ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ

Μηχανική Μάθηση (7ο εξάμηνο)

Δεύτερη εργασία: Clustering/Dimensionality Reduction

Ημερομηνία: 10/1/2025

Φοιτητής:

Δαδακίδης Γιώργος (iis22127)

Επιβλέπων Καθηγητής: Πρωτοπαπαδάκης Ευτύχιος

Τμήμα Εφαρμοσμένης Πληροφορικής Πληροφοριακά Συστήματα

Πανεπιστήμιο Μακεδονίας

Περιεχόμενα

Λίστα γραφημάτων

Εισαγωγή

Το παρόν πρόβλημα αφορά την ανάλυση του dataset **Fashion-MNIST**, ένα σύνολο δεδομένων εικόνων ρούχων και υποδημάτων που αποτελείται από 70.000 παραδείγματα διαμοιρασμένα σε 10 διαφορετικές κατηγορίες (π.χ., μπλουζάκι, παπούτσι, τσάντα). Η φύση του προβλήματος είναι ανάλυση συσταδοποίησης (clustering) και μείωση διάστασης (dimensionality reduction), με στόχο την αναπαράσταση των δεδομένων σε χαμηλότερες διαστάσεις και τη δημιουργία συστάδων (clusters) που αντικατοπτρίζουν τις δομές του dataset.

Στα πλαίσια της εργασίας εφαρμόζονται 3 διαφορετικές τεχνικές μείωσης διάστασης:

- **Principal Component Analysis (PCA)**,
- **Stacked Autoencoder (SAE)**,
- και **Uniform Manifold Approximation and Projection (UMAP)**,

καθώς και 3 διαφορετικοι αλγόριθμοι clustering:

- **MiniBatch K-Means**,
- **DBSCAN**,
- και **Agglomerative Clustering**.

Αξιολογούνται οι παραπάνω συνδυασμοί με βάση τέσσερις μετρικές απόδοσης:

- **Calinski–Harabasz Index**,
- **Davies–Bouldin Index**,
- **Silhouette Score**,
- και **Dunn Index**.

Τελικός στόχος της έκθεσης είναι η παρουσίαση αναλυτικών αποτελεσμάτων, τόσο με τη μορφή πινάκων όσο και γραφημάτων, ώστε να αξιολογηθεί ο βέλτιστος συνδυασμός μεθόδων για τη συγκεκριμένη περίπτωση. Παράλληλα, θα διερευνηθεί εάν ένας συνδυασμός επιτυγχάνει την καλύτερη απόδοση σε όλες τις μετρικές ή αν απαιτείται προσαρμογή ανάλογα με το κριτήριο που θεωρείται πιο σημαντικό.

Θεωρητικό υπόβαθρο

Στο πλαίσιο της εργασίας αυτής, εφαρμόζεται μια ολοκληρωμένη διαδικασία επεξεργασίας και ανάλυσης δεδομένων, με στόχο την συσταδοποίηση κάθε εικόνας στην κατηγορία που ανήκει (τσάντα, μπλούζα, κλπ)». Ο σκοπός της εργασίας είναι η αξιολόγηση διαφόρων τεχνικών μείωσης διάστασης και συσταδοποίησης, ώστε να κατανοηθεί καλύτερα η δομή

των δεδομένων και να βελτιστοποιηθεί η ποιότητα των clusters που παράγονται. Το θεωρητικό υπόβαθρο που αξιοποιείται για την επίτευξη αυτού του στόχου καλύπτει:

Τεχνικές Μείωσης Διάστασης

Το PCA είναι μια γραμμική μέθοδος που μετασχηματίζει τα δεδομένα σε έναν νέο χώρο μικρότερων διαστάσεων, διατηρώντας τη μέγιστη δυνατή διασπορά. Η μέθοδος βασίζεται στον υπολογισμό των ιδιοτιμών και ιδιοδιανυσμάτων του πίνακα συνδιακύμανσης. Χρησιμοποιήθηκε το **scikit-learn** για την εφαρμογή του PCA.

Παράμετροι:

: Ορίζει τον αριθμό των κύριων συνιστωσών. Χρησιμοποιήθηκαν δύο ρυθμίσεις:
: Για να διατηρηθεί το 95% της διακύμανσης των δεδομένων.
_components = 2 ή 3: Για τη δημιουργία 2D και 3D οπτικοποιήσεων.

Υλοποίηση: Τα δεδομένα κλιμακώθηκαν πριν την εφαρμογή του PCA. Μετά τον μετασχηματισμό, υπολογίστηκε η διασπορά που εξηγείται από κάθε κύρια συνιστώσα και οι ανακατασκευασμένες εικόνες συγκρίθηκαν με τις αρχικές.

Stacked Autoencoder (SAE)

Το SAE είναι ένα νευρωνικό δίκτυο που μαθαίνει να κωδικοποιεί τα δεδομένα σε έναν χώρο μικρότερων διαστάσεων (latent space). Περιλαμβάνει δύο μέρη:

: Μετασχηματίζει τις εισόδους σε έναν χώρο μικρότερων διαστάσεων.
: Ανακατασκευάζει τα δεδομένα από το latent space.

