2η Εργασία: Clustering Problems

Γεώργιος Καμίδης (dai17026)
Ευστάθιος Κωνσταντίνος Αναστασιάδης

21/ 1/ 2024

Μηχανική Μάθηση

Περιεχόμενα

1 Εισαγωγή

Σκοπός της παρούσας αναφοράς είναι η μελέτη διαφόρων τεχνικών μείωσης διάστασης δεδομένων και κατόπιν η χρήση τεχνικών συσταδοποίησης. Για τη μελέτη αυτή χρησιμοποιήθηκε το *Fashion MNIST* σύνολο δεδομένων, το οποίο περιλαμβάνει εικόνες ρούχων και αξεσουάρ από διάφορες κατηγορίες μόδας.

Τρεις τεχνικές μείωσης διάστασης εφαρμόστηκαν επάνω στα αρχικά δεδομένα: η *Principal Component Analysis* ή *PCA*, η *Linear Discriminant Analysis* ή *LDA* και η χρήση stacked autoencoders (*SAE*). Η δημιουργία συστάδων (clusters) έγινε εφαρμόζοντας τρεις ευρέως διαδεδομένες τεχνικές clustering: *MiniBatch KMeans*, *DBSCAN* και *Agglomerative Clustering*.

Τα αποτελέσματα της συσταδοποίησης αξιολογήθηκαν με τον υπολογισμό τεσσάρων evaluation metrics: δείκτης *Calinski-Harabasz*, δείκτης *Davies-Boudin*, *Silhouette score* και *Adjusted rand index*.

Η διάρθρωση της αναφοράς είναι η εξής: στο κεφάλαιο 2 παρατίθεται μια σύντομη θεωρητική εισαγωγή μαζί με την περιγραφή του dataset και των επιλεγμένων παραμέτρων για κάθε αλγόριθμο. Στο τρίτο κεφάλαιο παρουσιάζονται τα αποτελέσματα από την εφαρμογή των αλγορίθμων συσταδοποίησης και συζητώνται, ενώ στο τελευταίο κεφάλαιο (Κεφάλαιο 4) γίνεται μια σύνοψη των αποτελεσμάτων και η καταγραφή των τελικών συμπερασμάτων στα οποία καταλήξαμε.

2 Μεθοδολογία

2.1 Θεωρητικά προαπαιτούμενα

2.1.1 Τεχνικές μείωσης διάστασης

Τα σύνολα δεδομένων που χρησιμοποιούνται κατά τη μηχανική μάθηση (και όχι μόνο) συχνά αποτελούνται από έναν μεγάλο αριθμό features, καθιστώντας τους υπολογισμούς επίπονους και κοστοβόρους. Η ύπαρξη, επιπλέον, πολλών features δυσκολεύει την απεικόνιση των δεδομένων, εξαιτίας της πολυδιαστατικότητάς τους. Για το λόγο αυτό χρειάζεται να καταφύγουμε σε τεχνικές μείωσης διάστασης των δεδομένων, δημιουργώντας "συνθετικά" features τα οποία διατηρούν ένα σημαντικό μέρος της πληροφορίας των αρχικών. Για τους σκοπούς της αναφοράς, περιοριζόμαστε στις τεχνικές PCA, LDA και SAE.

Principal Component Analysis (PCA)

Η PCA προβάλλει τα αρχικά δεδομένα στις κύριες συνιστώσες τους (Principal Components). Οι συνιστώσες αυτές αντιστοιχούν στην κατεύθυνση όπου παρατηρείται η μέγιστη διασπορά στα δεδομένα.

Αποφεύγοντας την αυστηρή μαθηματική παρουσίαση της τεχνικής, αναφέρουμε πως η PCA δημιουργεί τις κύριες συνιστώσες έτσι, ώστε να αποτελεί η κάθε μία γραμμικό συνδυασμό των αρχικών features, υπολογίζοντας ιδιοτιμές και ιδιοδιανύσματα. Αποτελεί, δηλαδή, μια γραμμική μέθοδο.

Linear Discriminant Analysis (LDA)

Στόχος της LDA είναι να προβάλει τα δεδομένα σε ένα χώρο χαμηλότερης διάστασης με τέτοιο τρόπο, ώστε να μεγιστοποιείται ο διαχωρισμός των διαφορετικών κλάσεων, διατηρώντας παράλληλα ελάχιστη τη διακύμανση μεταξύ examples της ίδιας κλάσης. Για το λόγο αυτό, κατά την εφαρμογή της LDA χρειάζονται και οι ετικέτες των δεδομένων και όχι μόνο τα feature vectors.

Όπως και η PCA, η LDA βασίζεται στον υπολογισμό ιδιοτιμών και ιδιοδιανυσμάτων για τη μείωση της διάστασης.

Stacked Auto Encoders (SAE)

Οι SAE βασίζονται στη χρήση τεχνητών νευρωνικών δικτύων προκειμένου να δημιουργήσουν μια χαμηλής διάστασης αναπαράσταση των εισαγόμενων, στο μοντέλο, δεδομένων. Πιο συγκεκριμένα, αποτελούνται από πολλαπλά layer διαφόρων autoencoders, τοποθετημένων διαδοχικά.

