Πανεπιστήμιο Πειραιώς

Τμήμα Ψηφιακών Συστημάτων

Π.Μ.Σ. «Πληροφοριακά Συστήματα και Υπηρεσίες»

Κατεύθυνση: «Μεγάλα Δεδομένα και Αναλυτική»

Προσωποποιημένες Συνεργατικές Μέθοδοι Συστάσεων

Διπλωματική Εργασία

ΜΕΛΙΝΑ ΑΠΟΣΤΟΛΙΔΟΥ

Επιβλέπουσα:

ΜΑΡΙΑ ΧΑΛΚΙΔΗ

Αναπληρώτρια Καθηγήτρια

Αθήνα, Ιούνιος 2021

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω την επιβλέπουσα καθηγήτρια της διπλωματικής μου Καθηγήτρια κ. Μαρία Χαλκίδη για την ευκαιρία που μου έδωσε να ασχοληθώ με ένα τόσο ενδιαφέρον θέμα, καθώς και για την καθοδήγηση και την βοήθεια που μου παρείχε καθ' όλη τη διάρκεια της εκπόνησης αυτής της διπλωματικής εργασίας.

Επιπλέον, είμαι ευγνώμων στα υπόλοιπα μέλη της εξεταστικής επιτροπής της διπλωματικής εργασίας τους Καθηγητές κ. Χρήστο Δουλκερίδη και κ. Μιχαήλ Φιλιππάκη για την προσεκτική ανάγνωση της εργασίας μου.

Τέλος, είμαι ευγνώμων στην οικογένειά μου για την συμπαράσταση τους και την στήριξη που μου παρείχαν καθ' όλη τη διάρκεια των σπουδών μου.

Περίληψη

Η παρούσα διπλωματική εργασία έχει ως στόχο την παρουσίαση της συνεργατικής μάθησης ως έναν κατάλληλο τρόπο δημιουργίας μοντέλων μηχανικής μάθησης όπου τα δεδομένα των χρηστών προστατεύονται ακόμα και αν βρίσκονται κατανεμημένα σε απομακρυσμένες συσκευές. Επιπλέον, παρουσιάζονται τα συστήματα συστάσεων και ο τρόπος με τον οποίο εκπαιδεύονται και αποφασίζουν τις κατάλληλες προτάσεις για τον κάθε χρήστη. Στη συνέχεια, περιγράφεται το θεωρητικό πλαίσιο της υλοποίησης που έγινε, η οποία αφορά στην δημιουργία ενός συνεργατικού συστήματος συστάσεων το οποίο αποτελείται από ένα νευρωνικό δίκτυο ενσωμάτωσης το οποίο θα εκπαιδευτεί με συνεργατικό τρόπο. Το σύνολο δεδομένων που χρησιμοποιήθηκε αποτελείται από τις αξιολογήσεις που έδωσαν ορισμένοι χρήστες σε έναν αριθμό ταινιών. Μετά την εκπαίδευση το μοντέλο είναι σε θέση να προβλέπει την αξιολόγηση που είναι πιθανό να έδινε ένας χρήστης σε μία συγκεκριμένη ταινία ώστε αν είναι αρκετά υψηλή να την προτείνει σε αυτόν τον χρήστη για παρακολούθηση. Επιπλέον, χρησιμοποιείται μία προσωποποιημένη προσέγγιση όπου υπολογίζεται ο βαθμός στον οποίο το τοπικό μοντέλο του κάθε χρήστη αναμιγνύεται με το κεντρικό μοντέλο με στόχο τη βελτιστοποίηση της απόδοσης του μοντέλου.

Abstract

This thesis presents federated learning as an appropriate way to create machine learning models where user data is protected even if it is distributed on remote devices. In addition, there are presented the recommender systems and the way in which they are trained and decide the appropriate proposals for each user. Then, it is described the theoretical framework of the implementation which concerns the creation of a federated recommender system that consists of an embedding neural network that will be trained in a federated way. The data set that was used consists of the ratings given by some users to a number of movies. After the training, the model is able to predict the rating that a user is likely to give to a particular movie, so that if it is high enough the movie will be recommended to this user to watch. Furthermore, a personalized approach is used to calculate the degree to which each user's local model is mixed with the global model in order to optimize the performance of the model.

Περιεχόμενα

Κατάλογος Σχημάτων

Κεφάλαιο 1: Εισαγωγή

1.1 Αντικείμενο Διπλωματικής

Ο τομέας της μηχανικής μάθησης εξελίσσεται συνεχώς και καλείται να αντιμετωπίσει όλο και περισσότερες προκλήσεις. Μία βασική πρόκληση είναι ο όλο και αυξανόμενος όγκος δεδομένων που παράγεται σε απομακρυσμένες συσκευές και η ανάγκη επεξεργασίας σε πολλαπλά μηχανήματα είναι μεγάλη. Ένα άλλο βασικό πρόβλημα όσον αφορά την ανάλυση και επεξεργασία δεδομένων είναι η προστασία των ευαίσθητων προσωπικών δεδομένων των χρηστών. Για να εκπαιδευτούν τα μοντέλα μηχανικής μάθησης απαιτείται μία επαρκής ποσότητα δεδομένων των χρηστών. Καθώς το απόρρητο και η ασφάλεια των προσωπικών δεδομένων είναι πλέον κρίσιμης σημασίας αναπτύσσονται νέες τεχνικές ώστε να αντιμετωπιστεί αυτό το πρόβλημα.