Παράμετροι:

: Ρυθμίστηκε σε **64** διαστάσεις.
: Χρησιμοποιήθηκε **ReLU** για τα ενδιάμεσα επίπεδα και **sigmoid** για την έξοδο.
: Επιλέχθηκε το Adam για την ελαχιστοποίηση της μέσης τετραγωνικής απόκλισης

: Εφαρμόστηκε για την αποφυγή υπερεκπαίδευσης.

Υλοποίηση: Μετά την εκπαίδευση, το encoder απομονώθηκε για την παραγωγή των μειωμένων χαρακτηριστικών, ενώ οι ανακατασκευασμένες εικόνες συγκρίθηκαν με τις αρχικές.

Uniform Manifold Approximation and Projection (UMAP)

Το UMAP επιλέχθηκε ως 3η τεχνική dimensionality reduction ως μια μη γραμμική μέθοδος που διατηρεί τη γεωμετρία του τοπικού χώρου δεδομένων. Είναι ιδιαίτερα αποδοτική σε μεγάλα σύνολα δεδομένων.

Παράμετροι:

: Ρυθμίστηκε σε **2**, ώστε να διευκολυνθεί η οπτικοποίηση.
: Χρησιμοποιήθηκε το **42** για επαναληψιμότητα.

Υλοποίηση: Το UMAP εφαρμόστηκε σε τρισδιάστατα δεδομένα και χρησιμοποιήθηκε για τη δημιουργία χαρτών 2D με στόχο την ανάλυση της δομής των δεδομένων και τη βελτιστοποίηση των αλγορίθμων clustering.

Αλγόριθμοι Συσταδοποίησης

Η μέθοδος MiniBatch KMeans είναι μια παραλλαγή του KMeans που χρησιμοποιεί τυχαία υποσύνολα (mini-batches) για τη βελτιστοποίηση της απόδοσης σε μεγάλα δεδομένα. Ο αριθμός των clusters ρυθμίστηκε σε **10**, ισοδύναμο με τον αριθμό των κατηγοριών του dataset.

Παράμετροι:

DBSCAN

Το DBSCAN (Density-Based Spatial Clustering of Applications with Noise) αναγνωρίζει clusters με βάση την πυκνότητα δεδομένων, ανιχνεύοντας ταυτόχρονα και τις εκτός cluster παρατηρήσεις (outliers).

Παράμετροι:

: 0.5 (μέγιστη απόσταση για ένα σημείο να θεωρηθεί γείτονας).
: 10 (ελάχιστος αριθμός σημείων για να σχηματιστεί ένα cluster).

Επιπλέον, για την χρήση του DBSCAN χρησιμοποιήθηκε η τεχνική grid search για να βρεί τις βέλτιστες τιμές παραμέτρων. Ωστόσο, η χρήση αυτής της τεχνικής ήταν υπολογιστικά ακριβή και χρονοβόρα οπότε επιλέχθηκε η αφαίρεση της απο το πείραμα.

Agglomerative Clustering

Πρόκειται για μια ιεραρχική μέθοδο συσταδοποίησης που συνενώνει clusters με βάση τη μικρότερη απόσταση (linkage). Χρησιμοποιήθηκε η μέθοδος σύνδεσης **ward**.

Παράμετροι:

,average,complete

Για την παράμετρο linkage διεξάχθηκαν πειράματα σχετικά με το βέλτιστο τρόπο υπολογισμου απόστασης. Με την παράμετρο Linkage=complete σημειώθηκαν τα βέλτιστα αποτελέσματα στις 4 μετρικές.

Μετρικές Αξιολόγησης Clustering

Οι μετρικές που χρησιμοποιήθηκαν για την αξιολόγηση των αποτελεσμάτων συσταδοποίησης είναι:

alinski–Harabasz Index: Αξιολογεί τη διαφορά μεταξύ clusters.
avies–Bouldin Index: Μετρά τη συνοχή και τη διαχωρισιμότητα των clusters.
: Μετρά την ποιότητα των clusters συγκρίνοντας την απόσταση ενός σημείου από το cluster του με τα άλλα clusters.
: Υπολογίζει την αναλογία μεταξύ των ελάχιστων αποστάσεων διαφορετικών clusters και των μέγιστων αποστάσεων εντός των clusters.

Με τη χρήση αυτών των μεθόδων, διερευνήθηκαν οι διαφορές στην ποιότητα των clusters που προκύπτουν από τις διαφορετικές τεχνικές μείωσης διάστασης και τους αλγορίθμους συσταδοποίησης.

Πειραματικά αποτελέσματα

Στο παρόν κεφάλαιο παρουσιάζονται τα πειραματικά αποτελέσματα που προέκυψαν από τη χρήση τριών τεχνικών μείωσης διάστασης (PCA, SAE, UMAP) και την εφαρμογή τριών διαφορετικών αλγορίθμων clustering (MiniBatch K-Means, DBSCAN, Agglomerative Clustering). Τα αποτελέσματα αξιολογούνται με βάση τέσσερις μετρικές: Calinski–Harabasz, Davies–Bouldin, Silhouette και Dunn Index, προσφέροντας μια πλήρη συγκριτική εικόνα της απόδοσης.