Ένας autoencoder περιλαμβάνει έναν encoder, για την προβολή των δεδομένων σε ένα χώρο λιγότερων διαστάσεων, και έναν decoder, ο οποίος ανακατασκευάζει τα αρχικά δεδομένα με βάση τη νέα αναπαράσταση που έφτιαξε ο encoder. Όπως προαναφέραμε, ένας Stacked Autoencoder περιλαμβάνει πολλούς autoencoders. Κάθε autoencoder εκπαιδεύεται από την αναπαράσταση των δεδομένων που δημιούργησε ο προηγούμενος autoencoder. Με τη χρήση της τεχνικής αυτής μπορεί να εφαρμοστεί μη γραμμική μείωση διάστασης πάνω στα δεδομένα, σε αντίθεση με τις τεχνικές PCA και LDA.

2.1.2 Τεχνικές Clustering

Το clustering αποτελεί μια μέθοδο unsupervised learning και αποσκοπεί στην ταξινόμηση δεδομένων με παρόμοια χαρακτηριστικά σε διακριτές ομάδες. Μέσω της εφαρμογής clustering τεχνικών ανακαλύπτονται μη τετριμμένες σχέσεις μεταξύ των παρατηρήσεων του μελετώμενου συνόλου δεδομένων. Στην παρούσα αναφορά καταπιανόμαστε με τις ακόλουθες τεχνικές: MiniBatch KMeans, DBSCAN, Agglomerative clustering.

MiniBatch KMeans

Ο MiniBatch KMeans βασίζεται στην προεπιλογή του αριθμού των clusters και στη δημιουργία centroids ίσων με τον αριθμό των επιλεγμένων clusters. Κατόπιν υπολογίζονται οι αποστάσεις ενός υποσυνόλου παρατηρήσεων από τα επιλεγμένα centroids και το κοντινότερο centroid σε κάθε παρατήρηση καταγράφεται ως label της παρατήρησης. Για κάθε centroid, ύστερα, υπολογίζεται το μέσο feature διάνυσμα των παρατηρήσεων με label το συγκεκριμένο centroid. Οι μέσες τιμές αποτελούν τα νέα centroids. Η διαδικασία συνεχίζεται επαναληπτικά μέχρι να συγκλίνει ο αλγόριθμος σε κάποια τελικά centroids.

DBSCAN

Η μέθοδος DBSCAN ομαδοποιεί τις παρατηρήσεις ορίζοντας τους clusters ως περιοχές πυκνές σε αριθμό παρατηρήσεων, σε αντίθεση με άλλες περιοχές όπου οι παρατηρήσεις είναι πιο διάσπαρτες. Έτσι, ο DBSCAN αλγόριθμος χρησιμεύει και στον εντοπισμό outliers.

Ο αλγόριθμος επιλέγει αρχικά μια τυχαία παρατήρηση και την εντάσσει στον πρώτο cluster. Στη συνέχεια, καταμετράται ο αριθμός παρατηρήσεων που απέχει από την επιλεγμένη παρατήρηση απόσταση μικρότερη ή ίση του ϵ, όπου ϵ είναι μια υπερπαράμετρος του αλγορίθμου, ορισμένη από τον χρήστη. Αν ο αριθμός που καταμετρήθηκε είναι μεγαλύτερος ή ίσος του n (άλλη μια υπερπαράμετρος ορισμένη από το χρήστη), τότε όλες αυτές οι παρατηρήσεις που ικανοποιούν τις παραπάνω συνθήκες εντάσσονται στον cluster του αρχικού παραδείγματος. Η διαδικασία συνεχίζεται μέχρι να ταξινομηθούν όλα τα δεδομένα σε clusters ή να καταγραφούν ως outliers.

Agglomerative Clustering

Ο Agglomerative clustering αλγόριθμος είναι ένας αλγόριθμος ιεραρχικής συσταδοποίησης. Εκκινείται θεωρώντας κάθε παρατήρηση ως έναν ξεχωριστό cluster και μετά από κάθε επανάληψη ενώνει τους clusters που θεωρούνται κοντινότεροι. Η διαδικασία ενώσης των διαφορετικών clusters απεικονίζεται με δενδρόγραμμα.

2.2 Περιγραφή dataset και προεπεξεργασία

Το Fashion MNIST dataset είναι μια συλλογή grayscale εικόνων που αντιστοιχούν σε ρούχα ή αξεσουάρ μόδας, κάθε ένα από τα οποία ταξινομείται σε μια συγκεκριμένη κλάση. Υπάρχουν δέκα διαφορετικές κλάσεις στο dataset, οι οποίες συνοψίζονται στον παρακάτω πίνακα.

Table 1: Fashion MNIST κατηγορίες αντικειμένων.

Ετικέτες	**Περιγραφή**
0	T-shirt/top
1	Trouser
2	Pullover
3	Dress
4	Coat
5	Sandal
6	Shirt
7	Sneaker
8	Bag
9	Ankle boot

Κάθε εικόνα του dataset έχει διαστάσεις 28x28. Συνεπώς, κάθε παρατήρηση αναπαρίσταται με έναν πίνακα (numpy array) 28x28. Κάθε pixel είναι ένας ακέραιος αριθμός από το 0 έως το 255.

