Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Σχολή θετικών Επιστημών
Τμήμα Πληροφορικής

Έκθεση Αποτελεσμάτων στην εργασία με θέμα Dimensionality Reduction and Spectral Clustering

Βασίλειος Ασημακόπουλος

15 Ιανουαρίου 2023

Περιεχόμενα

Κατάλογος σχημάτων

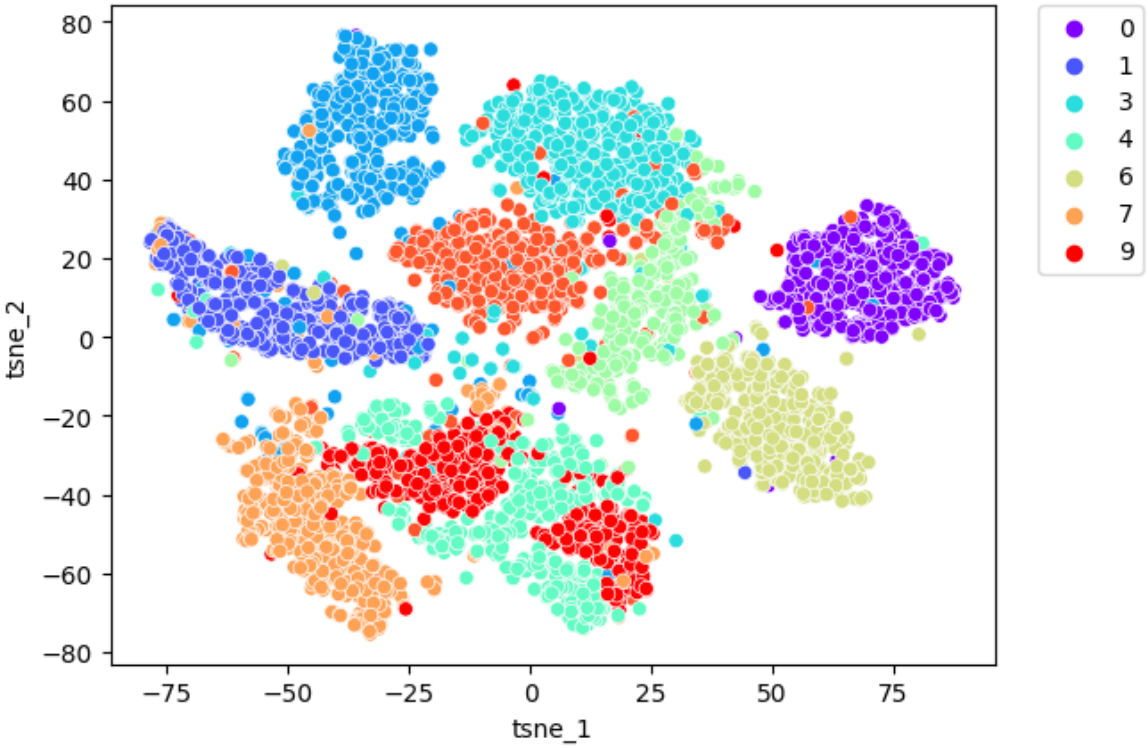
Κεφάλαιο 1

MNIST DIGIT

Σε αυτό το κεφάλαιο θα συγκριθούν αρχικά οι μέθοδοι που χρησιμοποιήθηκαν για classification των δεδομένων, μετά από μείωση των διαστάσεων μέσω t - SNE, και Isomap, και στην συνέχεια θα συγκριθούν οι δύο μέθοδοι μείωσης των διαστάσεω μεταξύ τους.

1.1 T-SNE

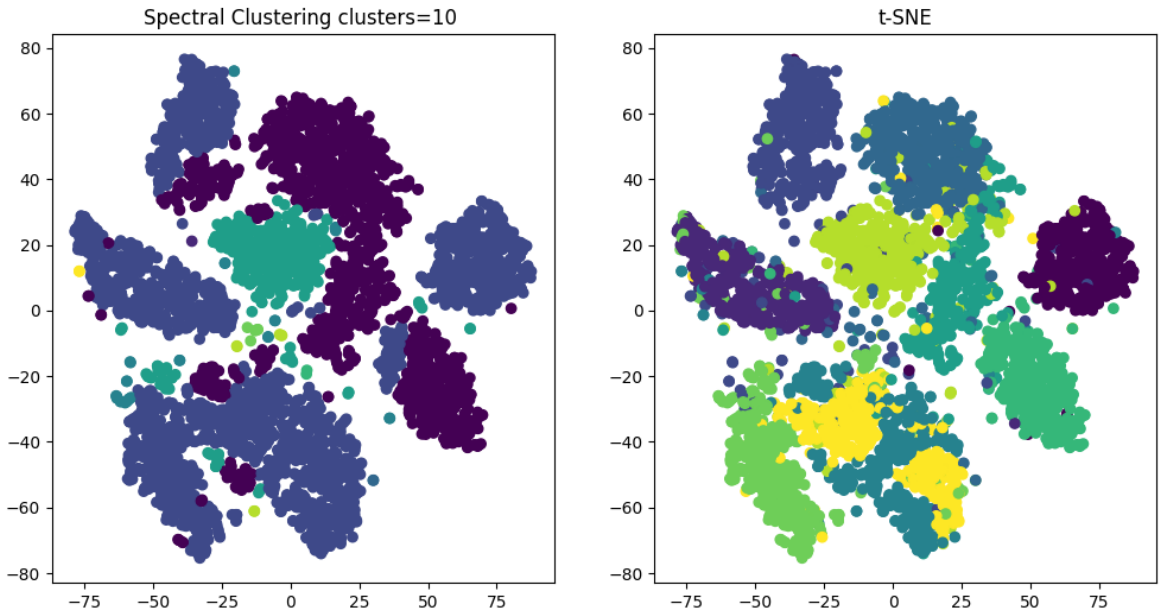
Το t - SNE διαχώρισε τα δεδομένα σχετικά καλά όπως φαίνεται και στην εικόνα 1.1



Σχήμα 1.1: Διάγραμμα απεικονισμού κλάσεων

1.1.1 Spectral clustering with rbf

Στο συγκεκριμένο αλγόριθμο ο αριθμός των clusters να ήταν δέκα και το rbf επιλέχθηκε ώς δείκτης ομοιότητας μεταξύ των σημείων των δεδομένων. Το αποτέλεσμα αυτού του αλγορίθμου φάνηκε σε δύο μετρικές, στην μετρική της ομοιογένειας και της σιλουέτας. Το Silhouette score: -0.405 και το Homogeneity score: 0.246, τα οποιά είναι πολύ χαμηλά, κάτι που φαίνεται και στο διάγραμμα 1.2.