3.1 Διαγράμματα Dataset

Το dataset Fashion-MNIST προεπεξεργάστηκε μέσω κανονικοποίησης (normalization) των τιμών των pixel και διαχωρίστηκε σε σύνολα εκπαίδευσης, επικύρωσης και ελέγχου. Κάθε τεχνική μείωσης διάστασης εφάρμοσε τις εξής μεθόδους συσταδοποίησης: MiniBatch K-Means, DBSCAN και Agglomerative Clustering. Οι μέσες τιμές των μετρικών υπολογίστηκαν για κάθε τεχνική, ώστε να γίνει μια συγκριτική αξιολόγηση των επιδόσεών τους.

Στο παρακάτω διάγραμμα παρουσιάζονται ενδεικτικά ορισμένες εικόνες του dataset, επιτρέποντας μια αρχική κατανόηση των κατηγοριών και της οπτικής ποικιλομορφίας τους:

Figure 1: Random images from dataset

Οι εικόνες αποτελούνται απο 28X28 pixels πριν από την μείωση διάστασης. Ορισμένες απο τ ι ς

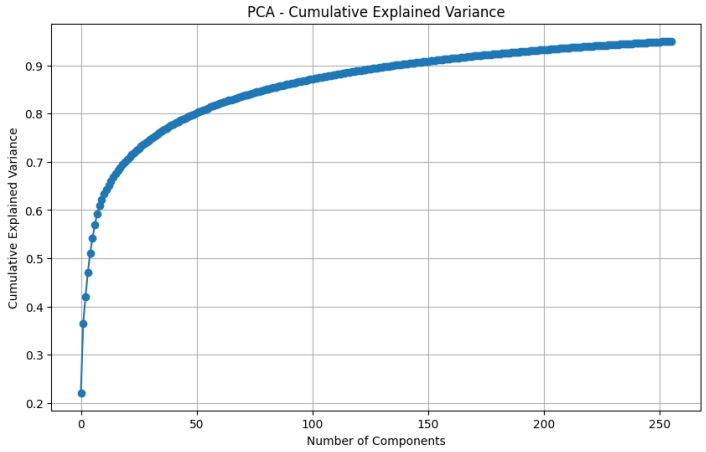
Το παρακάτω διάγραμμα απεικονίζει τον αθροιστικό λόγο εξηγούμενης διακύμανσης ως συνάρτηση του αριθμού των κύριων συνιστωσών (PCA) που εφαρμόζονται στο σύνολο δεδομένων Fashion-MNIST.

σ τ ά δ ε ς

ε ί ν α ι

ser,coat.

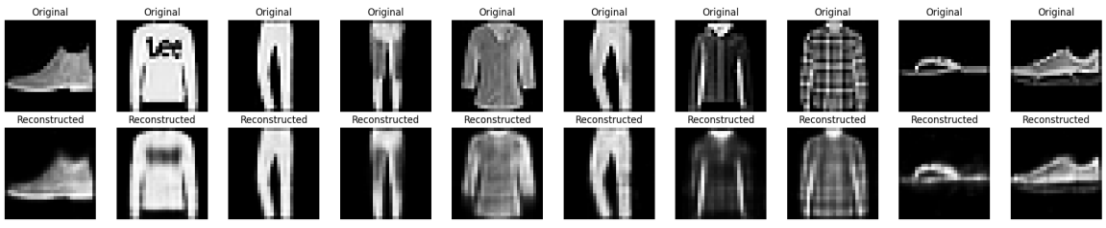
Figure 2: PCA- Cumulative Explained Variance



Η καμπύλη δείχνει ότι οι πρώτες συνιστώσες συλλαμβάνουν ένα σημαντικό μέρος της διακύμανσης, με περίπου 95% της συνολικής διακύμανσης να διατηρείται με τη χρήση λιγότερων από 200 συνιστωσών. Αυτό αποδεικνύει την αποτελεσματικότητα της PCA στη μείωση της διαστατικότητας, διατηρώντας παράλληλα την πλειονότητα των πληροφοριών του συνόλου δεδομένων. Η απότομη αρχική κλίση υποδεικνύει ότι ένας μικρός αριθμός συνιστωσών αντιπροσωπεύει ένα σημαντικό ποσό διακύμανσης, καθιστώντας την PCA ένα πολύτιμο εργαλείο για τη μείωση της διαστατικότητας σε αυτό το πλαίσιο. Πέρα από ένα ορισμένο σημείο, η καμπύλη ισοπεδώνεται, υποδεικνύοντας φθίνουσες αποδόσεις στην εξηγούμενη διακύμανση με πρόσθετες συνιστώσες.