Το σύνολο εκπαίδευσης αποτελείται από 60000 παρατηρήσεις, ενώ το test σύνολο είναι σημαντικά μικρότερο, αποτελούμενο από 10000 παρατηρήσεις. Το σύνολο εκπαίδευσης χωρίστηκε, επιπλέον, σε σύνολο εκπαίδευσης και επαλήθευσης (validation). Επιλέξαμε το validation σύνολο να είναι περίπου ίσο σε μέγεθος με το test σύνολο, για το λόγο αυτό από το αρχικό σύνολο εκπαίδευσης χρησιμοποιήσαμε 16% των παρατηρήσεων για τη δημιουργία του validation συνόλου.

Πριν την εφαρμογή τεχνικών μείωσης διάστασης και αλγορίθμων συσταδοποίησης κανονικοποιήσαμε τα δεδομένα στο διάστημα $[0, 1]$ και κατόπιν μετατρέψαμε τις εικόνες σε μονοδιάστατες numpy arrays, όπου κάθε παρατήρηση έχει 784 features.

2.3 Επιλογή παραμέτρων

Η επιλογή των καταληλλότερων παραμέτρων για τις τεχνικές μείωσης διάστασης και τους αλγορίθμους συσταδοποίησης έγινε συνδυάζοντας τυχαία αναζήτηση παραμέτρων (Randomized Search CV), τη διαίσθηση μέσω παρατήρησης των δεδομένων και την αξιοποίηση αποτελεσμάτων σχετικής δουλειάς που βρέθηκε κατά την αναζήτηση βιβλιογραφικών πηγών.

Πιο συγκεκριμένα, για τις τεχνικές μείωσης διάστασης PCA και LDA χρησιμοποιήθηκε η Randomized Search CV μέθοδος η οποία, για εξοικονόμηση χρόνου εκτέλεσης του κώδικα, πραγματοποιήθηκε σε σχετικά Jupyter notebooks. Η αρχιτεκτονική των SAE επιλέχθηκε να είναι συμμετρική, αποτελούμενη από δύο αρχικά layer και ένα bottleneck layer. Αντίστοιχα, ο decoder αποτελούνταν από δύο όμοια layers με αυτά του encoder. Ελέγχθηκαν τέσσερεις διαφορετικές αρχιτεκτονικές με διαφορετικό αριθμό νευρώνων σε κάθε layer και loss function. Όλοι οι SAE εκπαιδεύτηκαν για 10 εποχές και για batch size ίσο με 64. Τέλος, για τις τεχνικές συσταδοποίησης, χρησιμοποίηθηκαν οι default παράμετροι για την DBSCAN μέθοδο, ενώ για τις άλλες δύο εκτιμήσαμε πως ο αριθμός των labels (10) των παρατηρήσεων, αποτελεί μια καλή και λογική εκτίμηση για τον αρχικό αριθμό centroids. Οι παράμετροι που χρησιμοποιήθηκαν συνοψίζονται παρακάτω. Περισσότερες πληροφορίες σχετικά με την επιλογή παραμέτρων δίνονται στο Παράρτημα.

Table 2: Επιλεγμένες παράμετροι για τις PCA και LDA τεχνικές .

Μέθοδος	**Παράμετροι**
PCA	$components = 187$
LDA	$solver = svd$

Table 3: Αρχιτεκτονική SAE. Η loss function που ελαχιστοποιείται είναι η MSE.

Layer	**Νευρώνες**	**Activation Function**
Encoder 1	512	relu
Encoder 2	128	relu
Bottleneck	2	linear
Decoder 1	128	relu
Decoder 2	512	relu
Output	784	sigmoid

Table 4: Παράμετροι τεχνικών συσταδοποίησης.

Τεχνική	**Παράμετροι**
MiniBatch KMeans	$n - clusters = 10$
DBSCAN	$\epsilon = 0.5$, $n = 5$
Agglomerative	$n - clusters = 10$

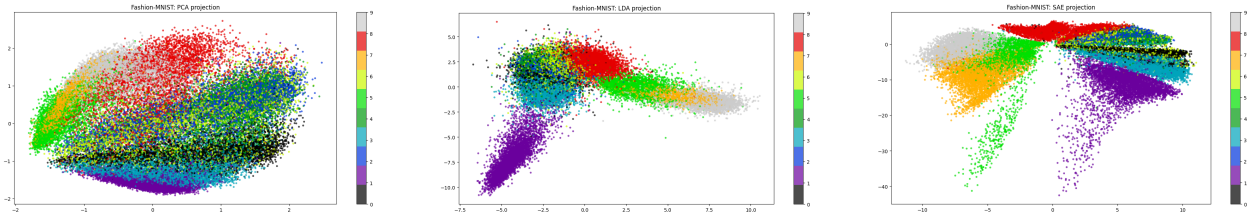
3 Αποτελέσματα συσταδοποίησης

3.1 Σύγκριση εκπαίδευσης τεχνικών μείωσης διάστασης

Στην Εικόνα 1, απεικονίζονται τα scatter plot που δημιουργήθηκαν έπειτα από την εφαρμογή των τεχνικών μείωσης διάστασης. Στην περίπτωση της PCA φαίνεται πως η συστάδα που αντιστοιχεί στην ετικέτα *Trouser* είναι διαχωρισμένη από τις υπόλοιπες συστάδες σε ικανοποιητικό βαθμό (σημεία με μωβ χρώμα). Παρομοίως ο συνδυασμός των συστάδων *Sandal* με *Sneaker*, καθώς και ο συνδυασμός *Bag* με *Ankle Boot* είναι καλώς διαχωρισμένος. Περαιτέρω διαχωρισμοί δεν παρατηρούνται, καθώς οι παρατηρήσεις που αντιστοιχούν σε διαφορετικές κλάσεις αναμειγνύονται μεταξύ τους.

