the five datasets are summarized in Table 1.

GEO series	Code	Cancer type	No. of Subtypes	No. of Samples
GSE51082	GEO1	Leukemia	6	140
GSE57162	GEO2	Renal Cell Carcinoma	4	191
GSE66354	GEO3	Brain Tumor	6	136
GSE85217	GEO4	Medulloblastoma	4	763
GSE94601	GEO5	Lung Carcinoma	5	137

Table 1: Table to test captions and labels

3.2 Data Analysis

3.2.1 Reproducing the previous analysis

One of the main goals of this project is to reproduce the analysis in [3]. Following the setup of the comparative study in the paper, we applied those eleven clustering algorithms to the five datasets.

To get the results consistent with those in [3] K-Means++, K-Means, SOM, SC and Neural Gas algorithms were run 100 times on these five datasets and the average ARI scores obtained are shown for the comparisons.

3.2.2 New Depth-based Clustering

Depth-based clustering for cancer subtype prediction will work in the following way:

Assignment Step Assign each data point to the cluster where the data point has the deepest depth with respect to cluster's center (the center has the depth of 1).

Update Step Calculate the new centroids of the observations in the new clusters.

To be more specific, the traditional clustering methods usually use "distance" as the measurement in the assignment step. However, depth based clustering methods employ "depth" which makes the methods more robust and hold more great mathematical properties.

Here we consider about the depth function that we previously introduced to construct depth based clustering:

$$D(x,P) = \frac{1}{1+O(x,P)}. \tag{2}$$

where $O(x,P)$ is outlying function which measures the degree of outlyingness with respect to a specific "center".

We here use Mahalanobis depth in the depth based clustering. Then the outlyingness function is:

$$O(X_i,U_i) = (X_i - U_i)\Sigma^{-1}(X_i - U_i)^T \tag{3}$$

Where Σ is the covariance matrix for the data matrix X. We can use robust covariance matrix to make this method more robust.

Then we applied this new clustering method to the same five datasets and summarized the corresponding ARI scores.

4 Results

Here five line charts are displayed in Figure 2 in order to evaluate the extent to which our results are similar to those in [3].

For each dataset, there are some deviations between the ARI values in [3] and our reproduced results. The results of Kmeans++ and SOM are the most biased.

However, in general, these deviations are not significant, which means our results are pretty similar to those in the original paper. The deviations are probably due to the fact that some algorithms such as K-Means++, K-Means, SOM, SC and Neural Gas algorithms may produce different results on different runs for the same dataset.

Although the new proposed depth clustering is not the best among these twelve clustering algorithms, it outperforms most of the clustering methods.

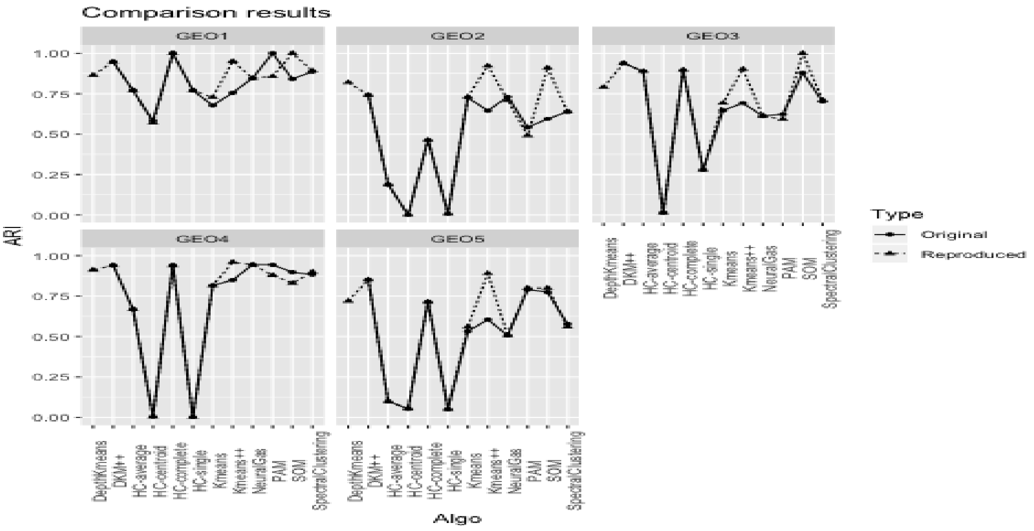


Figure 2: Comparisons of clustering algorithms

5 Summary and Discussion

In this paper, we reproduced the analysis published in the [3]. Furthermore, we presented the theoretical feasibility of depth based clustering method and proposed a new depth-based Kmeans method. All comparisons results are shown in the figure 2. More details about our results will be given in the table 2 in the Appendix. In summary, we showed the critical role of clustering techniques in cancer subtype prediction areas.