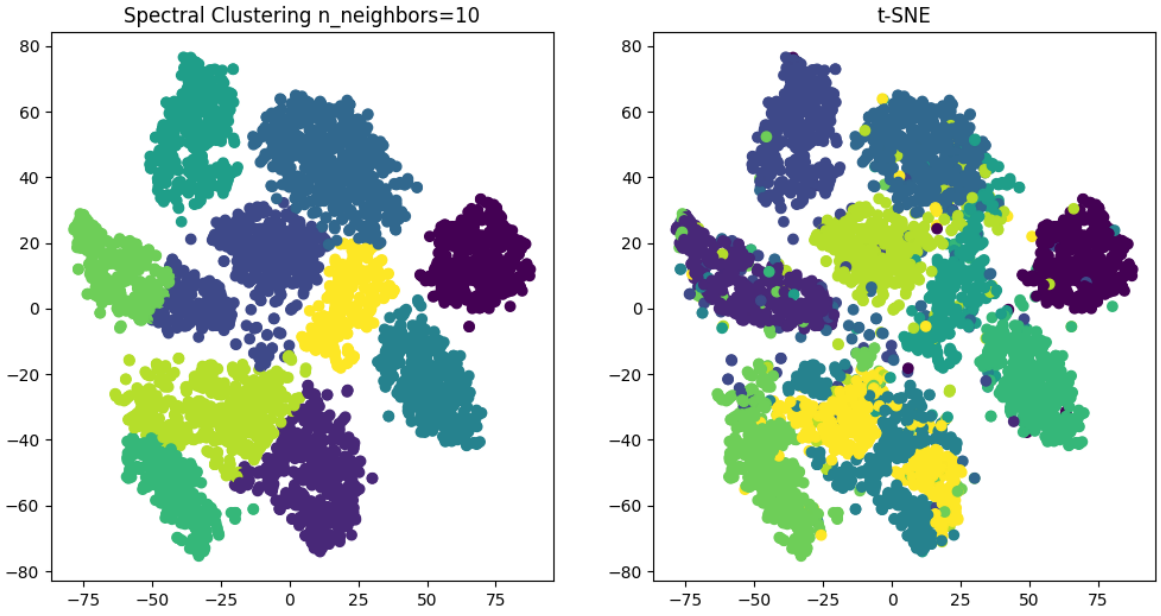
Σχήμα 1.2: Διαγράμματα σύγκρισεις απεικόνισης κλάσεων

1.1.2 Spectral clustering with nearest neighbors

Στην συνέχεια αναπτύχθηκε ο ίδιος αλγόριθμος με διαφορετικό δείκτη ομοιότητας (affinity), ο οποίος είναι ο nearest neighbors. Επιλέχθηκε να αλλάζει ο αριθμός των cluster και ο αριθμός των γειτόνων απο το 5 μέχρι το 40 με βήμα 5. Τα αποτελέσματα φάνηκαν στις δύο μετρικές στον πίνακα 1.1. Το καλύτερο αποτέλεσμα για την πρώτη μετρική δίνεται όταν οι γείτονες και οι cluster είναι 15 1.3, και για την δεύτερη όταν είναι 35 1.4.

Πίνακας 1.1: Μετρικές για διαφορετικές τιμές cluster/γειτόνων

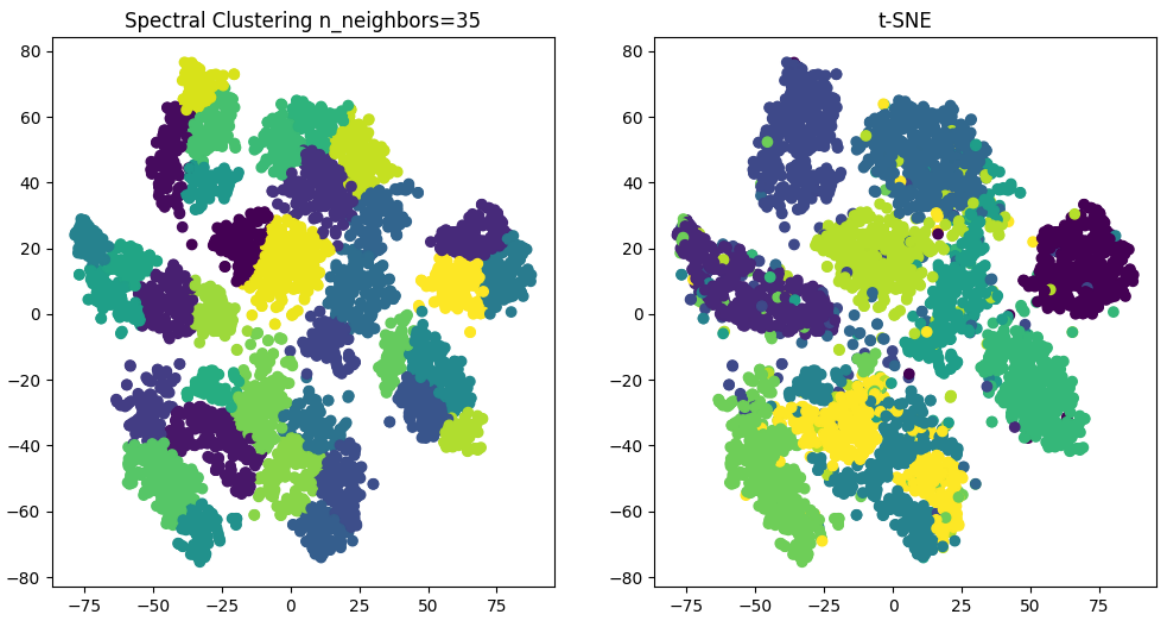
Τιμή cluster/γειτόνων	Silhouette score	Homogeneity score
5	-0.158	0.192
10	0.443	0.728
15	0.428	0.785
20	0.389	0.789
25	0.369	0.793
30	0.354	0.799
35	0.360	0.820