Οι κλάσεις εμφανίζονται περισσότερο διαχωρισμένες μετά από την εφαρμογή της LDA στα δεδομένα. Οι κλάσεις *Trouser*, *Ankle boot*, *Sandal*, *Bag* και *Sneaker* είναι ευκρινώς διαχωρισμένες και ελάχιστα σημεία τους δεν ανήκουν στον αντίστοιχο cluster. Εξακολουθεί να υπάρχει, ωστόσο, ισχυρή ανάμειξη μεταξύ των παρατηρήσεων που ανήκουν στις ομάδες *T-shirt/top*, *Pullover*, *Coat* και *Dress*.

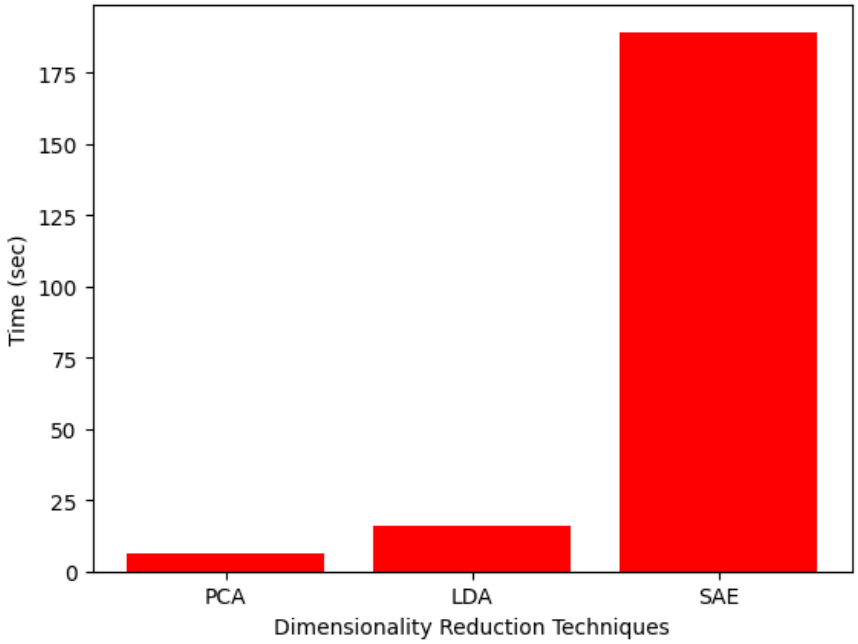
Τέλος, η χρήση τεχνικής μείωσης διάστασης μέσω SAE δίνει τα καλύτερα αποτελέσματα συσταδοποίησης, καθώς όπως φαίνεται και στην Εικόνα 3c, υπάρχει μικρή ανάμιξη μεταξύ των παρατηρήσεων που ανήκουν σε διαφορετικές κλάσεις.



(a) PCA (b) LDA (c) SAE

Εικόνα 1: Τεχνικές μείωσης διάστασης στο επίπεδο.

Οι χρόνοι εκπαίδευσης που χρειάστηκε κάθε μία από τις επιλεγμένες μεθόδους παρουσιάζονται στην Εικόνα 2.



Εικόνα 2: Χρόνοι εκπαίδευσης για κάθε μία τεχνική μείωσης διάστασης σε δευτερόλεπτα.

Όπως είναι αναμενόμενο, η μείωση διάστασης με τη χρήση SAE είναι και η πιο χρονοβόρα. Ο χρόνος που απαιτεί, ωστόσο, η εκπαίδευση με νευρωνικά δίκτυα επηρεάζεται άμεσα από τις εποχές που ορίσαμε για εκπαίδευση, αλλά και φυσικά και από την αρχιτεκτονική του μοντέλου. Παρατηρούμε, επίσης, πως η LDA απαιτεί περισσότερο χρόνο εκτέλεσης σε σχέση με την PCA, παρόλο που είναι και οι δύο γραμμικές μέθοδοι βασισμένες στον υπολογισμό ιδιοτιμών και ιδιοδιανυσμάτων. Εικάζουμε πως η διαφορά στην εκτέλεση οφείλεται και στο γεγονός ότι η LDA κατά την εκπαίδευση χρησιμοποιεί και τα labels των δεδομένων.

Παρά το γεγονός ότι η μείωση διάστασης με SAE κοστίζει σε χρόνο, τα αποτελέσματά της εί-

ναι καλύτερα από τις υπόλοιπες μεθόδους, όσον αφορά το σωστό διαχωρισμό των παρατηρήσεων. Ωστόσο, και η LDA αποδίδει ικανοποιητικά και έχει το πλεονέκτημα ότι είναι σαφώς γρηγορότερη μέθοδος από τον SAE. Μειονεκτεί, όμως, στο ότι απαιτεί να έχουμε γνωστά και τα true labels των δεδομένων, κάτι που πολύ συχνά δεν είναι εφικτό.