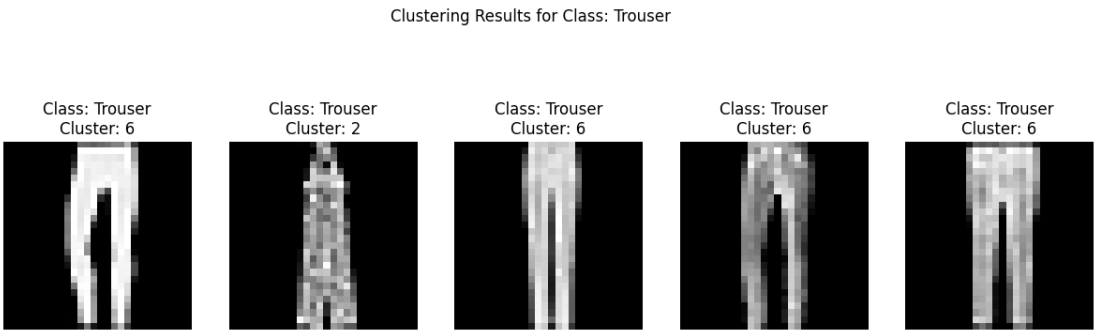
Μετα την χρήση των μοντέλων dimensionality reduction οι εικόνες μειώθηκαν στα 256 χαρακτηριστικά απο τα 784. Ωστόσο οι κλάσεις παραμένουν ακόμα ευδιάκριτες μετά την μείωση διάστασης. Ενδεικτικα παρατίθεται 10 τυχαίες εικόνες πριν και μετά την μείωση της διάστασης τους με την μεθοδο PCA.

Figure 3: Random pictures before and after reconstructrion (PCA)



Τέλος, παρουσιάζονται ενδεικτικά αποτελέσματα συσταδοποίησης με την χρήση UMAP με μοντέλο clustering MiniBatch KMeans για την κλάση “Trouser”:

Figure 4: Clustering results for class "trouser"



Η εικόνα απεικονίζει τα αποτελέσματα της συσταδοποίησης για την κλάση «Trouser», με προβλεπόμενες ετικέτες συστάδων όπως 6 και 2. Η πλειονότητα των παντελονιών (τέσσερις από τις πέντε εικόνες) αντιστοιχίζονται στη συστάδα 6, υποδεικνύοντας ότι ο αλγόριθμος συσταδοποίησης ομαδοποίησε επιτυχώς τις περισσότερες εικόνες στην ίδια συστάδα, πιθανότατα λόγω των παρόμοιων οπτικών χαρακτηριστικών τους. Ωστόσο, η 2η εικόνα τοποθετείται εσφαλμένα στη συστάδα 2, γεγονός που μπορεί να αντικατοπτρίζει μια μικρή απόκλιση στα χαρακτηριστικά της σε σύγκριση με τις άλλες. Αυτό υποδηλώνει ότι ο αλγόριθμος αποδίδει γενικά καλά για αυτή την κατηγορία, αλλά ορισμένες περιπτώσεις μπορεί να έχουν ταξινομηθεί εσφαλμένα ή να αντιπροσωπεύουν ακραίες τιμές μέσα στα δεδομένα.

Ενώ ορισμένα δείγματα ομαδοποιούνται σταθερά στην ίδια συστάδα, άλλα αποκλίνουν, υποδεικνύοντας πιθανή επικάλυψη στην αναπαράσταση των χαρακτηριστικών ή θόρυβο στη διαδικασία ομαδοποίησης. Αυτή η μεταβλητότητα υπογραμμίζει τη σημασία της προσεκτικής ρύθμισης των παραμέτρων ομαδοποίησης για την επίτευξη ακριβέστερων ομαδοποιήσεων.

Διαγράμματα μοντέλων Dimensionality Reduction

Στην συγκεκριμένη ενότητα παρουσιάζονται αναλυτικά τα αποτελέσματα για κάθε μια από τις τεχνικές dimensionality reduction: PCA,SAE,UMAP. Η κάθε μια τεχνική συγκρίνεται με τις μετρικές CalinskiHarabasz, DaviesBouldin, Silhouette, DunnIndex. Τέλος, παρουσιάζονται γραφήματα που συγκρίνουν την κάθε τεχνική με σκοπό την ανάδειξη της βέλτιστης τεχνικής