Σχήμα 1.3: Διαγράμματα σύγκρισεις απεικόνισης κλάσεων

1.1.3 Kmeans

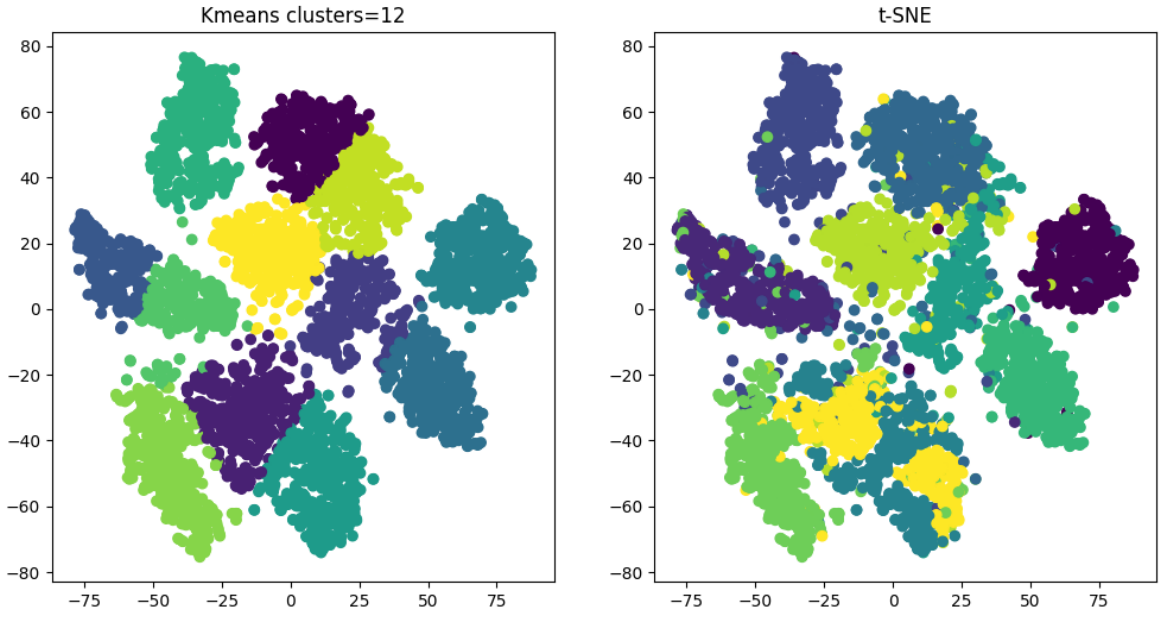
Ακόμη χρησιμοποιήθηκε ο αλγόριθμος Kmeans, στον οποίο ελέγχθηκαν οι ίδιες μετρικές με πρίν για διαφορετικό αριθμο clusters, από 2 μέχρι 28 με βήμα 5. Τα αποτελέσματα φαίνονται στον πίνακα 1.2 παρακάτω. Το καλύτερο αποτέλεσμα για την πρώτη μετρική δίνεται όταν οι cluster είναι 12 1.5, και για την δεύτερη όταν είναι 27 1.6.



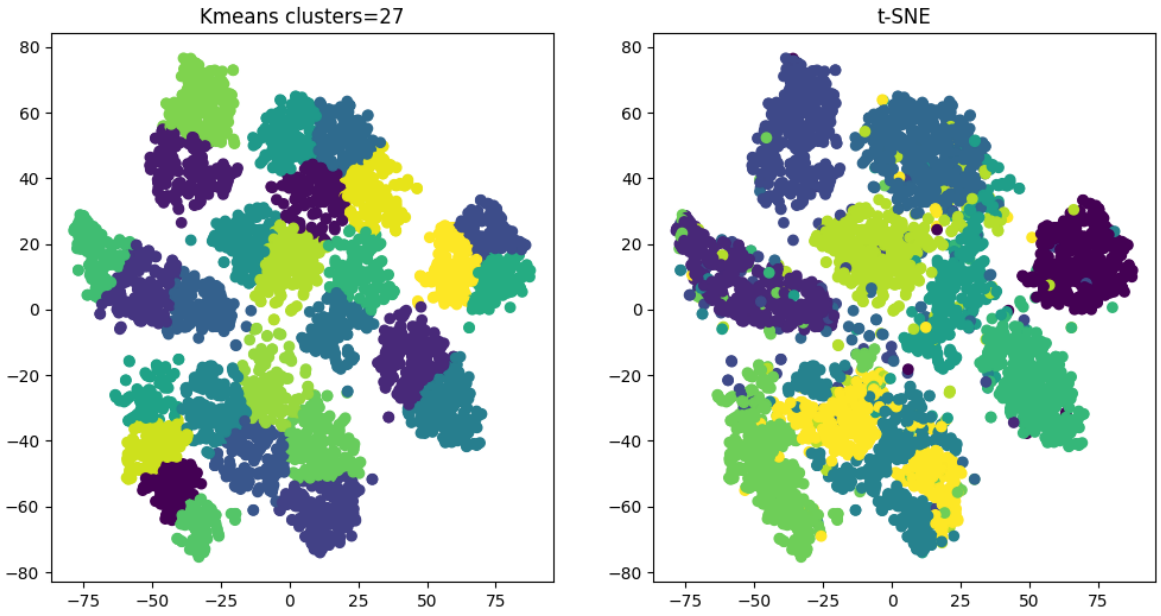
Σχήμα 1.4: Διαγράμματα σύγκρισεις απεικόνισης κλάσεων

Πίνακας 1.2: Μετρικές για διαφορετικές τιμές cluster

Τιμή cluster	Silhouette score	Homogeneity score
2	0.378	0.244
7	0.436	0.616
12	0.457	0.766
17	0.411	0.775
22	0.405	0.796
27	0.401	0.797



Σχήμα 1.5: Διαγράμματα σύγκρισεις απεικόνισης κλάσεων



Σχήμα 1.6: Διαγράμματα σύγκρισεις απεικόνισης κλάσεων

1.1.4 Σύγκριση μεταξύ αλγορίθμων

Τα αποτελεσμάτα των μετρικών είναι παρόμοια στο spectral clustering με nearest neighbors με τα αντίστοιχα του αλγορίθμου kmeans, αλλά για rbf δείκτη , τα σκόρ είναι πολύ χαμηλά.

1.2 Isomap

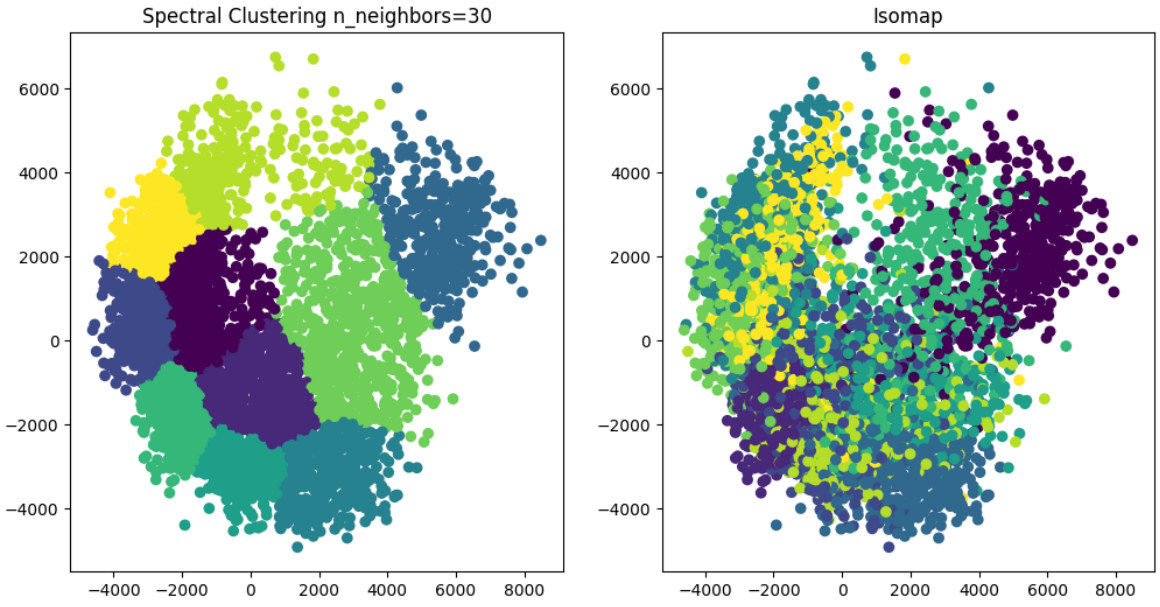
Στην συνέχεια αναπτύχθηκε ο αλγόριθμος Isomap για την μείωση των διαστάσεων σε δύο στο dataset. Ελέγχθηκαν διαφορετικές τιμές components στο isomap και όλες βγάλανε παρόμοια αποτελέσματα στις απεικονίσεις των διαγραμμάτων, για αυτό τον λόγο η διερεύνηση των αλγορίθμων Spectral clustering και kmeans έγινε με components = 10 στο Isomap