Η προσέγγιση της συνεργατικής μάθησης αποτελεί μία αποτελεσματική λύση σε αυτά τα προβλήματα. Συγκεκριμένα, επιτρέπει την δημιουργία ενός κεντρικού μοντέλου μηχανικής μάθησης το οποίο εκπαιδεύεται με τα δεδομένα που βρίσκονται σε κάθε συσκευή χωρίς αυτά να δεσμεύονται σε κάποιον κεντρικό διακομιστή. Αυτό επιτυγχάνεται καθώς το μοντέλο εκπαιδεύεται σε κάθε συσκευή τοπικά με τα δεδομένα που βρίσκονται αποθηκευμένα εκεί και στο κεντρικό μοντέλο στέλνονται μόνο οι ενημερώσεις των τοπικών μοντέλων.

Αντικείμενο της παρούσας διπλωματικής εργασίας αποτελεί η δημιουργία συστημάτων συστάσεων που εκπαιδεύονται με τεχνικές συνεργατικής μάθησης και θα προτείνουν στους χρήστες ταινίες που είναι πιθανό να επιθυμούν να παρακολουθήσουν. Αρχικά περιγράφεται η έννοια της συνεργατικής μάθησης αλλά και των συστημάτων συστάσεων. Στη συνέχεια παρουσιάζονται λεπτομερώς οι τεχνικές που χρησιμοποιήθηκαν για την υλοποίηση της εργασίας αυτής. Τέλος, περιγράφεται η πειραματική μελέτη που έγινε μαζί με τα αποτελέσματα και τα συμπεράσματα που προέκυψαν.

1.2 Οργάνωση Κειμένου

Η διπλωματική εργασία αποτελείται από τις παρακάτω θεματικές ενότητες:

- Στο Κεφάλαιο 2 τίθενται οι θεωρητικές βάσεις που αφορούν το αντικείμενο της διπλωματικής.

- Στο Κεφάλαιο 3 παρουσιάζονται οι τεχνικές που αξιοποιήθηκαν για την υλοποίηση της εργασίας μαζί με τους μαθηματικούς ορισμούς των εννοιών που χρησιμοποιήθηκαν.

- Στο Κεφάλαιο 4 παρουσιάζεται η πειραματική μελέτη του μοντέλου που κατασκευάστηκε στο πλαίσιο των συνεργατικών συστημάτων συστάσεων με στόχο την δημιουργία προτάσεων σε κάθε χρήστη όσον αφορά τις ταινίες. Επιπλέον, παρουσιάζονται και αναλύονται τα αποτελέσματα που προέκυψαν.

- Τέλος, στο Κεφάλαιο 5 περιλαμβάνονται η σύνοψη και τα συμπεράσματα που εξήχθησαν κατά την εκπόνηση αυτής της διπλωματικής και προτείνονται πιθανές μελλοντικές επεκτάσεις.

Κεφάλαιο 2: Συνεργατική Μάθηση και Συστήματα Συστάσεων: Θεωρητικό Υπόβαθρο

2.1 Εισαγωγή

Σε αυτό το κεφάλαιο παρουσιάζεται αρχικά η συνεργατική μάθηση και συγκεκριμένα περιγράφονται λεπτομερώς οι ανάγκες που οδηγούν στην δημιουργία μοντέλων συνεργατικής μάθησης, πως ορίζονται αυτά τα μοντέλα και διάφορες προκλήσεις που μπορεί να προκύψουν. Στη συνέχεια περιγράφονται τα συστήματα συστάσεων, τα βασικά κίνητρα για τη δημιουργία τους και διάφοροι τύποι συστημάτων συστάσεων.

