UNIVERSITY OF THESSALY
SCHOOL OF MEDICINE
LABORATORY OF
BIOMATHEMATICS

MASTER PROGRAM IN
"Research Methodology in Biomedicine,
Biostatistics and Clinical Bioinformatics"

Develop Software in Python for Performing Nearest Neighbor and Furthest Neighbor Analysis Using the Squared Euclidean Distance

By

Michail Gkorgkolis

June 2017

EVALUATION COMMITTEE

Kowald Axel, Professor, Supervisor
Zintzaras Elias, Professor
Raxiotis Georgios, Professor

Institutional Repository - Library & Information Centre - University of Thessaly
21/09/2024 08:23:17 EEST - 18.117.7.43

Institutional Repository - Library & Information Centre - University of Thessaly
21/09/2024 08:23:17 EEST - 18.117.7.43

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΙΑΤΡΙΚΗΣ
ΕΡΓΑΣΤΗΡΙΟ
ΒΙΟΜΑΘΗΜΑΤΙΚΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«Μεθοδολογία Βιοϊατρικής Έρευνας,
Βιοστατιστική και Κλινική Βιοπληροφορική»

Ανέπτυξε ένα πρόγραμμα στην python χρησιμοποιώντας τη μέθοδο του κοντινότερου γείτονα και του μακρύτερου γείτονα μέσω της τετραγωνικής Ευκλείδειας απόστασης.

Από τον

Μιχαήλ Γκοργκόλη

Ιούνιος 2017

ΤΡΙΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

Kowald Axel, Καθηγητής, Επιβλέπων
Ζιντζαράς Ηλίας, Καθηγητής
Ραχιώτης Γεώργιος, Καθηγητής

Institutional Repository - Library & Information Centre - University of Thessaly
21/09/2024 08:23:17 EEST - 18.117.7.43

Institutional Repository - Library & Information Centre - University of Thessaly
21/09/2024 08:23:17 EEST - 18.117.7.43

Contents

Institutional Repository - Library & Information Centre - University of Thessaly
21/09/2024 08:23:17 EEST - 18.117.7.43

Summary

In this article, we explore two common clustering algorithms: nearest neighbor algorithm and furthest neighbor algorithm. We apply these two algorithms in the area of pelvic ring injuries diagnose. We implement these algorithms using Python SciPy library and perform our experiments.

Introduction

In data mining and statistics, cluster analysis or clustering is the task of grouping a set of objects in a way that the objects in the same cluster are similar than these in the other clusters [1]. Hierarchical clustering, which is also called hierarchical cluster analysis or HCA is a method of cluster analysis which seeks to build a hierarchy of clusters. Clustering can be well-applied in medical area [2] . For example, we can explore groups of patients with similar morphometric profile in the sacrum. We guess that the identification of groups of different patients may direct us in choosing the appropriate surgical technique in patients with pelvic ring injuries. However, we can use different methods to cluster different patients into different groups. Nearest neighbor analysis and furthest neighbor analysis are two common algorithms for cluster analysis. Nearest neighbor analysis, also called single linkage, defines the distance between two clusters as the distance of the two closest objects. Furthest neighbor analysis, also called complete linkage, defines the distance between two clusters as the distance of the two furthest objects [3]. In this article, we use Python to develop software to implement these two algorithms and compare them.

Institutional Repository - Library & Information Centre - University of Thessaly
21/09/2024 08:23:17 EEST - 18.117.7.43