In this project, depth convex clustering only uses Mahalanobis depth function for simple interpretation and computation. However, there are more depth functions which can be basis of depth based convex clustering methods. Besides, the great statistical properties are not proved in this paper such as the robustness of dept based clustering. These untouched considerations are worthy of further investigation.

Appendix

Summary Table

Algorithms	GEO1	GEO2	GEO3	GEO4	GEO5
DKM++	0.95 (0.95)	0.74 (0.74)	0.94 (0.94)	0.94 (0.94)	0.85 (0.85)
Kmeans++	0.95 (0.76)	0.92 (0.65)	0.90 (0.69)	0.96 (0.85)	0.89 (0.61)
Kmeans	0.73 (0.68)	0.73 (0.73)	0.69 (0.65)	0.82 (0.82)	0.56 (0.53)
SOM	1 (0.84)	0.91 (0.59)	1 (0.88)	0.83 (0.90)	0.80 (0.78)
NeuralGas	0.85 (0.85)	0.73 (0.73)	0.61 (0.61)	0.95 (0.95)	0.51 (0.51)
HC-single	0.77 (0.77)	0.007 (0.007)	0.28 (0.28)	0.002 (0.002)	0.051 (0.051)
HC-centroid	0.57 (0.58)	0 (0)	0.013 (0.014)	0.004 (0.004)	0.054 (0.054)
HC-complete	1 (1)	0.46 (0.046)	0.90 (0.90)	0.94 (0.94)	0.72 (0.72)
HC-average	0.77 (0.77)	0.19 (0.19)	0.89 (0.89)	0.67 (0.67)	0.1 (0.1)
PAM	0.86 (1)	0.49 (0.54)	0.59 (0.62)	0.88 (0.94)	0.80 (0.79)
SpectralClustering	0.89 (0.89)	0.64 (0.63)	0.71 (0.70)	0.90 (0.88)	0.56 (0.58)
DepthKmeans	0.86 (NA)	0.82 (NA)	0.79 (NA)	0.91 (NA)	0.72 (NA)

Table 2: Summary table of comparison results

*The values in the parentheses are obtained from the original paper.

R script (GSE51082 as an example)

The corresponding R codes are available on: `https://github.com/PeixuanZ/PUBH7445project.git`

References

[1] Teuvo Kohonen, Jussi Hynninen, Jari Kangas, and Jorma Laaksonen. Som pak: The self-organizing map program package. *Report A31, Helsinki University of Technology, Laboratory of Computer and Information Science*, 1(1-27):39–40, 1996.

[2] Thomas Martinetz, Klaus Schulten, et al. A" neural-gas" network learns topologies. 1991.

[3] N Nidheesh, KA Abdul Nazeer, and PM Ameer. An enhanced deterministic k-means clustering algorithm for cancer subtype prediction from gene expression data. *Computers in biology and medicine*, 91:213–221, 2017.

[4] Ainhoa Perez-Diez, A Morgun, and N Shulzhenko. Microarrays for cancer diagnosis and classification. *Madame Curie Bioscience Database [Internet]. Landes Bioscience, Austin (TX)(2000–2013)*, 2000.

[5] Yixin Wang, Jan GM Klijn, Yi Zhang, Anieta M Sieuwerts, Maxime P Look, Fei Yang, Dmitri Talantov, Mieke Timmermans, Marion E Meijer-van Gelder, Jack Yu, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet*, 365(9460):671–679, 2005.

[6] Yijun Zuo and Robert Serfling. General notions of statistical depth function. *Annals of statistics*, pages 461–482, 2000.