1.2.1 Spectral clustering with nearest neighbors

Στην συνέχεια αναπτύχθηκε ο ίδιος αλγόριθμος με διαφορετικό δείκτη ομοιότητας (affinity), ο οποίος είναι ο nearest neighbors. Επιλέχθηκε να αλλάζει ο αριθμός των cluster και ο αριθμός των γειτόνων απο το 5 μέχρι το 40 με βήμα 5. Τα αποτελέσματα φάνηκαν στις δύο μετρικές στον πίνακα 1.3. Το καλύτερο αποτέλεσμα και για τις δύο μετρικές δίνεται όταν οι clusters και οι γείτονες είναι 30 1.7.

Πίνακας 1.3: Μετρικές για διαφορετικές τιμές cluster/γειτόνων

Τιμή cluster/γειτόνων	Silhouette score	Homogeneity score
5	0.228	0.390
10	0.308	0.415
15	0.308	0.410
20	0.322	0.420
25	0.322	0.421
30	0.323	0.421
35	0.322	0.421



Σχήμα 1.7: Διαγράμματα σύγκρισεις απεικόνισης κλάσεων

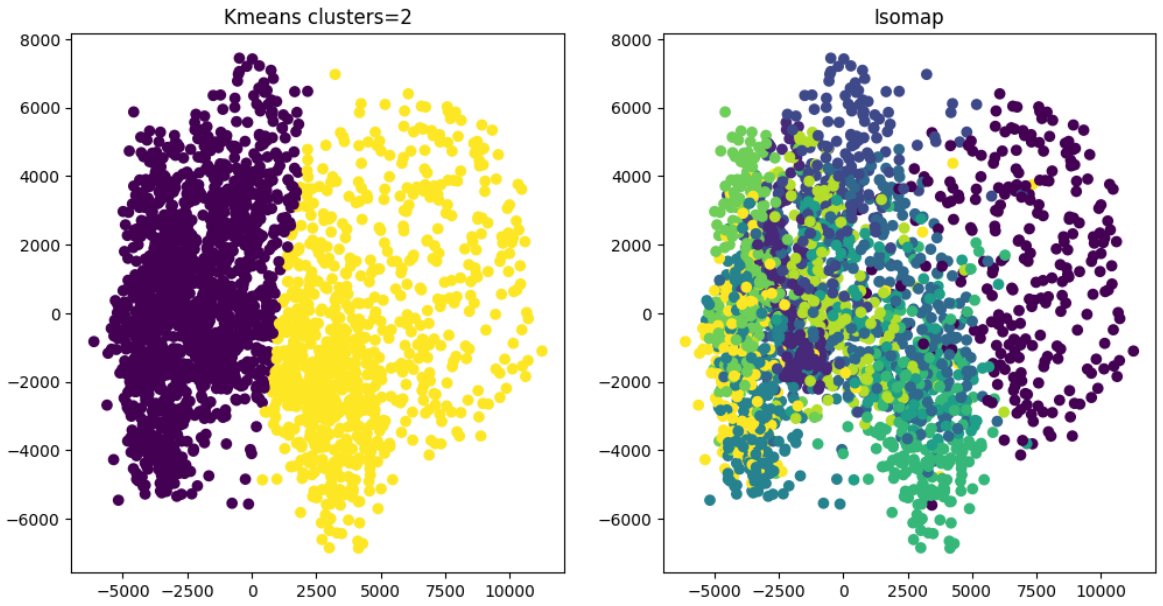
1.2.2 Kmeans

Ακόμη χρησιμοποιήθηκε ο αλγόριθμος Kmeans, στον οποίο ελέγχθηκαν οι ίδιες μετρικές με πρίν για διαφορετικό αριθμο clusters, από 2 μέχρι 28 με βήμα 5. Τα αποτελέσματα

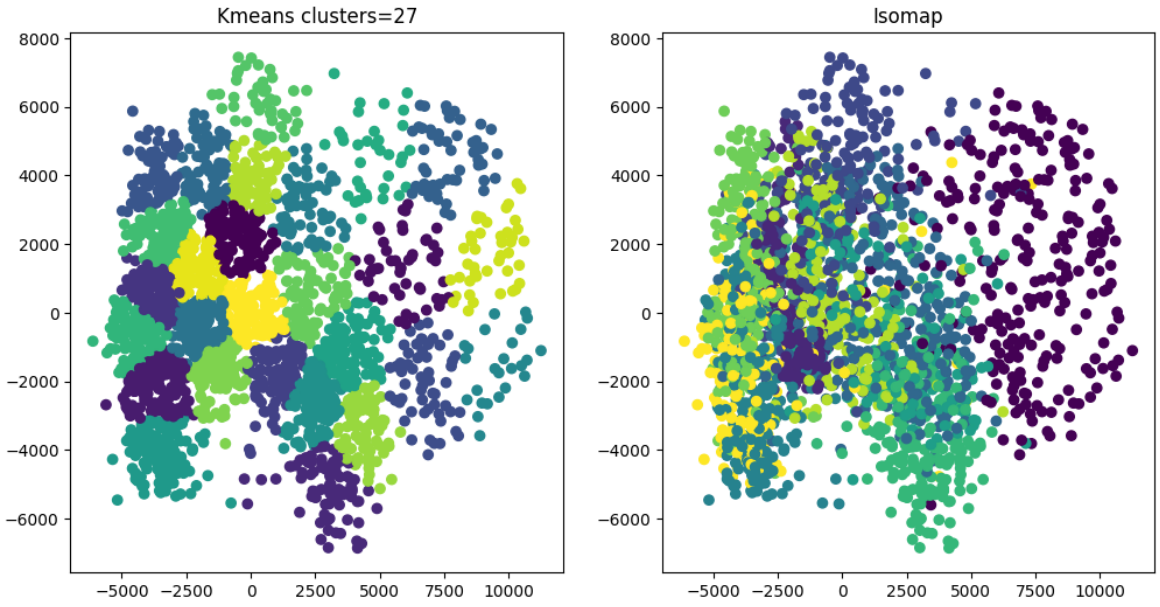
φαίνονται στον πίνακα 1.4 παρακάτω. Το καλύτερο αποτέλεσμα για την πρώτη μετρική δίνεται όταν οι cluster είναι 2 1.8, και για την δεύτερη όταν είναι 27 1.9.