Τέλος, στην παρακάτω εικόνα παρουσιάζονται οι ανακατασκευασμένες εικόνες για τυχαία παραδείγματα του συνόλου δεδομένων, όπως εξήχθησαν από τον decoder του SAE. Τα ανακατασκευασμένα παραδείγματα παρουσιάζονται πιο θολά, ωστόσο, είναι εύκολο να καταλάβει κανείς την κατηγορία στην οποία ανήκει το κάθε αντικείμενο.

Εικόνα 3: Ανακατασκευασμένα τυχαία παραδείγματα για κάθε κλάση έπειτα από χρήση του SAE.

3.2 Σύγκριση τεχνικών συσταδοποίησης

Στον Πίνακα 5 (*Table 5*) παρουσιάζονται οι αριθμοί clusters που προέβλεψε κάθε μέθοδος συσταδοποίησης. Τα αποτελέσματα περιλαμβάνουν και τα αρχικά δεδομένα (raw).

Table 5: Προβλεπόμενοι αριθμοί clusters για κάθε μέθοδο συσταδοποίησης. Οι μέθοδοι εφαρμόστηκαν και στα αρχικά δεδομένα και σε αυτά που προέκυψαν από τη μείωση διάστασης.

Αλγόριθμος Συσταδοποίησης	**raw**	**PCA**	**LDA**	**SAE**
MiniBatch KMeans	10	10	10	10
DBSCAN	1	1	8	24
Agglomerative	10	10	10	10

Παρατηρούμε ότι η DBSCAN αδυνατεί να διαχωρίσει τα raw δεδομένα και τα δεδομένα που έχουν υποστεί μείωση διάστασης με PCA (πάντα σε συνάρτηση με τις παραμέτρους που επιλέξαμε). Ωστόσο, η DBSCAN μέθοδος εφαρμοσμένη σε δεδομένα που υπέστησαν μείωση διάστασης με SAE μοντέλο, είναι η μόνη μέθοδος που προβλέπει σημαντικά περισσότερους clusters σε σχέση με τους αναμενόμενους (24). Αντίθετα, όλοι οι άλλοι αλγόριθμοι συσταδοποίησης καταφέρνουν να δημιουργήσουν τον αναμενόμενο αριθμό συστάδων.

Στην Εικόνα 4 απεικονίζονται οι τιμές των δεικτών αξιολόγησης των αλγορίθμων συσταδοποίησης. Ο οριζόντιος άξονας αφορά τους αλγορίθμους συσταδοποίησης, ενώ ο κάθετος άξονας τα δεδομένα που χρησιμοποιήθηκαν (raw, PCA transformed, κ.ο.κ.). Κάθε heatmap αντιστοιχεί σε έναν από τους 4 δείκτες αξιολόγησης που μελετήθηκαν. Κίτρινες τιμές στα heatmaps συνεπάγονται μεγάλες τιμές του εκάστοτε δείκτη, ενώ οι πιο σκοτεινές αποχρώσεις

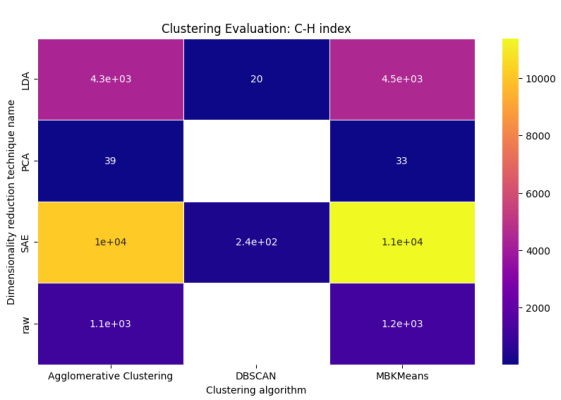
(προς το μωβ) αντιστοιχούν σε χαμηλές τιμές. Τα κενά κελιά (άσπρο χρώμα) αντιστοιχούν σε τιμές όπου ο δείκτης δεν υπολογίστηκε, καθώς η τεχνική συσταδοποίησης δε δημιούργησε παραπάνω από έναν clusters.

Εξετάζοντας τον C-H δείκτη, παρατηρούμε πως την καλύτερη απόδοση (υψηλές τιμές) τις έχει ο συνδυασμός MiniBatch KMeans με SAE και ακολουθεί ως δεύτερος ο συνδυασμός Agglomerative Clustering με SAE. Οι δύο αυτοί συνδυασμοί δημιουργούν τους καλύτερα ορισμένους clusters με το λιγότερο δυνατό overlap μεταξύ τους. Την χειρότερη απόδοση την παρουσιάζει η DBSCAN μέθοδος σε συνδυασμό με την LDA (λαμβάνοντας υπόψη τις περιπτώσεις που δημιούργησε πάνω από έναν cluster η DBSCAN).