2.2 Συνεργατική Μάθηση

2.2.1 Εισαγωγή

Όπως αναφέρεται στο [1], οι κινητές συσκευές αποτελούν πλέον την κύρια υπολογιστική πηγή για δισεκατομμύρια χρήστες παγκοσμίως και δισεκατομμύρια ακόμα συσκευές IoT αναμένεται να συνδεθούν διαδικτυακά τα επόμενα χρόνια. Τα κινητά τηλέφωνα, οι φορητές συσκευές και τα αυτόνομα οχήματα είναι μόνο μερικά από τα σύγχρονα κατανεμημένα δίκτυα τα οποία παράγουν μία τεράστια ποσότητα πολύτιμων δεδομένων καθημερινά. Για τη δημιουργία εξυπνότερων εφαρμογών και υπηρεσιών για τις συσκευές αυτές είναι όλο και αυξανόμενη η χρήση τεχνικών μηχανικής μάθησης που χρησιμοποιούν αυτά τα δεδομένα για την εκπαίδευσή τους. Ωστόσο, η πλειονότητα των τεχνικών μηχανικής μάθησης απαιτούν μία μεγάλη ποσότητα δεδομένων από τον χρήστη, με ευαίσθητες και απόρρητες πληροφορίες που συγκεντρώνονται σε έναν κεντρικό διακομιστή για να εκπαιδευτεί το μοντέλο. Αυτό μπορεί να είναι αδύνατο ή ανεπιθύμητο από την άποψη του απόρρητου και της ασφάλειας. Επομένως, οι προσεγγίσεις που διατηρούν τα δεδομένα στη συσκευή και μοιράζονται το μοντέλο γίνονται ολοένα και πιο δημοφιλείς.

Η έννοια του edge computing δεν είναι καινούρια. Πράγματι, ο υπολογισμός απλών ερωτημάτων σε κατανεμημένες και χαμηλής ισχύος συσκευές είναι μία έρευνα που

διερευνάται εδώ και δεκαετίες στο πλαίσιο της επεξεργασίας ερωτημάτων σε δίκτυα αισθητήρων και edge computing. Σε αρκετές πρόσφατες εργασίες έχει μελετηθεί η εκπαίδευση μοντέλων μηχανικής μάθησης που γίνεται κεντρικά, ενώ η αποθήκευση δεδομένων γίνεται τοπικά σε κάθε συσκευή.

Ωστόσο, καθώς αυξάνεται η υπολογιστική ισχύς και η αποθηκευτική ικανότητα των συσκευών σε κατανεμημένα δίκτυα υπάρχει η δυνατότητα αξιοποίησης των βελτιωμένων τοπικών πόρων της κάθε συσκευής. Έτσι, προτάθηκε πρόσφατα από την Google ένα παράδειγμα αποκεντρωμένης μηχανικής μάθησης που ονομάζεται Συνεργατική Μάθηση (Federated Learning). Η συνεργατική μάθηση επιτρέπει σε κινητές συσκευές (π.χ. κινητά τηλέφωνα, οχήματα) να εκπαιδεύουν συνεργατικά ένα κεντρικό μοντέλο με αποκεντρωμένο τρόπο χωρίς να απαιτείται η αποθήκευση ακατέργαστων δεδομένων εκπαίδευσης σε έναν κεντρικό διακομιστή. Πιο συγκεκριμένα, η εκπαίδευση στατιστικών μοντέλων γίνεται απευθείας στις απομακρυσμένες συσκευές.

Πολλοί μεγάλοι πάροχοι υπηρεσιών αναπτύσσουν πλέον τεχνικές συνεργατικής μάθησης, καθώς είναι απαραίτητες στην ανάπτυξη εφαρμογών όπου πρέπει να προστατευθούν τα ευαίσθητα προσωπικά δεδομένα των χρηστών, τα οποία είναι κατανεμημένα στην άκρη.

Υπάρχουν αρκετά παραδείγματα εφαρμογών συνεργατικής μάθησης, όπως για παράδειγμα στα κινητά τηλέφωνα όπου μπορούν να δημιουργηθούν στατιστικά μοντέλα τα οποία καθώς εκπαιδεύονται από τις συμπεριφορές πολλών διαφορετικών χρηστών, σε ένα δίκτυο κινητών συσκευών, μπορούν να προβλέπουν την επόμενη λέξη που θέλει να γράψει ο χρήστης. Εντούτοις, πολλοί χρήστες δεν είναι διατεθειμένοι να μοιραστούν τα προσωπικά τους δεδομένα, ώστε να προστατέψουν την ιδιωτικότητά τους. Η συνεργατική μάθηση δίνει τη δυνατότητα να κατασκευάζονται τέτοια στατιστικά μοντέλα χωρίς να δεσμεύονται αυτά τα δεδομένα σε έναν κεντρικό διακομιστή, αλλά και χωρίς να μειώνεται η απόδοσή του μοντέλου.

Άλλο παράδειγμα όπου η συνεργατική μάθηση μπορεί να παίξει σημαντικό ρόλο είναι η ανάπτυξη εφαρμογών σε οργανισμούς με ευαίσθητα προσωπικά δεδομένα. Συγκεκριμένα, τα νοσοκομεία προστατεύουν το ιατρικό απόρρητο των ασθενών και απαγορεύεται να διαρρεύσουν οποιαδήποτε δεδομένα, γεγονός που καθιστά δύσκολη την δημιουργία ιατρικών εφαρμογών χωρίς τις τεχνικές της συνεργατικής μάθησης.