Πίνακας 1.4: Μετρικές για διαφορετικές τιμές cluster

Τιμή cluster	Silhouette score	Homogeneity score
2	0.441	0.173
7	0.369	0.374
12	0.380	0.432
17	0.372	0.451
22	0.352	0.472
27	0.341	0.479



Σχήμα 1.8: Διαγράμματα σύγκρισεις απεικόνισης κλάσεων



Σχήμα 1.9: Διαγράμματα σύγκρισεις απεικόνισης κλάσεων

1.2.3 Σύγκριση μεταξύ αλγορίθμων

Τα αποτελεσμάτα των μετρικών είναι παρόμοια στο spectral clustering με nearest neighbors με τα αντίστοιχα του αλγορίθμου kmeans. Στο πρώτο η διαφορές μεταξύ των βημάτων είναι μικρές στις μετρικές.

1.3 Σύγκριση μεταξύ t-SNE και Isomap

Παρατηρείται μεγάλη διαφορά στα αποτελέσματα των μετρικών μεταξύ των δύο αλγορίθμων μείωσης της διάστασης των δεδομένων. Συγκεκριμένα φαινεται ότι η μέθοδος t-SNE ταιριάζει περισσότερο στην MNIST digits, αφού οι clustering αλγόριθμοι δουλεύουν πιο αποτελσματικά στον διαχωρισμό των κλάσεων. Επίσης παρατηρείται ότι κανένας απο των δύο αλγορίθμους μείωσης των διαστάσεων, δεν έβγαλε ένα πολυ ικανοποιητικό αποτέλεσμα. Αυτό ίσως να φταίει στο γεγονός ότι συνήθως όταν αναπτύσσονται αλγόριθμοι σαν τον t-SNE , χρησιμοποιούνται πρίν απο αυτούς και άλλες μέθοδοι μείωσης των διαστάσεων όπως η PCA (βλέπε προηγούμενη εργασία).

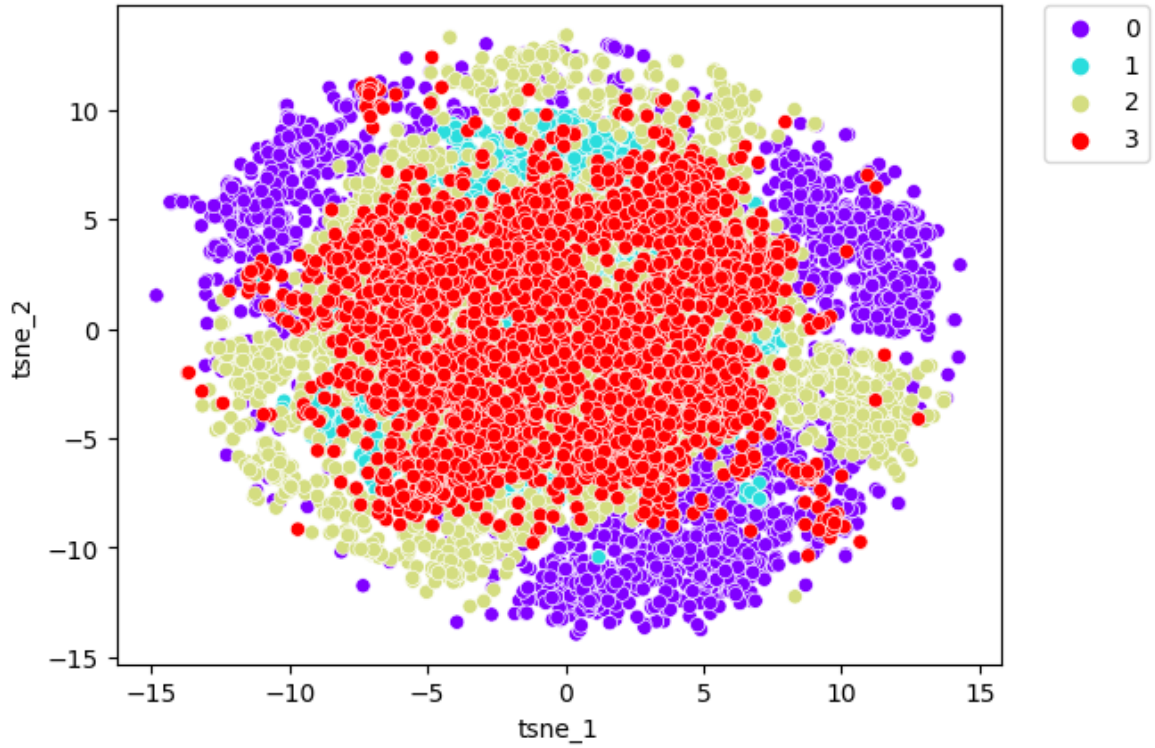
Κεφάλαιο 2

Muscle Activity Dataset

Σε αυτό το κεφάλαιο θα συγκριθούν αρχικά οι μέθοδοι που χρησιμοποιήθηκαν για διαχωρισμό των κλάσεων των δεδομένων, μετά από μείωση των διαστάσεων μέσω t - SNE, και Isomap, και στην συνέχεια θα συγκριθούν οι δύο μέθοδοι μείωσης των διαστάσεω μεταξύ τους.

2.1 T-SNE

Το t - SNE διαχώρισε τα δεδομένα, όπως φαίνεται και στην εικόνα 2.1.



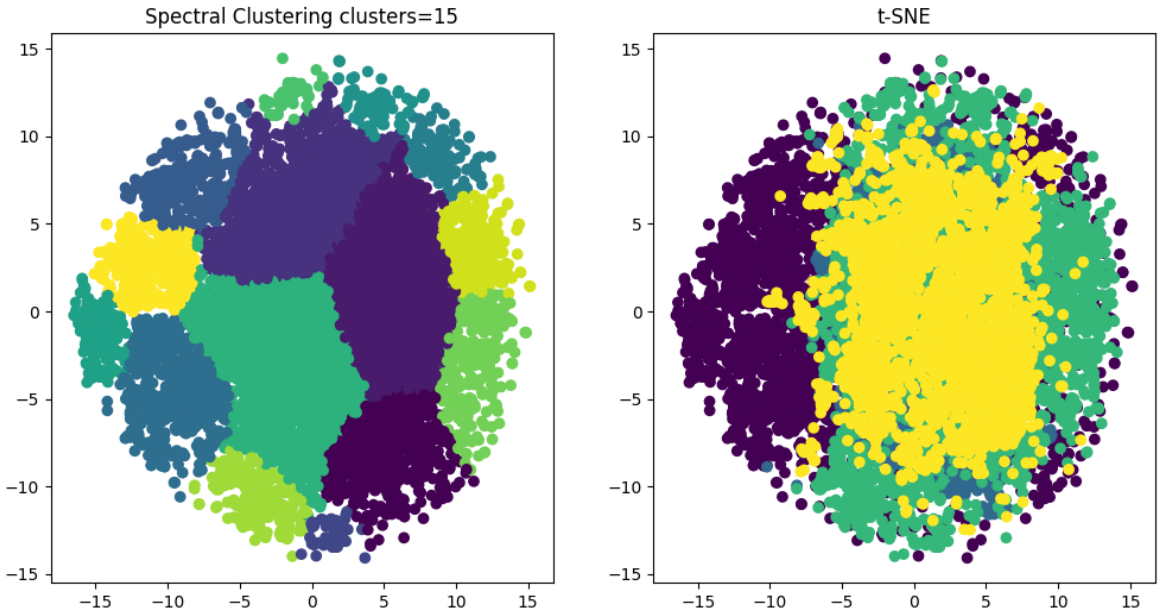
Σχήμα 2.1: Διάγραμμα απεικονισμού κλάσεων