:PCA 2D Projection

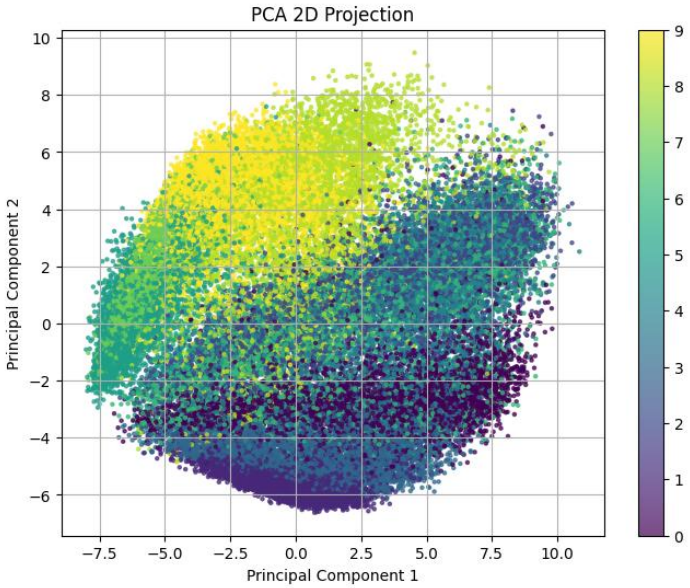
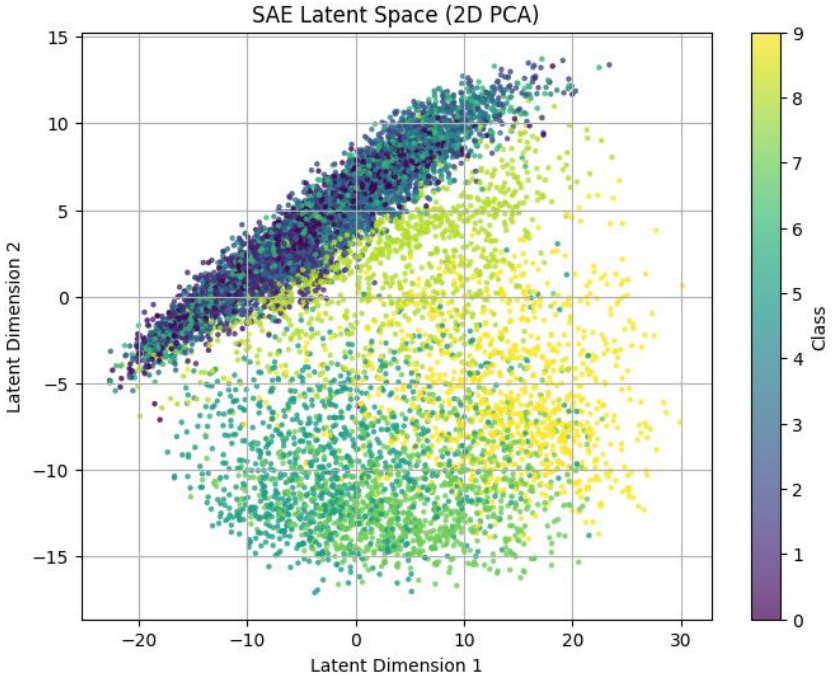
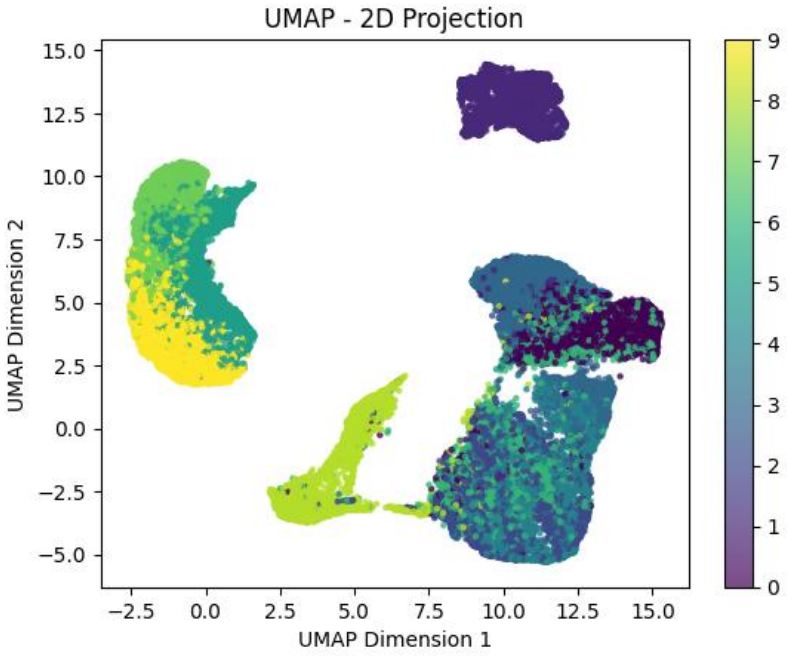


Figure 6: SAE Latent space 2D



: UMAP 2D Projection



Οι τρεις απεικονίσεις που παρέχονται αντιπροσωπεύουν προβολές 2D του συνόλου δεδομένων με τη χρήση τριών διαφορετικών τεχνικών μείωσης της διαστατικότητας: PCA, SAE και UMAP. Κάθε προβολή αναδεικνύει την ικανότητα αυτών των τεχνικών να δομούν το σύνολο δεδομένων σε χώρο χαμηλότερης διάστασης, αναδεικνύοντας διαχωρισμούς ή επικαλύψεις μεταξύ κλάσεων.

Α (Ανάλυση Κύριων Συνιστωσών): Η προβολή PCA 2D δείχνει μια σε γενικές γραμμές κυκλική κατανομή των δεδομένων. Ενώ υπάρχει κάποιος διαχωρισμός μεταξύ των συστάδων, τα όρια δεν είναι ευδιάκριτα. Αυτό το αποτέλεσμα είναι αναμενόμενο από την PCA, καθώς επιδιώκει τη μεγιστοποίηση της διακύμανσης χωρίς να εστιάζει ρητά στο διαχωρισμό των κλάσεων.