Επιπλέον, μία άλλη κατηγορία είναι οι συσκευές IoT (Internet of Things), όπως τα αυτόνομα οχήματα και τα έξυπνα σπίτια που μπορεί να περιέχουν αισθητήρες μέσω των οποίων να συλλέγουν ευαίσθητα προσωπικά δεδομένα. Για παράδειγμα, ένα αυτόνομο όχημα χρειάζεται να συλλέγει δεδομένα που αφορούν την συμπεριφορά των πεζών. Επομένως, η χρήση των μεθόδων συνεργατικής μάθησης σε τέτοιου είδους συσκευές μπορεί να αποτελέσει τη λύση σε αυτό το πρόβλημα.

Ένα παράδειγμα εφαρμογής της συνεργατικής μάθησης σε κινητές συσκευές, είναι η πρόβλεψης της επόμενης λέξης που θέλει να πληκτρολογήσει ο χρήστης. Αντί να αποστέλλονται σε κάποιον κεντρικό διακομιστή τα μη επεξεργασμένα δεδομένα των χρηστών, η εκπαίδευση του μοντέλου γίνεται με κατανεμημένο τρόπο προκειμένου να διατηρηθεί το απόρρητο των κειμένων που πληκτρολογούν οι χρήστες. Στην εικόνα 1 φαίνεται η επικοινωνία των απομακρυσμένων συσκευών με τον κεντρικό διακομιστή με περιοδικό τρόπο ώστε να εκπαιδευτεί ένα κεντρικό μοντέλο. Σε κάθε γύρο επικοινωνίας ένα υποσύνολο επιλεγμένων κινητών τηλεφώνων εκτελεί τοπικές εκπαιδεύσεις στα μη πανομοιότυπα κατανεμημένα δεδομένα των χρηστών και αποστέλλει τις τοπικές ενημερώσεις στον διακομιστή. Ο διακομιστής αφού ενσωματώσει αυτές τις ενημερώσεις στέλνει πίσω το νέο κεντρικό μοντέλο σε ένα άλλο υποσύνολο συσκευών. Αυτή η επαναληπτική διαδικασία εκπαίδευσης συνεχίζεται σε όλο το δίκτυο έως ότου επιτευχθεί η σύγκλιση ή ικανοποιηθεί κάποιο κριτήριο διακοπής.

Εικόνα 1. Επικοινωνία απομακρυσμένων συσκευών με τον κεντρικό διακομιστή για την πρόβλεψη επόμενης λέξης

2.2.2 Βασικές Προκλήσεις

Κατά την δημιουργία ενός μοντέλου συνεργατικής μάθησης παρουσιάζονται αρκετές προκλήσεις που πρέπει να ληφθούν υπόψη και να αντιμετωπιστούν. Παρακάτω, παρουσιάζονται οι βασικότερες:

- **Ακριβή Επικοινωνία.** Η διαδικασία της επικοινωνίας μεταξύ του κεντρικού διακομιστή και των απομακρυσμένων συσκευών μπορεί να είναι μία ακριβή διαδικασία, λαμβάνοντας υπόψη ότι τα δεδομένα που παράγονται σε κάθε συσκευή παραμένουν σε αυτή. Τα δίκτυα στα

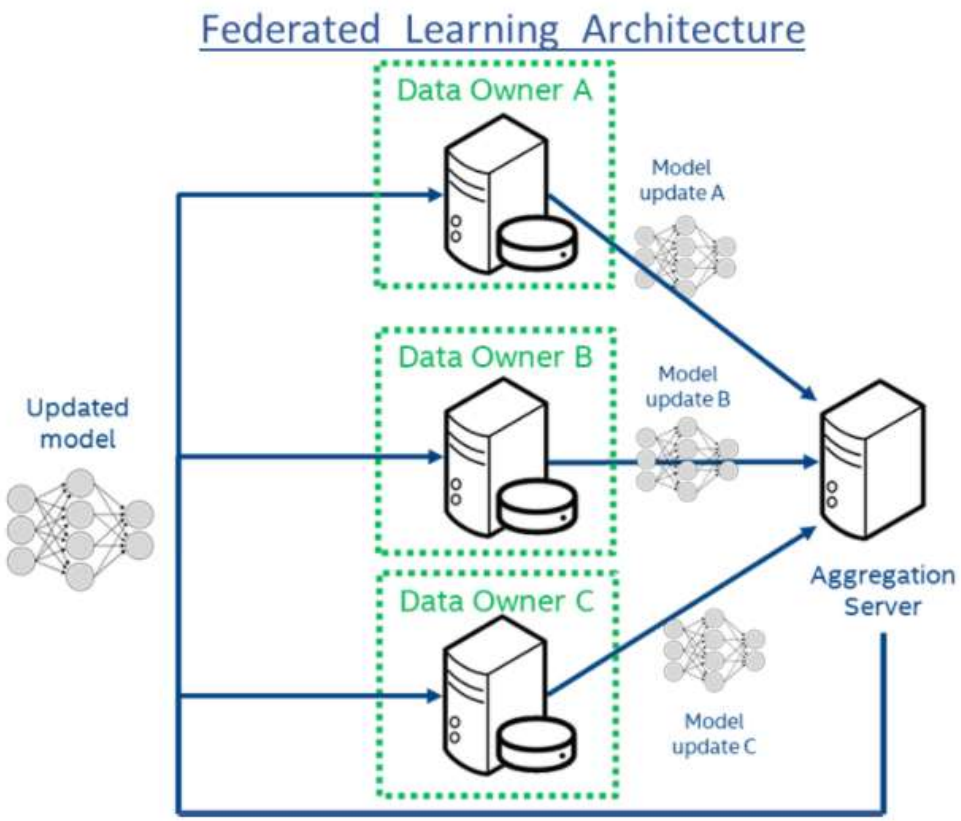
οποία μπορεί να εφαρμοστούν οι τεχνικές της συνεργατικής μάθησης μπορεί να αποτελούνται από εκατομμύρια κινητά τηλέφωνα. Επομένως, η επικοινωνία της κάθε συσκευής με τον κεντρικό διακομιστή μπορεί να καθυστερεί περισσότερο την διαδικασία της εκπαίδευσης από ότι αν συγκεντρώνονταν όλα τα δεδομένα στον κεντρικό διακομιστή και εκπαιδεύονταν εκεί. Για την εκπαίδευση ενός μοντέλου με δεδομένα τα οποία παράγονται από απομακρυσμένες συσκευές, είναι σημαντική η δημιουργία τεχνικών βελτίωσης της επικοινωνίας, όπως η αποστολή μικρών ενημερώσεων του μοντέλου από την κάθε συσκευή στον κεντρικό διακομιστή. Αν χρειαστεί μεγαλύτερη βελτίωση της επικοινωνίας οι μέθοδοι που θα πρέπει να χρησιμοποιηθούν είναι η μείωση του αριθμού των επαναλήψεων επικοινωνίας καθώς και η μείωση του μεγέθους των ενημερώσεων που μεταφέρονται σε κάθε επανάληψη.