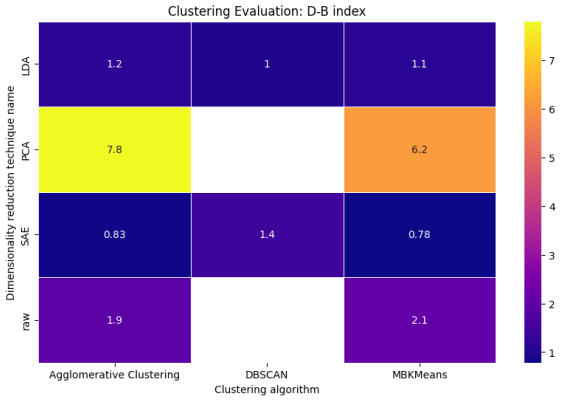
Όσον αφορά τον D-B δείκτη, την καλύτερη επίδοση (χαμηλές τιμές) την έχει και πάλι ο συνδυασμός MiniBatch KMeans με SAE, επιτυγχάνοντας τη δημιουργία clusters με λίγες ομοιότητες μεταξύ τους. Δεύτερος καλύτερος συνδυασμός σε επίδοση είναι και πάλι ο συνδυασμός Agglomerative Clustering με SAE. Ο χειρότερος συνδυασμός είναι Agglomerative clustering με PCA, οποίος δημιουργεί clusters με μεγάλη σχετικά διασπορά οδηγώντας, συνεπώς, σε overlapping.

Τα υπολογισμένα Silhouette scores υποδεικνύουν πως για ακόμη μια φορά ο συνδυασμός MiniBatch KMeans με SAE επιτυγχάνει την καλύτερη συσταδοποίηση των δεδομένων. Αντίθετα, οι αρνητικές τιμές που παρουσιάζουν οι συνδυασμοί Agglomerative με PCA, DBSCAN με LDA και MiniBatch KMeans με PCA συνεπάγονται την ύπαρξη παρατηρήσεων που δεν καταχωρούνται ξεκάθαρα και με βεβαιότητα σε κάποιον cluster, γεγονός που ενισχύει το overlapping μεταξύ clusters.

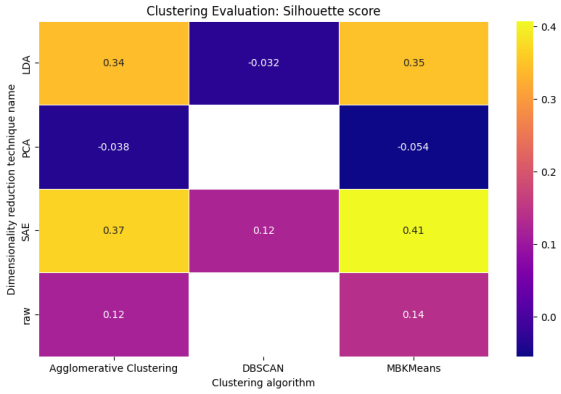
Τέλος, οι τιμές του δείκτη AR δείχνουν πως οι καλύτεροι συνδυασμοί είναι Agglomerative με LDA και MiniBatch KMeans με LDA. Οι δύο αυτοί συνδυασμοί επιτυγχάνουν μια καλή συμφωνία μεταξύ των πραγματικών κλάσεων και τον δημιουργημένων clusters. Η συσταδοποίηση σε δεδομένα που υπέστησαν μείωση διάστασης με SAE δεν είναι τόσο καλή, κρίνοντας μονάχα από τον AR δείκτη. Σαφώς, όπως είναι αναμενόμενο, η DBSCAN με τα raw δεδομένα και τα PCA δεδομένα, είναι η χειρότερη μέθοδος, καθώς ομαδοποιεί όλα τα δεδομένα στον ίδιο cluster.



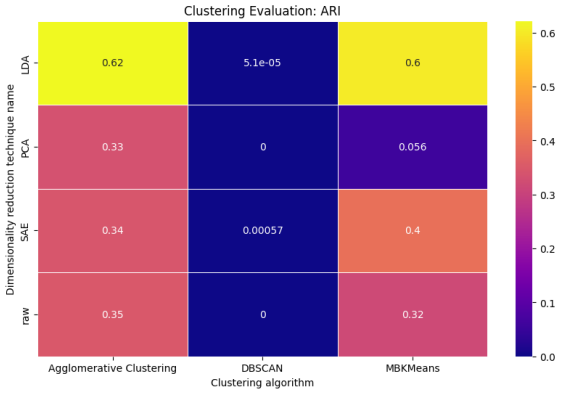
(a) C-H index



(b) D-B index



(c) Silhouette score



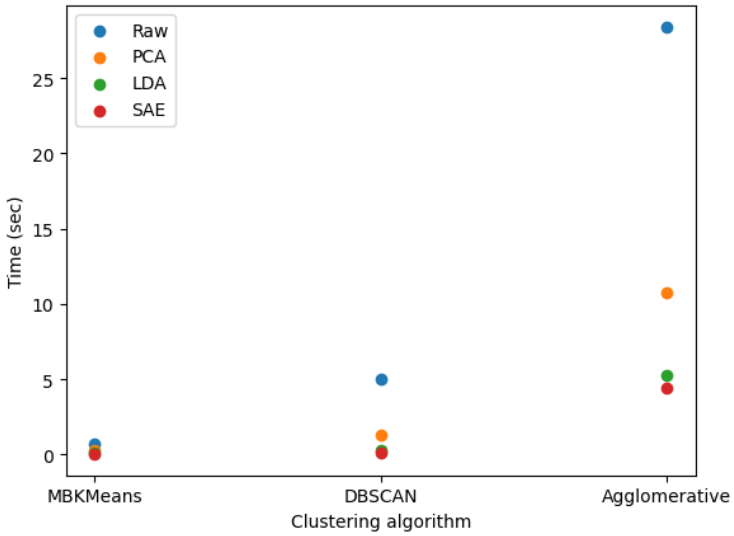
(d) ARI

Εικόνα 4: Heatmaps των μετρικών αξιολόγησης των διαφόρων αλγορίθμων συσταδοποίησης για κάθε τεχνική μείωσης διάστασης και τα raw δεδομένα.