Η προβολή 2D PCA του λανθάνοντος χώρου του SAE παρουσιάζει βελτιωμένη γραμμική διαχωρισιμότητα σε σύγκριση με την PCA. Οι συστάδες του λανθάνοντος χώρου εμφανίζονται επιμηκυμένες και ευθυγραμμισμένες, γεγονός που υποδηλώνει ότι ο SAE συλλαμβάνει αποτελεσματικότερα τις μη γραμμικές σχέσεις, ενισχύοντας την υποκείμενη δομή των κλάσεων.

Η προβολή UMAP παρουσιάζει τον πιο έντονο διαχωρισμό των συστάδων, με σαφώς καθορισμένα όρια και διακριτές ομαδοποιήσεις. Αυτό υποδηλώνει ότι η UMAP συλλαμβάνει αποτελεσματικά τόσο τις παγκόσμιες όσο και τις τοπικές δομές, καθιστώντας την ιδιαίτερα κατάλληλη για εργασίες ομαδοποίησης.

Συνοπτικά, ενώ η PCA παρέχει μια γενική επισκόπηση της διακύμανσης, η SAE εισάγει πιο εκλεπτυσμένες δομές ομαδοποίησης και η UMAP επιτυγχάνει την υψηλότερη διαχωρισιμότητα κλάσεων, γεγονός που αντανακλά τη δύναμή της στη διατήρηση τόσο των τοπικών όσο και των παγκόσμιων σχέσεων των δεδομένων.

Figure 8:Comparison of the 3 DimRed techniques

AVG METRICS FOR DIMRED:	CalinskiHarabasz	DaviesBouldin	Silhouette	DunnIndex
PCA	8186.042	0.884	0.338	0.006
SAE	1302.769	1.910	0.134	0.083
UMAP	24469.693	0.699	0.481	0.011
best:	UMAP	UMAP	UMAP	SAE

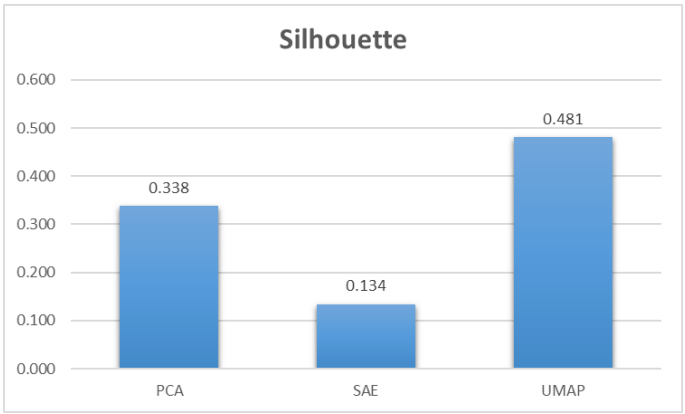
Ο πίνακας παρέχει μια επισκόπηση των μέσων μετρικών απόδοσης για κάθε τεχνική μείωσης διαστάσεων (PCA, SAE και UMAP) σε όλες τις μεθόδους ομαδοποίησης. Η UMAP έχει τις καλύτερες επιδόσεις σε τρεις μετρικές: Calinski-Harabasz Index (24469,693), Davies- (0,699) και Silhouette Score (0,481), καθιστώντας την συνολικά την πιο αποτελεσματική τεχνική για τη δημιουργία διακριτών και καλά διαχωρισμένων συστάδων. Ωστόσο, η SAE ξεχωρίζει στο δείκτη Dunn (0,083), υποδεικνύοντας τη δύναμή της στην επίτευξη μιας καλής ισορροπίας μεταξύ του διαχωρισμού μεταξύ των συστάδων και της συμπίεσης εντός των συστάδων.

Κατά τη σύγκριση της PCA και της SAE, η PCA υπερτερεί της SAE τόσο στο δείκτη Calinski-Harabasz (8186,042) όσο και στο Silhouette Score (0,338), υποδεικνύοντας ότι η PCA σχηματίζει συνολικά καλύτερα καθορισμένες συστάδες σε σύγκριση με την SAE. Ωστόσο, η ισχυρή επίδοση της SAE στο δείκτη Dunn υποδηλώνει ότι διατηρεί καλύτερο διαχωρισμό μεταξύ των συστάδων σε ορισμένα σενάρια.

Με βάση μια προσέγγιση ψηφοφορίας, η UMAP αναδεικνύεται ως η καλύτερη τεχνική μείωσης διαστάσεων, υπερέχοντας σε τρεις από τις τέσσερις μετρικές. Η PCA, αν και δεν είναι τόσο ισχυρή όσο η UMAP, εξακολουθεί να έχει καλύτερες επιδόσεις από την SAE σε δύο κρίσιμες μετρικές, αναδεικνύοντας τη χρησιμότητά της σε εργασίες ομαδοποίησης όπου ο υψηλός διαχωρισμός μεταξύ των συστάδων δεν είναι το κύριο μέλημα.