- **Ετερογένεια των Συστημάτων.** Σε ένα δίκτυο ενδεχομένως να υπάρχει μεγάλη διαφοροποίηση μεταξύ των συσκευών όσον αφορά το hardware (επεξεργαστής, μνήμη), τη συνδεσιμότητα του δικτύου (wifi) και την ισχύ (επίπεδο μπαταρίας). Συνεπώς, οι περιορισμοί αυτοί που μπορεί να έχει η κάθε συσκευή, καθώς και το μέγεθος του δικτύου μπορεί να έχουν ως συνέπεια ένα μόνο ποσοστό των συσκευών του δικτύου να είναι ταυτόχρονα ενεργό και να μπορεί να συμμετέχει στην εκπαίδευση (π.χ. 0,01%). Επίσης, είναι αρκετά συχνό μία συσκευή να σταματήσει να είναι ενεργή κατά τη διάρκεια της εκπαίδευσης λόγω κακής συνδεσιμότητας ή χαμηλής ισχύος. Συμπερασματικά, σε ένα μοντέλο συνεργατικής μάθησης θα πρέπει να ληφθεί υπόψη ότι το ποσοστό συμμετοχής στην εκπαίδευση ενδέχεται να είναι χαμηλό και θα πρέπει να έχει ανοχή στην ετερογένεια των συστημάτων των συσκευών, όπως και στις συσκευές που σταματούν να είναι ενεργές κατά τη διάρκεια της εκπαίδευσης.
- **Στατιστική Ετερογένεια.** Σε ένα δίκτυο οι συσκευές μπορεί να παράγουν και να αποθηκεύουν δεδομένα μη πανομοιότυπα κατανεμημένα, για παράδειγμα μπορεί τα χαρακτηριστικά των χρηστών των κινητών συσκευών να ποικίλουν. Επίσης, μπορεί να υπάρχει μεγάλη απόκλιση μεταξύ των συσκευών όσων αφορά την ποσότητα των δεδομένων που παράγουν. Σε αυτή την περίπτωση τα δεδομένα είναι μη ανεξάρτητα και πανομοιότυπα κατανεμημένα (non I.I.D.), γεγονός που μπορεί να αυξήσει την πολυπλοκότητα στην επίλυση ενός προβλήματος μηχανικής μάθησης. Συνεπώς, για την διαχείριση της στατιστικής ετερογένειας των δεδομένων κατά τη δημιουργία ενός μοντέλου συνεργατικής μάθησης συχνά απαιτούνται διαφορετικές τεχνικές όπως η δημιουργία προσωποποιημένων μοντέλων.
- **Ζητήματα Απορρήτου.** Η συνεργατική μάθηση αποτελεί μία προσπάθεια προστασίας των ευαίσθητων προσωπικών δεδομένων