Κλείνουμε την ανάλυση των αποτελεσμάτων με μια σύγκριση των χρόνων εκτέλεσης κάθε αλγορίθμου συσταδοποίησης, όπως παρουσιάζονται στην Εικόνα 5.

Ο ταχύτερος αλγόριθμος συσταδοποίησης είναι ο MiniBatch KMeans. Εικάζουμε πως στην ταχύτητά του βοήθησε και η γνώση των labels των δεδομένων, η οποία οδήγησε σε σωστή εκτίμηση του αρχικού αριθμού centroids. Ο Agglomerative αλγόριθμος είναι ο πιο αργός, κάτι που αναμένεται, καθώς απαιτεί $O(n^3)$ χρόνο, σε αντίθεση με τους υπόλοιπους αλγορίθμους.

Τέλος, παρατηρούμε πως σε κάθε αλγόριθμο συσταδοποίησης ο μικρότερος χρόνος εκτέλεσης αντιστοιχεί στην περίπτωση μείωσης διάστασης με SAE. Πιθανόν αυτό το αποτέλεσμα να οφείλεται στην πολύπλοκη αρχιτεκτονική τους οι οποίοι, σε αντίθεση με τις γραμμικές μεθόδους PCA και LDA, αξιοποιούν καλύτερα την πληροφορία που παρέχεται από τα δεδομένα και εξερευνούν και μη γραμμικές σχέσεις μεταξύ των features.



Εικόνα 5: Χρόνοι εκτέλεσης αλγορίθμων συσταδοποίησης πάνω στα raw δεδομένα και σε δεδομένα που υπέστησαν μείωση διάστασης. Ο χρόνος μετριέται σε δευτερόλεπτα.

4 Συμπεράσματα

Παρατηρήσαμε πως η μείωση διάστασης δεδομένων με Stacked Autoencoders ταξινομεί αρκετά καλά τα δεδομένα σε διακριτούς clusters, πρωτού καν εφαρμοστούν οι τεχνικές συσταδοποίησης και χωρίς να χρειαστούν πολλές εποχές εκπαίδευσης του μοντέλου. Αυτό σημαίνει πως ένας Stacked Autoencoder μαθαίνει πολύ γρήγορα τα δεδομένα του Fashion MNIST dataset και μπορεί να ομαδοποιήσει σωστά τις παρατηρήσεις. Ωστόσο, και η LDA μέθοδος δημιούργησε από μόνη της πολύ καλή ταξινόμηση των δεδομένων, κάτι που πιθανόν να οφείλεται στο γεγονός ότι στην εκπαίδευση χρησιμοποιεί και τα labels των δεδομένων.

Είδαμε, επιπλέον, πως ο μοναδικός συνδυασμός που προέβλεψε μεγαλύτερο αριθμό clusters από τον αριθμό των πραγματικών labels είναι ο DBSCAN με SAE. Ο συνδυασμός αυτός, επομένως, κατάφερε να εντοπίσει επιπλέον σχέσεις μεταξύ των δεδομένων, πέρα από αυτές που αναμέναμε να ανιχνευθούν.

Ο συνδυασμός MiniBatch KMeans με SAE είναι ο πιο αποδοτικός σε τρεις από τις τέσσερεις μετρικές αξιολόγησης που εξετάστηκαν. Στην περίπτωση του AR δείκτη, οι LDA μέθοδοι με MiniBatch KMeans και Agglomerative clustering είναι οι πιο αποδοτικοί συνδυασμοί. Η επικράτησή τους μονάχα στην περίπτωση του δείκτη AR ίσως οφείλεται στο γεγονός ότι ο AR αξιολογεί με βάση τα true labels και η LDA μέθοδος εκπαιδεύει τα δεδομένα χρησιμοποιώντας αυτή την πληροφορία.