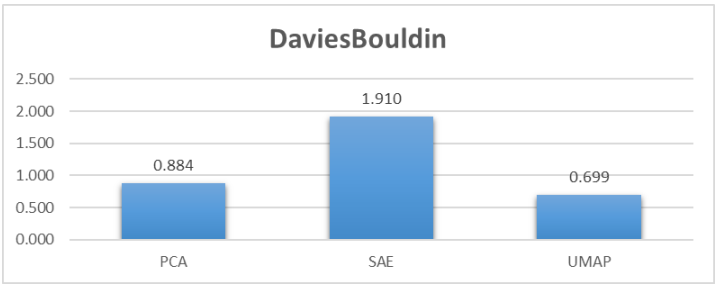
Τέλος προσθέτονται και τα αντίστοιχα γραφήματα για τις μετρικές που αναλύσαμε:

F
i



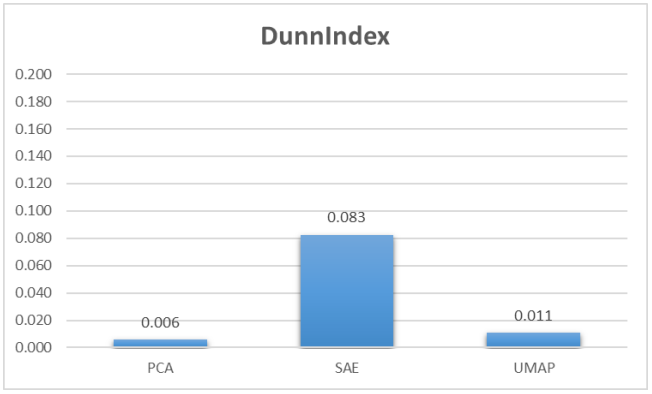
Σχόλιο: Με βάση το Silhouette η τεχνικη UMAP είναι η καλύτερη.

: Best DimRed based on DaviesBouldin



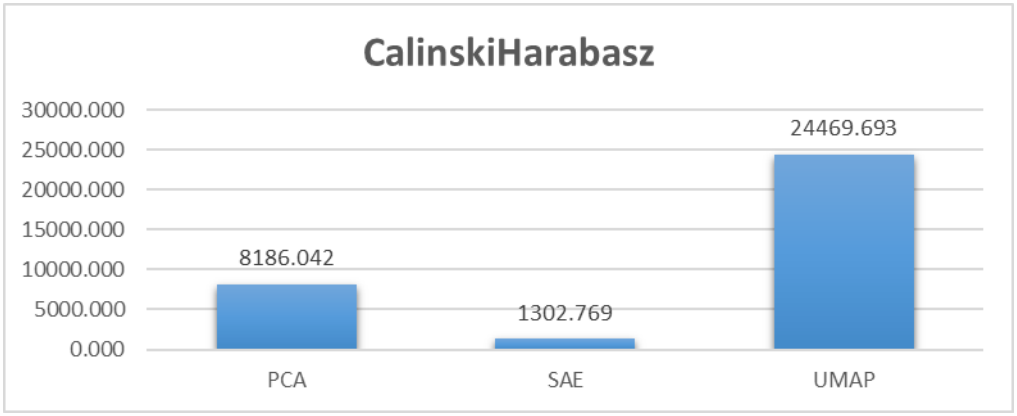
Σχόλιο: Με βάση το DaviesBouldin η τεχνικη UMAP είναι η καλύτερη (η μικρότερη).

F
i



Σχόλιο: Με βάση το Dunnindex η τεχνικη SAE είναι η καλύτερη.

F
i



Σχόλιο: Με βάση το CalinskiHarabasz η τεχνικη UMAP είναι η καλύτερη.

Διαγράμματα τεχνικών Clustering

Παρακάτω παραθέτονται ένας πίνακας με σκοπό να συγκρίνει την απόδοση κάθε αλγορίθμου Clustering λαμβάνοντας υπόψη τις 4 μετρικές απόδοσης:

Figure

AVG METRICS FOR CLUSTERING	CalinskiHarabasz	DaviesBouldin	Silhouette	DunnIndex
MiniBatch KMeans	13557.205	1.105	0.346	0.019
DBSCAN	7088.917	0.841	0.360	0.013
Agglomerative Clustering(WARD)	13230.303	1.158	0.325	0.035
Agglomerative Clustering (AVERAGE)	12421.096	1.018	0.338	0.036
Agglomerative Clustering (COMPLETE)	12228.702	1.344	0.296	0.040
best(Average for 3DimRed):	MiniBatch Kmeans	DBSCAN	DBSCAN	Agg(COMPLETE)
best(excluding DBSCAN):	MiniBatch Kmeans	Agg(AVG)	MiniBatch Kmeans	Agg(COMPLETE)
Best using UMAP	Agg(COMPLETE)	Agg(COMPLETE)	Agg(COMPLETE)	Agg(Ward)

Comparison of the 4 clustering algorithms

Παρόλο που ο DBSCAN παρουσιάζει τις καλύτερες μέσες τιμές, η απόδοσή του είναι παραπλανητική, καθώς αποτυγχάνει να ομαδοποιήσει πλήρως τα δεδομένα (π.χ., σχηματίζει μόνο 2/10 συστάδες με PCA). Για το λόγο αυτό, εξαιρείται από τη σύγκριση για την ανάδειξη του καλύτερου αλγορίθμου.