που παράγονται από την κάθε συσκευή. Αυτό επιτυγχάνεται με την δημιουργία ενός κεντρικού μοντέλου σε έναν κεντρικό διακομιστή, στον οποίο αποστέλλονται ενημερώσεις από την κάθε συσκευή στην οποία το μοντέλο έχει εκπαιδευτεί με τα τοπικά δεδομένα κ έτσι δεν αποστέλλονται ακατέργαστα δεδομένα. Ωστόσο, αυτή η μεταφορά των ενημερώσεων ενδέχεται να αποκαλύψει απόρρητες πληροφορίες είτε στον κεντρικό διακομιστή ή σε άλλους χρήστες που συμμετέχουν στην εκπαίδευση. Οι ήδη υπάρχουσες τεχνικές για τη βελτίωση της προστασίας του απορρήτου στη συνεργατική μάθηση, όπως για παράδειγμα το διαφορικό απόρρητο, ενδέχεται να οδηγήσουν σε μειωμένη απόδοση του μοντέλου. Επομένως, η εύρεση μεθόδων για την κατασκευή μοντέλων συνεργατικής μάθησης αποτελεί μία βασική πρόκληση.

2.2.3 Αρχιτεκτονική Μοντέλου Συνεργατικής Μάθησης

Στην εικόνα 2 φαίνεται η αρχιτεκτονική ενός παραδείγματος συνεργατικής μάθησης.



Εικόνα 2. Αρχιτεκτονική μοντέλου συνεργατικής μάθησης

Στην παραπάνω εικόνα απεικονίζεται ο κεντρικός διακομιστής ως Aggregation Server καθώς εκεί συγκεντρώνονται όλες οι ενημερώσεις από τα τοπικά μοντέλα. Στο παράδειγμα αυτό απεικονίζονται τρεις χρήστες ως Data Owner A, Data Owner B και Data Owner C, όπου στις συσκευές τους βρίσκονται αποθηκευμένα τα τοπικά δεδομένα του κάθε χρήστη. Το μοντέλο απεικονίζεται ως ένα νευρωνικό δίκτυο το οποίο αποστέλλεται από τον κεντρικό διακομιστή στους χρήστες. Εκεί, αφού εκπαιδευτούν τα τοπικά μοντέλα, η κάθε συσκευή στέλνει πίσω στον διακομιστή τις ενημερώσεις των τοπικά εκπαιδευμένων μοντέλων. Στη συνέχεια, αφού συγκεντρωθούν οι ενημερώσεις υπολογίζεται το νέο ενημερωμένο κεντρικό μοντέλο το οποίο αποστέλλεται εκ νέου στον κάθε χρήστη ώστε να επαναληφθεί η διαδικασία. Η συνεργατική εκπαίδευση τερματίζεται όταν ικανοποιηθεί κάποια συνθήκη ή ολοκληρωθούν οι γύροι επικοινωνίας που έχουν οριστεί.

2.3 Συστήματα Συστάσεων

2.3.1 Κίνητρα Δημιουργίας Συστημάτων Συστάσεων

Σύμφωνα με το [2], η συνεχής αύξηση του καταναλωτισμού σε συνδυασμό με την συμμετοχή του διαδικτύου στο εμπόριο έχει σαν αποτέλεσμα ο κάθε καταναλωτής να έρχεται αντιμέτωπος με μία πληθώρα προϊόντων και υπηρεσιών που είναι διαθέσιμα προς κατανάλωση. Καθώς οι συστάσεις που δέχεται κάθε άνθρωπος παίζουν σημαντικό ρόλο στη λήψη αποφάσεων, οι πωλητές καλούνται να δημιουργήσουν προσωποποιημένες συστάσεις για τον κάθε πιθανό πελάτη.

Επίσης, είναι πλέον γεγονός πως όλοι οι μεγάλοι οργανισμοί και επιχειρήσεις συλλέγουν μεγάλο όγκο δεδομένων που αφορά τους πελάτες τους και τον τρόπο που αλληλεπιδρούν με τα προϊόντα, με στόχο την περαιτέρω επεξεργασία και ανάλυση αυτών των δεδομένων. Μέσα από αυτή την ανάλυση δημιουργούνται τα συστήματα συστάσεων ώστε να εξυπηρετηθούν καλύτερα οι ανάγκες τόσο των επιχειρήσεων με την αύξηση των πωλήσεών τους όσο και των καταναλωτών με την ικανοποίηση τους για τα προϊόντα που επιθυμούν.