2.1.1 Spectral clustering with rbf

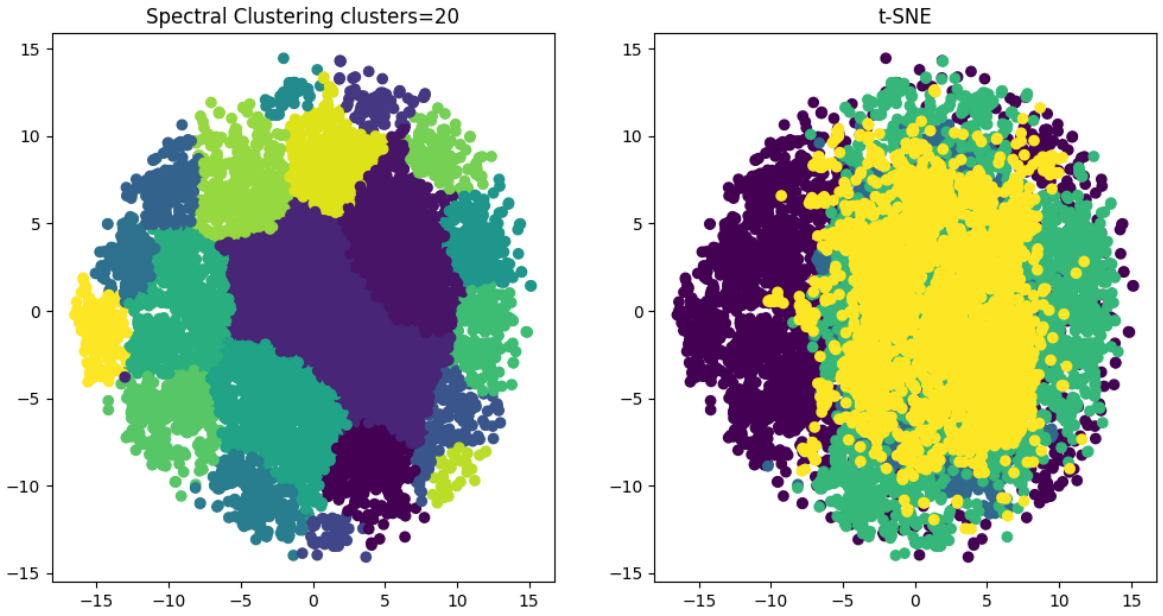
Στο συγκεκριμένο αλγόριθμο, επιλέχθηκε να αλλάζει ο αριθμός των cluster και ορίστηκε ώς δείκτης ομοιότητας μεταξύ των σημείων των δεδομένων το rbf. Τα αποτελέσματα φάνηκαν στις δύο μετρικές στον πίνακα 2.1. Το καλύτερο αποτέλεσμα για την πρώτη μετρική δίνεται όταν οι cluster είναι 15 2.2, και για την δεύτερη όταν είναι 35 2.3.

Πίνακας 2.1: Μετρικές για διαφορετικές τιμές cluster

Τιμή cluster	Silhouette score	Homogeneity score
10	0.208	0.240
15	0.227	0.243
20	0.182	0.260



Σχήμα 2.2: Διαγράμματα σύγκρισεις απεικόνισης κλάσεων



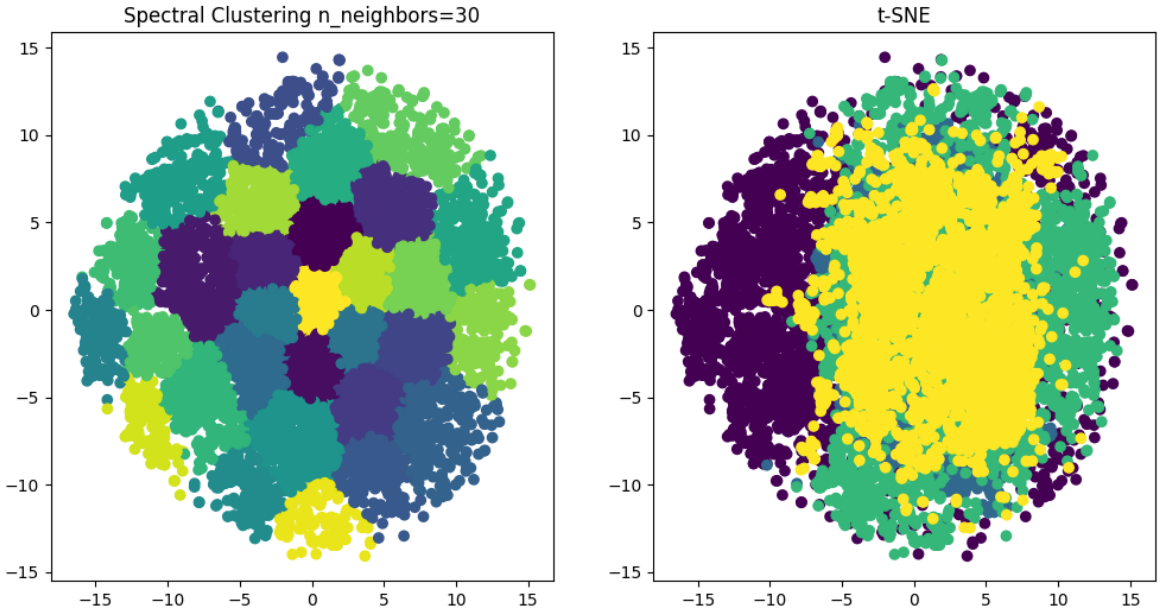
Σχήμα 2.3: Διαγράμματα σύγκρισεις απεικόνισης κλάσεων

2.1.2 Spectral clustering with nearest neighbors

Στην συνέχεια αναπτύχθηκε ο ίδιος αλγόριθμος με διαφορετικό δείκτη ομοιότητας (affinity), ο οποίος είναι ο nearest neighbors. Επιλέχθηκε να αλλάζει ο αριθμός των cluster και ο αριθμός των γειτόνων απο το 5 μέχρι το 40 με βήμα 5. Τα αποτελέσματα φάνηκαν στις δύο μετρικές στον πίνακα 2.2. Το καλύτερο αποτέλεσμα για την πρώτη μετρική δίνεται όταν οι γείτονες και οι cluster είναι 30 2.4, και για την δεύτερη όταν είναι 35 2.5.