Όσον αφορά τα άλλα 2 μοντέλα ο MiniBatch Kmeans αποδίδει καλύτερα στις περισσότερες μετρικές. Αν λάμβουμε υπόψη όμως μόνο την βέλτιστη τεχνική μείωσης διάστασης(UMAP) τότε διακρίνεται ξεκάθαρα η τεχνική Agglomerative Clustering(με linktype:Complete) να είναι η καλύτερη επιλογή συσταδοποίησης του συγκεκριμένου dataset.

Συμπεράσματα

Συμπερασματικά, η ανάλυση των πειραματικών αποτελεσμάτων αποκαλύπτει ότι η τεχνική **UMAP** αναδείχθηκε ως η πλέον αποτελεσματική μέθοδος μείωσης διάστασης για το συγκεκριμένο πρόβλημα. Τα αποτελέσματά της ξεπέρασαν εκείνα των PCA και SAE σε τρεις από τις τέσσερις μετρικές απόδοσης (Calinski–Harabasz, Davies–Bouldin, Silhouette). Ειδικότερα, η ικανότητα της UMAP να διατηρεί τόσο τις τοπικές όσο και τις παγκόσμιες σχέσεις του dataset την καθιστά ιδιαίτερα κατάλληλη για εργασίες συσταδοποίησης όπως αυτή. Παρόλα αυτά, η SAE παρουσίασε την υψηλότερη απόδοση στον δείκτη Dunn, γεγονός που υποδηλώνει ότι μπορεί να παρέχει καλύτερη ισορροπία μεταξύ του διαχωρισμού των συστάδων και της συνοχής εντός τους σε συγκεκριμένα σενάρια.

Όσον αφορά τις τεχνικές συσταδοποίησης, ο **MiniBatch K-Means** κατέδειξε συνολικά καλύτερες επιδόσεις στις περισσότερες μετρικές, καθιστώντας τον έναν αξιόπιστο αλγόριθμο για το πρόβλημα. Ωστόσο, όταν λήφθηκε υπόψη μόνο η καλύτερη τεχνική μείωσης διάστασης (UMAP), ο **Agglomerative Clustering** (με σύνδεση "Complete") ανέδειξε εξαιρετικά αποτελέσματα, με καλά καθορισμένες και διαχωρισμένες συστάδες. Η μέθοδος DBSCAN, αν και παρουσιάζει υψηλές βαθμολογίες σε ορισμένες μετρικές, απέτυχε να σχηματίσει τον πλήρη αριθμό των απαιτούμενων συστάδων, γεγονός που την καθιστά λιγότερο κατάλληλη για το συγκεκριμένο πρόβλημα.

Για τη βελτίωση της απόδοσης των τεχνικών, προτείνονται τα εξής:

> **ιερεύνηση υπερπαραμέτρων (hyperparameter tuning)**: Ειδικότερα για το DBSCAN, η επιλογή διαφορετικών τιμών για τις παραμέτρους eps και min_samples ενδέχεται να οδηγήσει σε πιο πλήρη συστάδες.
> **υνδυασμός τεχνικών**: Η υβριδική χρήση διαφορετικών τεχνικών μείωσης διάστασης, όπως η εφαρμογή PCA ως προεπεξεργασία πριν την UMAP, θα μπορούσε να ενισχύσει την απόδοση.
> **ξέταση εναλλακτικών clustering αλγορίθμων**: Η χρήση πιο σύγχρονων μεθόδων, όπως Spectral Clustering ή Gaussian Mixture Models, μπορεί να προσφέρει καλύτερες επιδόσεις.
> **νσωμάτωση περισσότερων μετρικών αξιολόγησης**: Μετρικές όπως το Adjusted Rand Index (ARI) ή το Normalized Mutual Information (NMI) θα μπορούσαν να παρέχουν μια πληρέστερη εικόνα της απόδοσης.

Συνολικά, τα αποτελέσματα υποδεικνύουν ότι η σωστή επιλογή τεχνικής μείωσης διάστασης, σε συνδυασμό με την κατάλληλη μέθοδο clustering, μπορεί να επιτύχει υψηλής ποιότητας συστάδες, διατηρώντας την πληροφορία του dataset. Το UMAP, σε συνδυασμό με το Agglomerative Clustering, προτείνεται ως η βέλτιστη επιλογή για το συγκεκριμένο πρόβλημα, ενώ η περαιτέρω βελτιστοποίηση των υπερπαραμέτρων μπορεί να ενισχύσει τα αποτελέσματα ακόμη περισσότερο.