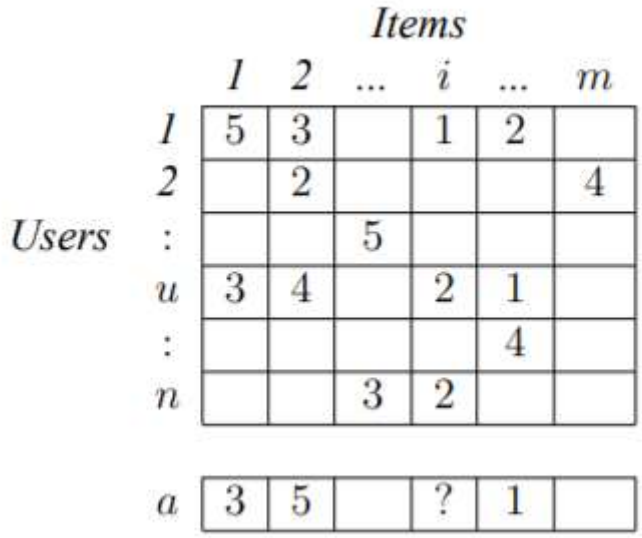
2.3.2 Γενικά

Τα συστήματα συστάσεων επεξεργάζονται δεδομένα σχετικά με τις προτιμήσεις των χρηστών για κάποια συγκεκριμένα αντικείμενα, όπως για παράδειγμα ταινίες, τραγούδια, ρούχα κλπ. Η συλλογή αυτών των πληροφοριών μπορεί να γίνει είτε άμεσα, δηλαδή από αξιολογήσεις χρηστών για τα προϊόντα, ή έμμεσα από τη

συμπεριφορά των χρηστών απέναντι στα προϊόντα, για παράδειγμα οι ταινίες που επιλέγει να δει και τα τραγούδια που επιλέγει να ακούσει. Τα συστήματα συστάσεων επεξεργάζοντας αυτές τις πληροφορίες προβλέπουν και προτείνουν στο χρήστη προϊόντα που είναι πιθανό να επιθυμεί.

2.3.3 Δομή των Συστημάτων Συστάσεων

Το γενικότερο πλαίσιο μελέτης των συστημάτων συστάσεων που μελετάται φαίνεται στην εικόνα 3.



Εικόνα 3. Πίνακας αξιολογήσεων χρηστών όπου κάθε κελί $r_{u,i}$ αντιστοιχεί στην αξιολόγηση του αντικειμένου i από τον χρήστη u. Ο στόχος είναι η πρόβλεψη της αξιολόγησης $r_{a,i}$ για τον ενεργό χρήστη a.

Οι προτιμήσεις των χρηστών που είναι γνωστές, αναπαριστώνται ως ένας πίνακας n χρηστών και m αντικειμένων, όπου κάθε κελί $r_{u,i}$ αντιστοιχεί στην αξιολόγηση του αντικειμένου m από τον χρήστη n. Αυτός ο πίνακας είναι συνήθως αραιός, καθώς οι περισσότεροι χρήστες δεν αξιολογούν τα περισσότερα αντικείμενα. Ο στόχος είναι να γίνει η πρόβλεψη για το ποια αξιολόγηση θα έδινε ένας χρήστης σε ένα αντικείμενο που προηγουμένως δεν είχε αξιολογήσει. Συνήθως, γίνονται οι προβλέψεις των αξιολογήσεων για όλα τα αντικείμενα που δεν έχουν παρατηρηθεί από έναν χρήστη και τα αντικείμενα με την υψηλότερη αξιολόγηση παρουσιάζονται ως προτάσεις. Ο χρήστης για τον οποίο υπολογίζονται οι συστάσεις αναφέρεται ως ενεργός χρήστης.

Οι προσεγγίσεις στα συστήματα συστάσεων μπορούν να κατηγοριοποιηθούν ως εξής:

- **Συνεργατικό Φιλτράρισμα.** Στο συνεργατικό φιλτράρισμα προτείνονται σε έναν χρήστη αντικείμενα με βάση τις προηγούμενες αξιολογήσεις όλων των χρηστών συλλογικά.
- **Σύσταση Βάσει Περιεχομένου.** Σε αυτές τις προσεγγίσεις προτείνονται σε ένα χρήστη αντικείμενα που είναι παρόμοια ως προς το περιεχόμενο με αντικείμενα που άρεσαν στον συγκεκριμένο χρήστη στο παρελθόν ή ταιριάζουν με τα χαρακτηριστικά του χρήστη.
- **Υβριδικές προσεγγίσεις.** Αυτές οι προσεγγίσεις συνδυάζουν το συνεργατικό φιλτράρισμα με τις συστάσεις βάσει περιεχομένου.

2.3.4 Συνεργατικό Φιλτράρισμα

Στα συστήματα συνεργατικού φιλτραρίσματος τα δεδομένα είναι αξιολογήσεις αντικειμένων από χρήστες και η επεξεργασία τους αφορά την ανάλυση της ομοιότητας που παρουσιάζουν οι χρήστες στη συμπεριφορά τους καθώς αξιολογούν τα αντικείμενα. Ο τελικός στόχος είναι η εύρεση ενός αποτελεσματικού τρόπου σύστασης των αντικειμένων. Υπάρχουν δύο προσεγγίσεις συνεργατικού φιλτραρίσματος, μία που βασίζεται στη μνήμη και μία που βασίζεται στο μοντέλο.