Πίνακας 2.2: Μετρικές για διαφορετικές τιμές cluster/γειτόνων

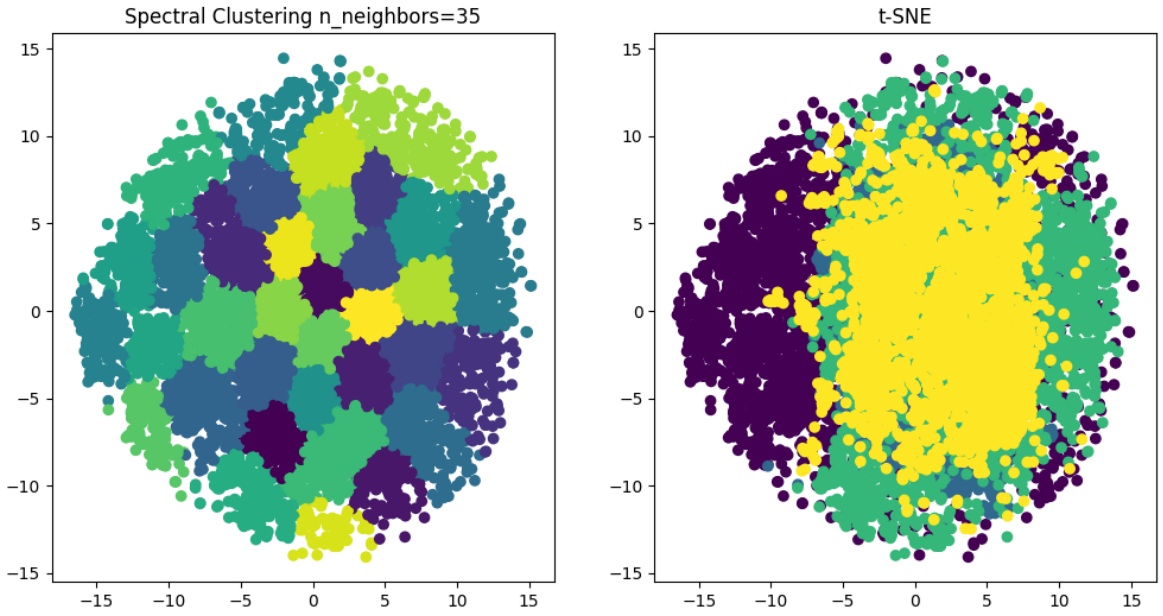
Τιμή cluster/γειτόνων	Silhouette score	Homogeneity score
5	-0.191	0.013
10	0.319	0.179
15	0.328	0.227
20	0.321	0.266
25	0.336	0.290
30	0.337	0.303
35	0.333	0.306



Σχήμα 2.4: Διαγράμματα σύγκρισεις απεικόνισης κλάσεων

2.1.3 Σύγκριση μεταξύ αλγορίθμων

Τα αποτελεσμάτα των μετρικών είναι παρόμοια στο spectral clustering με nearest neighbors και με rbf, και εξίσου χαμηλά.



Σχήμα 2.5: Διαγράμματα σύγκρισεις απεικόνισης κλάσεων

2.2 Isomap

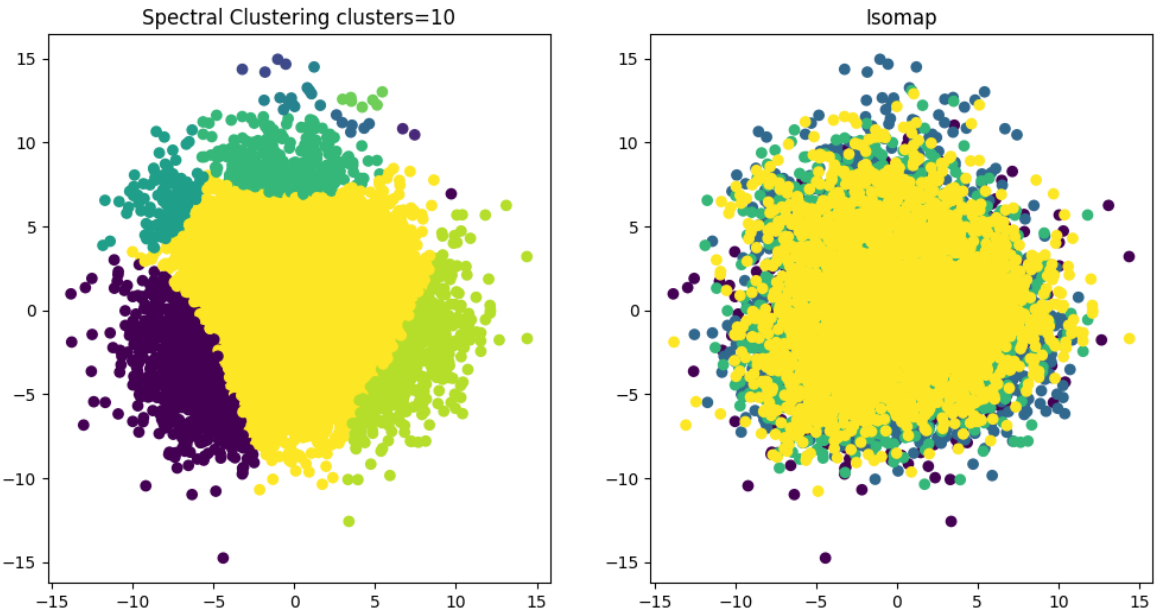
Στην συνέχεια αναπτύχθηκε ο αλγόριθμος Isomap για την μείωση των διαστάσεων σε δύο στο dataset. Ελέγχθηκαν διαφορετικές τιμές components στο isomap και όλες βγάλανε παρόμοια αποτελέσματα στις απεικονίσεις των διαγραμμάτων, για αυτό τον λόγο η διερεύνηση των αλγορίθμων Spectral clustering και kmeans έγινε με components = 2 στο Isomap.

2.2.1 Spectral clustering with rbf

Στο συγκεκριμένο αλγόριθμο, επιλέχθηκε να αλλάζει ο αριθμός των cluster και ορίστηκε ώς δείκτης ομοιότητας μεταξύ των σημείων των δεδομένων το rbf. Τα αποτελέσματα φάνηκαν στις δύο μετρικές στον πίνακα 2.3. Το καλύτερο αποτέλεσμα για την πρώτη μετρική δίνεται όταν οι cluster είναι 10 2.6, και για την δεύτερη το αποτέσμα παραμένει το ίδιο.