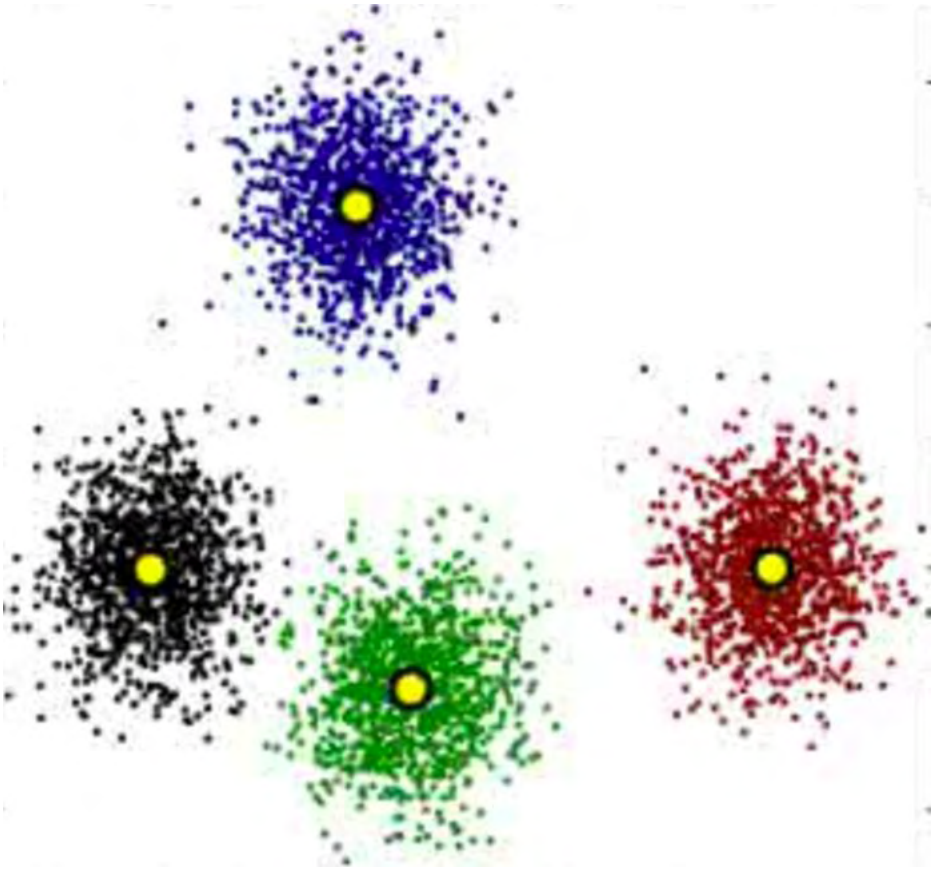

Figure 1

Methods

In this section, we introduce the methods we used and the experiments we designed in details.

Institutional Repository - Library & Information Centre - University of Thessaly
21/09/2024 08:23:17 EEST - 18.117.7.43

Cluster Analysis

Cluster analysis is the task of grouping a set of objects is such a way that objects in the same group, which is also called cluster, are more similar than the objects in the others groups. It is one of the main tasks of data mining and machine learning and it is used in many areas such as image analysis, information retrieval, pattern recognition and bioinformatics. As shown in Figure 1, there are four clusters in this figure.

Cluster is just a common denominator for a group of objects, so cluster can be defined using many different criteria in different circumstances. We call them models: connectivity models build clusters based on distance connectivity; distribution models build clusters using statistical distributions; density models define clusters as dense regions in the whole data space. Cluster algorithms are algorithms that calculating clusters. Each cluster model has its corresponding cluster algorithm.

In this article, we mainly explore the application of cluster analysis in the area of pelvic ring injuries. Morphometric profile in the sacrum influence the surgical technique choice badly. If we cluster the patients according their morphometric profile, we can find the most appropriate surgical technique for them, which could accelerate the process of diagnose a lot. To cluster different patients, we should firstly find the appropriate metrics to measure the similarity between different pelvic rings.

Institutional Repository - Library & Information Centre - University of Thessaly
21/09/2024 08:23:17 EEST - 18.117.7.43