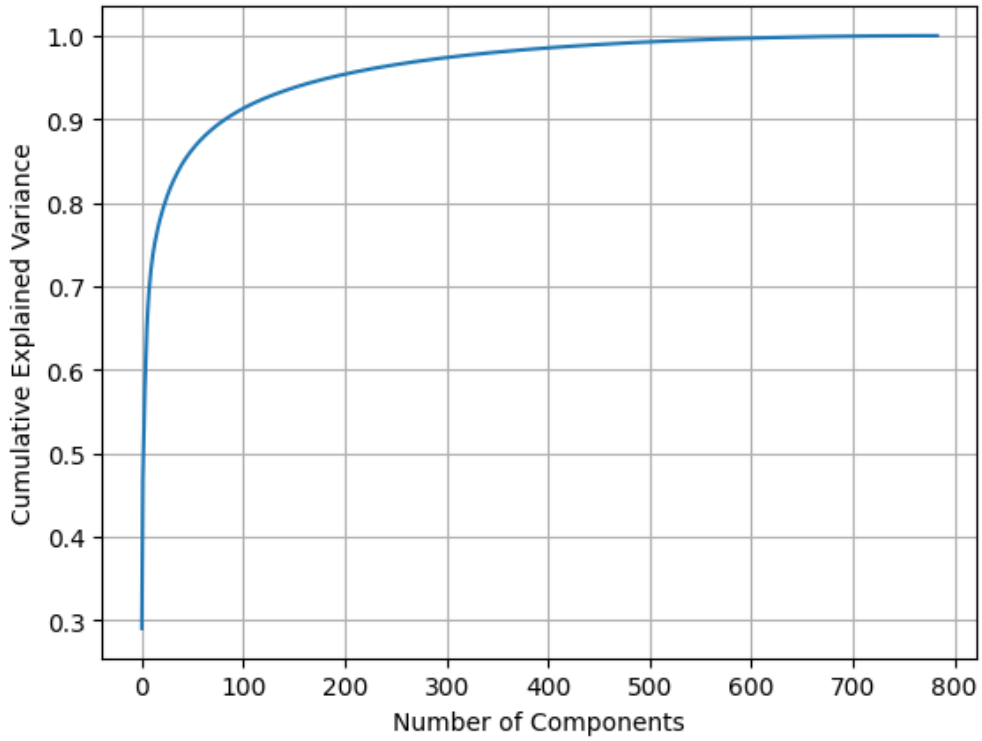
Τέλος, παρατηρήσαμε πως υπάρχουν περιπτώσεις όπου η χρήση raw δεδομένων μπορεί να επιφέρει καλύτερα αποτελέσματα από τη χρήση δεδομένων που έχουν υποστεί μείωση διάστασης. Αυτό φαίνεται στα Silhouette scores, όπου η χρήση raw δεδομένων επιφέρει καλύτερα αποτελέσματα από τη χρήση δεδομένων που προέρχονται από PCA.

Εν ολίγοις, παρά το μεγάλο χρόνο εκπαίδευσης, η μείωση διάστασης SAE μπορεί να οδηγήσει σε καλύτερη συσταδοποίηση των δεδομένων και καλύτερη απόδοση των αλγορίθμων συσταδοποίησης, απαιτώντας παράλληλα και λιγότερο χρόνο εκτέλεσης των αλγορίθμων αυτών.

Τα παραπάνω αποτελέσματα, βέβαια, δεν είναι robust, καθώς στην αναφορά μας απουσιάζει η στατιστική ανάλυση των αποτελεσμάτων. Για να γίνει αυτό θα χρειαζόταν να τρέξουμε τους αλγορίθμους αρκετές φορές σε διάφορα folds και να διεξαγάγουμε στατιστικό έλεγχο υποθέσεων. Επιπλέον, θα μπορούσαμε να εξετάσουμε έναν μεγαλύτερο αριθμό υπερπαραμέτρων τόσο για την κάθε τεχνική μείωσης διάστασης όσο και για τους αλγορίθμους συσταδοποίησης, ώστε να έχουμε μια καλύτερη εικόνα για το ποιοι συνδυασμοί αποδίδουν καλύτερα. Ένα τέτοιο εγχείρημα, ωστόσο, έχει και μεγάλο υπολογιστικό κόστος.

A Παράρτημα

Στο παράρτημα αυτό παρουσιάζονται οι υποψήφιοι παράμετροι για τις τεχνικές μείωσης διάστασης.



Εικόνα 6: Αθροιστική διασπορά συναρτήσει του αριθμού συνιστωσών. Το 0.95% της διασποράς περιγράφεται από 187 συνιστώσες, ενώ το 0.98% περιγράφεται από 348 συνιστώσες. Η RandomizedSearch CV επέστρεψε ως ιδανική παράμετρο 187 συνιστώσες.

Για την LDA μέθοδο εξετάστηκαν τρεις διαφορετικοί solvers, συγκεκριμένα οι Singular

Value Decomposition (svd), Least Squares (lsqr) και Eigendecomposition (eigen). Η RandomizedSearch CV επέλεξε ως καταλληλότερο solver τον svd.

Στους εναπομείναντες πίνακες περιγράφονται οι επιπλέον διαφορετικές αρχιτεκτονικές SAE που εξετάστηκαν.

Table 6: Αρχιτεκτονική SAE. Η loss function που ελαχιστοποιείται είναι η MSE.

Layer	**Νευρώνες**	**Activation Function**
Encoder 1	392	relu
Encoder 2	196	relu
Bottleneck	98	linear
Decoder 1	196	relu
Decoder 2	392	relu
Output	784	sigmoid

Table 7: Αρχιτεκτονική SAE. Η loss function που ελαχιστοποιείται είναι η binary cross entropy.

Layer	**Νευρώνες**	**Activation Function**
Encoder 1	392	relu
Encoder 2	196	relu
Bottleneck	98	linear
Decoder 1	196	relu
Decoder 2	392	relu
Output	784	sigmoid

Table 8: Αρχιτεκτονική SAE. Η loss function που ελαχιστοποιείται είναι η binary cross entropy.

Layer	**Νευρώνες**	**Activation Function**
Encoder 1	512	relu
Encoder 2	128	relu
Bottleneck	2	linear
Decoder 1	128	relu
Decoder 2	512	relu
Output	784	sigmoid

B Βιβλιογραφία

[1] Burkov, A. The Hundred-page machine learning book, 2019
[2] Witten, D.; James, G.M.; Hastie, T.; Tibshirani, R. An Introduction to Statistical Learning, 2023
[3] https://www.tensorflow.org/tutorials/generative/autoencoder
[4] https://www.joshcheema.io/post/2022-04-26-t-sne-on-the-mnist-dataset