Πίνακας 2.3: Μετρικές για διαφορετικές τιμές cluster

Τιμή cluster	Silhouette score	Homogeneity score
10	0.103	0.012
15	0.004	0.012
20	0.013	0.012



Σχήμα 2.6: Διαγράμματα σύγκρισεις απεικόνισης κλάσεων

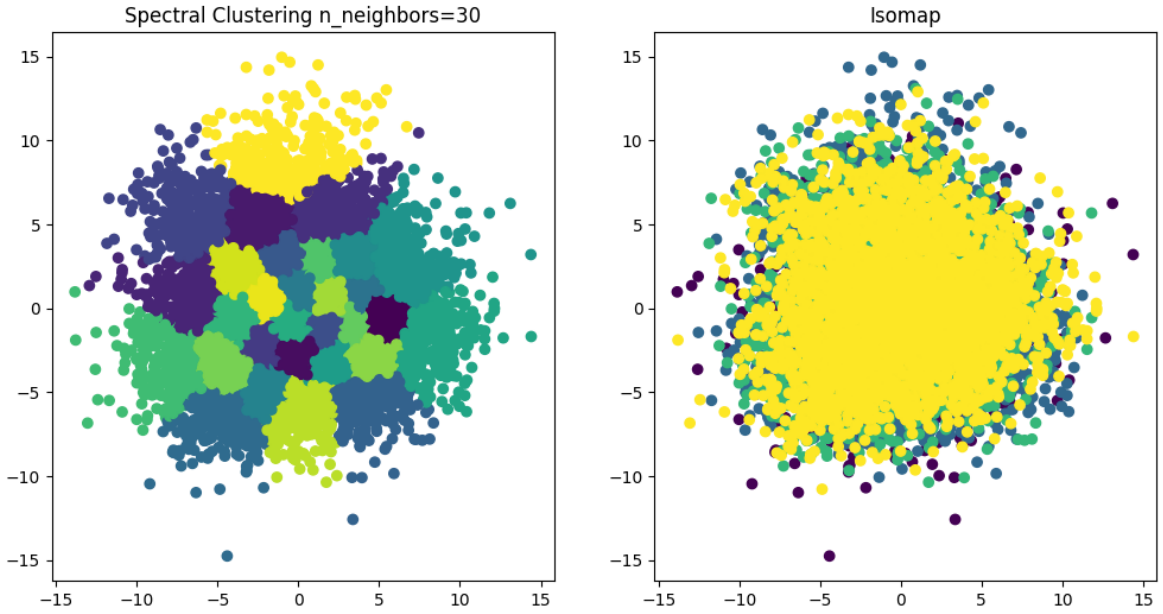
2.2.2 Spectral clustering with nearest neighbors

Στην συνέχεια αναπτύχθηκε ο ίδιος αλγόριθμος με διαφορετικό δείκτη ομοιότητας (affinity), ο οποίος είναι ο nearest neighbors. Επιλέχθηκε να αλλάζει ο αριθμός των cluster και ο αριθμός των γειτόνων απο το 5 μέχρι το 40 με βήμα 5. Τα αποτελέσματα φάνηκαν στις δύο μετρικές στον πίνακα 2.4. Το καλύτερο αποτέλεσμα και για τη 1η μετρική δίνεται

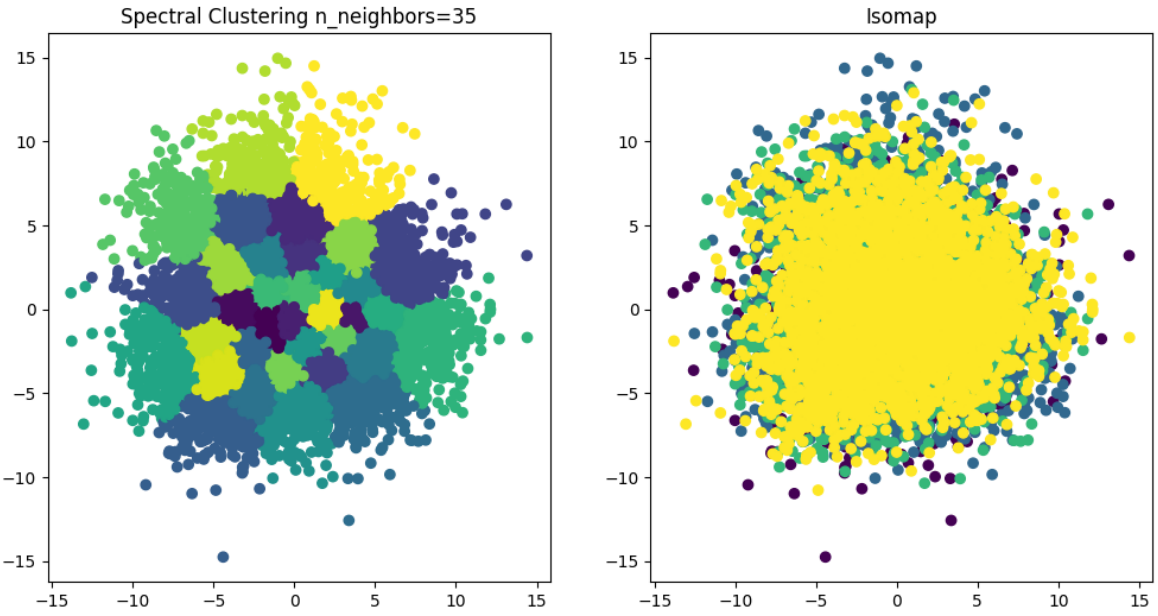
όταν ο αριθμός των cluster και των γειτόνων είναι 30 2.7 και για την 2η όταν είναι 35 2.8

Πίνακας 2.4: Μετρικές για διαφορετικές τιμές cluster/γειτόνων

Τιμή cluster/γειτόνων	Silhouette score	Homogeneity score
5	-0.036	0.003
10	0.277	0.016
15	0.301	0.020
20	0.282	0.022
25	0.288	0.024
30	0.289	0.025
35	0.288	0.026



Σχήμα 2.7: Διαγράμματα σύγκρισεις απεικόνισης κλάσεων



Σχήμα 2.8: Διαγράμματα σύγκρισεις απεικόνισης κλάσεων

2.2.3 Σύγκριση μεταξύ αλγορίθμων

Τα αποτελεσμάτα των μετρικών είναι παρόμοια στο spectral clustering με nearest neighbors έχουν καλύτέρα αποτελέσματα απο αυτά του rbf.

2.3 Σύγκριση μεταξύ t-SNE και Isomap

Το dataset στο οποίο αναπτύχθηκαν οι παραπάνω αλγόριθμοι δεν μας βοηθάει να βγάλουμε κάποια ερμηνεία για τα αποτελέσματα των δύο μετρικών. Αυτό συμβαίνει διότι τα σημεία των δεδομένω είναι πολύ κοντά μεταξύ τους και δεν μπορεί να γίνει καλός διαχωρισμός μετά απο μείωση διαστάσεων με τους αλγορίθμους t-SNE και Isomap. Οπότε παρόλο που τα αποτελέσματα είναι καλύτερα για τον πρώτο αλγόριθμο παραμένουν πάρα πολυ χαμηλά. Αυτο διακρίνεται σε μεγάλο βαθμό στα διαγράμματα απεικονισεις των κλάσεων.