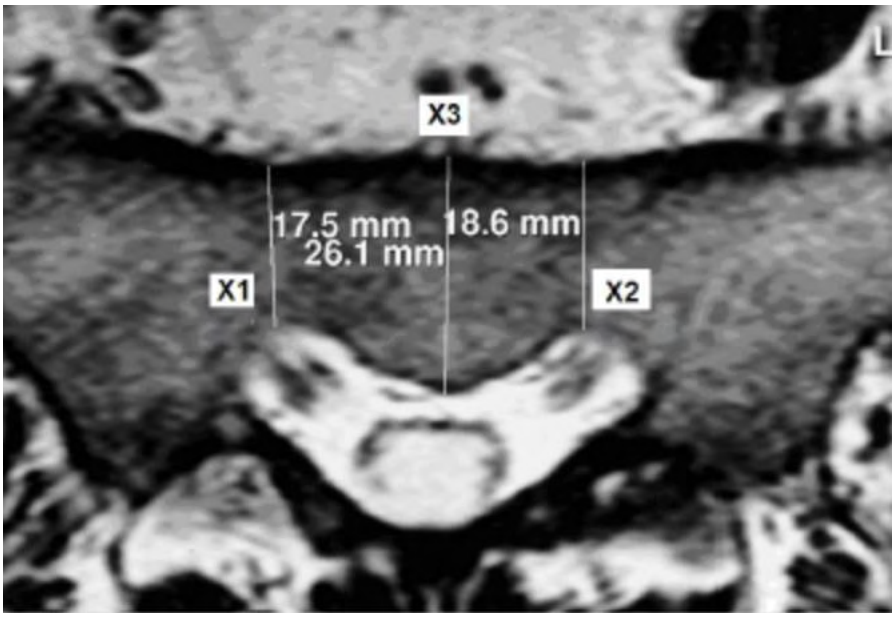

Figure 2

As shown in Figure 1, X1 represents the diameter of right pedicle, X2 represents the diameter of left pedicle and X3 represents the diameter of the vertebra body. These three factors are the main morphological variables of the sacrum. We use a vector of these three factors to represent the main feature of the sacrum of each patient.

Nearest Neighbor Analysis

The distance between two objects can be easily defined using given measurement. However, once several objects are linked together, how could we determine the distance two clusters? That is to say, we also need to define how two measure the distance between two clusters. Nearest neighbor is one the measurements for the distance of two clusters.

As describe above, in nearest neighbor analysis, we define the distance between two clusters as the distance between two nearest objects from these two clusters. That is:

Institutional Repository - Library & Information Centre - University of Thessaly
21/09/2024 08:23:17 EEST - 18.117.7.43

$$D(x, y) = \min_{x \in X, y \in Y} d(x, y)$$

Where

- d(x, y) is the distance between elements $x \in X \; and \; y \in Y$
- X and Y are two clusters

Furthest Neighbor Analysis

Opposite to the nearest neighbor analysis, in furthest neighbor analysis, the distance between two clusters are represented by the distance of the two furthest objects from these two clusters. That is:

$$D(x, y) = \max_{x \in X, y \in Y} d(x, y)$$

Where

- d(x, y) is the distance between elements $x \in X \; and \; y \in Y$
- X and Y are two clusters

Furthest neighbor analysis is also called complete-linkage clustering. It avoids a drawback of nearest neighbor analysis. The drawback is the chaining phenomenon: if most of the objects from two clusters are far away from each other, but accidentally some objects are near to each other, then these two clusters might be force to be linked. Furthest neighbor analysis can lead to compact clusters with approximately equal diameters.

Institutional Repository - Library & Information Centre - University of Thessaly
21/09/2024 08:23:17 EEST - 18.117.7.43

Python and SciKit-learn

Python is a widely used programming language created by Guido van Rossum and first released in 1991[4]. It provides a syntax which can express concepts with fewer lines of code that language such as C, C++ and Java. It also improves the readability of program. Due to its simplicity and readability, it is widely used in data mining, machine learning and many other areas.
As shown in Figure 3, Python is the fourth most popular language according to the newest TIOBE Index for June, 2017[5].

Jun 2017	Jun 2016	Change	Programming Language	Ratings	Change
1	1		Java	14.493%	-6.30%
2	2		C	6.848%	-5.53%
3	3		C++	5.723%	-0.48%
4	4		Python	4.333%	+0.43%
5	5		C#	3.530%	-0.26%
6	9	^	Visual Basic .NET	3.111%	+0.76%
7	7		JavaScript	3.025%	+0.44%
8	6	v	PHP	2.774%	-0.45%
9	8	v	Perl	2.309%	-0.09%
10	12	^	Assembly language	2.252%	+0.13%
11	10	v	Ruby	2.222%	-0.11%
12	14	^	Swift	2.209%	+0.38%
13	13		Delphi/Object Pascal	2.158%	+0.22%
14	16	^	R	2.150%	+0.61%
15	48	^^	Go	2.044%	+1.83%
16	11	vv	Visual Basic	2.011%	-0.24%
17	17		MATLAB	1.996%	+0.55%
18	15	v	Objective-C	1.957%	+0.25%
19	22	^	Scratch	1.710%	+0.76%
20	18	v	PL/SQL	1.566%	+0.22%

Figure 3

SciPy is an open source Python library used for scientific computing and technical computing. It contains modules for optimization, image

Institutional Repository - Library & Information Centre - University of Thessaly
21/09/2024 08:23:17 EEST - 18.117.7.43

processing and especially clustering. It has become one of the most popular libraries for researchers and data analysts.

SciPy is built on Numpy array object and is one important part of Numpy technical stack. There are similar libraries such as Matlab, Scilab and R, but due to the simpler and more readable code of SciPy, we choose SciPy to implement our experiments.

Experiments

To test and verify our idea, we design an experiment with the patient data in Figure 4.

Patient	X1	X2	X3
1	x11= 22	x12=21	x13=28
2	x21= 20	x22=22	x23=30
3	x31= 14	x32=15	x33=21
4	x41= 1 6	x42=16	x43=24
5	x51= 18	x52=19	x53=26

Figure 4

We have the pedicle data of five patients here. X1 represents the diameter of right pedicle, X2 represents the diameter of left pedicle and X3 represents the diameter of the vertebra body. Each patient object is represented by the three factors and becomes one element in our clustering algorithm. We try out the nearest neighbor clustering algorithm and furthest neighbor clustering algorithm based on our experiment data.

Institutional Repository - Library & Information Centre - University of Thessaly
21/09/2024 08:23:17 EEST - 18.117.7.43

Implementation

```
import numpy as np
from scipy.cluster.hierarchy import dendrogram, linkage
from scipy.spatial.distance import squareform

import matplotlib.pyplot as plt


X = np.array([[22.0, 21.0, 28.0], [20.0, 22.0, 30.0],
    [14.0, 15.0, 21.0], [16.0, 16.0, 24.0], [18.0, 19.0, 26.0]])


Z = linkage(X, method='single', metric='euclidean')


# calculate full dendrogram
plt.figure(figsize=(25, 10))
plt.title('Simple Linkage\n Euclidean distances')
plt.xlabel('')
plt.ylabel('Distance')
dendrogram(Z)
plt.show()
```

Figure 5

As shown in Figure 5, we implemented the algorithm using SciPy and finished the experiment using our patient data.

Institutional Repository - Library & Information Centre - University of Thessaly
21/09/2024 08:23:17 EEST - 18.117.7.43

Results

The process of clustering is normally hierarchical, so the result of clustering is usually represented in a dendrogram. The dendrogram of our experiment is shown in Figure 6.

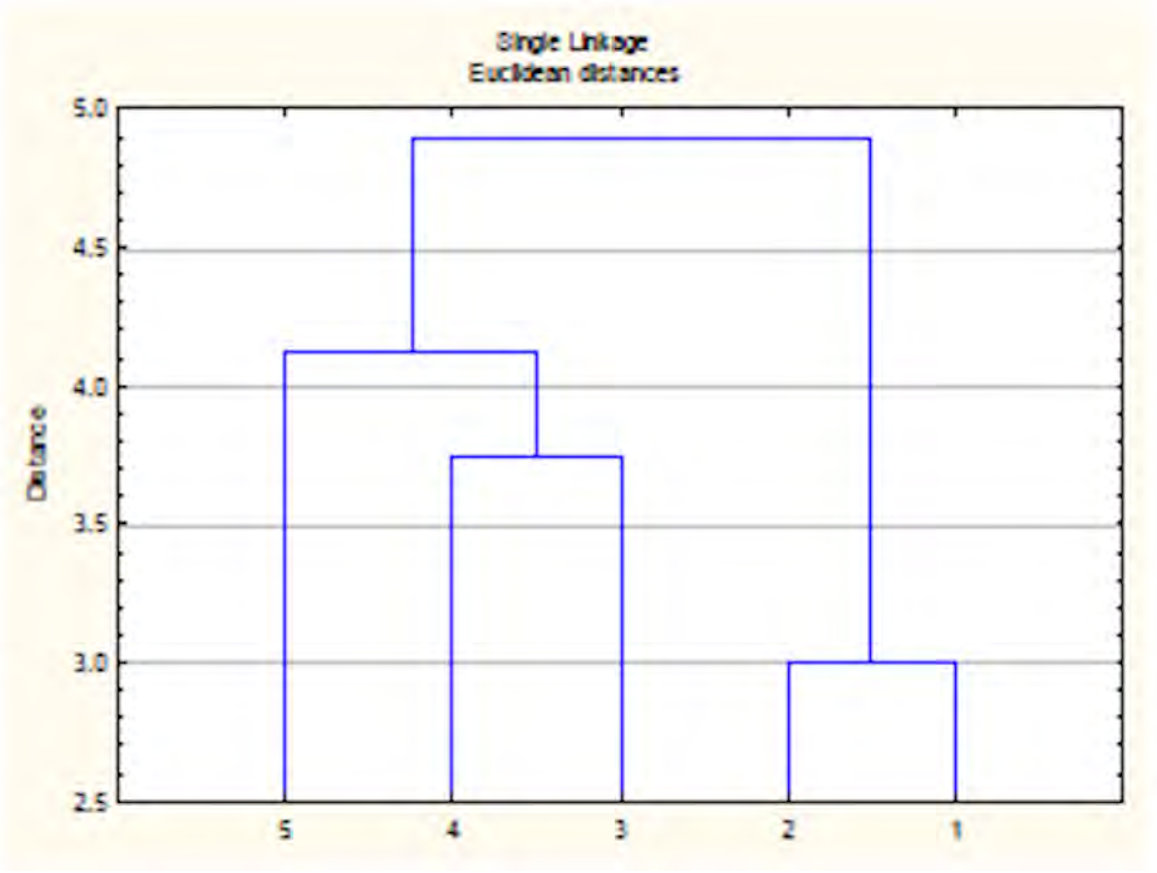

Figure 6

Institutional Repository - Library & Information Centre - University of Thessaly
21/09/2024 08:23:17 EEST - 18.117.7.43

Conclusion

As show in Figure 6, Patient 1 and Patient 2 are most similar to each other and they tend to use the similar surgical technique. Patient 3, Patient 4 and Patient 5 are more similar to each other and they tend to use similar surgical technique.

As a conclusion, we can use clustering algorithm to help direct us in choosing the appropriate surgical technique for different patients. However, because we lack enough data samples for the clustering algorithm, we cannot compare the nearest neighbor algorithm and furthest neighbor algorithm well.

Institutional Repository - Library & Information Centre - University of Thessaly
21/09/2024 08:23:17 EEST - 18.117.7.43

References

[1] Everitt, Brian (2011). *Cluster analysis*. Chichester, West Sussex, U.K: Wiley. ISBN 9780470749913.
[2] Rokach, Lior, and Oded Maimon. "Clustering methods." Data mining and knowledge discovery handbook. Springer US, 2005. 321-352.
[3] *Ward, Joe H. (1963). "Hierarchical Grouping to Optimize an Objective Function". Journal of the American Statistical Association.* ***58*** *(301): 236–244. JSTOR 2282967. MR 0148188. doi:10.2307/2282967*
[4] https://www.python.org
[5] https://www.tiobe.com/tiobe-index/
[6] https://www.scipy.org

Institutional Repository - Library & Information Centre - University of Thessaly
21/09/2024 08:23:17 EEST - 18.117.7.43