Συνεργατικό Φιλτράρισμα με βάση τη μνήμη

Στη μέθοδο συνεργατικού φιλτραρίσματος με βάση τη μνήμη προκειμένου να γίνουν προβλέψεις για έναν συγκεκριμένο χρήστη επιλέγεται ένα σύνολο χρηστών που παρουσιάζουν αρκετή ομοιότητα με αυτόν και μέσω ενός συνδυασμού των αξιολογήσεών τους και με κάποια βάρη υπολογίζεται το ζητούμενο αποτέλεσμα. Παρακάτω παρουσιάζονται συνοπτικά τα βήματα αυτής της μεθόδου:

1. Ορίζεται ένα βάρος για όλους τους χρήστες, το οποίο εκφράζει την ομοιότητα με τον αρχικό χρήστη για τον οποίο θα γίνουν οι προβλέψεις.
2. Επιλέγεται ένα σύνολο k χρηστών που παρουσιάζουν την μεγαλύτερη ομοιότητα με τον αρχικό χρήστη. Το σύνολο αυτό μπορεί να ονομαστεί και ως «γειτονιά».

Γίνεται η πρόβλεψη μέσω του συνδυασμού των αξιολογήσεων και των βαρών των k χρηστών που επιλέχθηκαν.

Το βάρος $w_{a,u}$ είναι το μέτρο ομοιότητας μεταξύ του χρήστη a (ενεργός χρήστης) και του χρήστη u (χρήστης από το σύνολο που επιλέχθηκε). Ο συντελεστής συσχέτισης Pearson αποτελεί ένα μέτρο ομοιότητα μεταξύ δύο χρηστών και ορίζεται παρακάτω:

$$w_{a,u} = \frac{\Sigma_{i \in I}(r_{a,i} - \bar{r}_a)(r_{u,i} - \bar{r}_u)}{\sqrt{\Sigma_{i \in I}(r_{a,i} - \bar{r}_a)^2 \; \Sigma_{i \in I}(r_{u,i} - \bar{r}_u)^2}} \quad (1)$$

όπου I είναι το σύνολο των αντικειμένων που έχουν βαθμολογηθεί και από τους δύο χρήστες, $r_{u,i}$ είναι η αξιολόγηση του αντικειμένου i από τον χρήστη u και $\bar{r}_u$ είναι η μέση αξιολόγηση που έχει γίνει από τον χρήστη u.

Η πρόβλεψη της αξιολόγησης του αντικειμένου i από έναν χρήστη a υπολογίζεται από την παρακάτω εξίσωση:

$$p_{a,i} = \bar{r}_a + \frac{\Sigma_{u \in K}\,(r_{u,i} - \bar{r}_u)\; x \; w_{a,u}}{\Sigma_{u \in K} \; w_{a,u}} \quad (2)$$

όπου $p_{a,i}$ είναι η πρόβλεψη για την αξιολόγηση που θα δώσει ο χρήστης a στο αντικείμενο i, $w_{a,u}$ είναι η ομοιότητα μεταξύ του χρήστη a και u και K είναι το σύνολο των χρηστών που παρουσιάζουν την μεγαλύτερη ομοιότητα με τον χρήστη a.

Το μέτρο ομοιότητας που παρουσιάστηκε παραπάνω εξετάζει τη γραμμική εξάρτηση μεταξύ δύο μεταβλητών. Ένα διαφορετικό μέτρο ομοιότητας είναι η ομοιότητα συνημίτονου. Σε αυτή την περίπτωση οι αξιολογήσεις των δύο χρηστών αντιμετωπίζονται ως διανύσματα με m διαστάσεις και υπολογίζεται το συνημίτονο της μεταξύ τους γωνίας. Στην παρακάτω εξίσωση ορίζεται αυτή η ομοιότητα:

$$w_{a,u} = cos(\vec{r}_a, \vec{r}_u) = \frac{\vec{r}_a \cdot \vec{r}_u}{\|\vec{r}_a\|_2 \; x \; \|\vec{r}_u\|_2} = \frac{\Sigma_{i=1}^{m} r_{a,i} r_{u,i}}{\sqrt{\Sigma_{i=1}^{m} r_{a,i}^2} \sqrt{\Sigma_{i=1}^{m} r_{u,i}^2}} \quad (3)$$

Στον υπολογισμό της ομοιότητας συνημίτονων οι αξιολογήσεις πρέπει να έχουν μη αρνητικές τιμές, ενώ στις περιπτώσεις που δεν υπάρχει αξιολόγηση, τότε παίρνει την τιμή μηδέν.

Συνεργατικό φιλτράρισμα με βάση το αντικείμενο

Η μέθοδος συνεργατικού φιλτραρίσματος με βάση τη μνήμη δεν είναι αποτελεσματική όταν εφαρμόζεται σε εκατομμύρια χρήστες και αντικείμενα. Αυτό συμβαίνει λόγω της πολυπλοκότητας που έχει η εύρεση παρόμοιων χρηστών. Μία λύση σε αυτό το πρόβλημα αποτελεί η μέθοδος συνεργατικού φιλτραρίσματος με

βάση το αντικείμενο, όπου αντί να υπολογίζονται παρόμοιοι χρήστες, υπολογίζονται παρόμοια αντικείμενα με αυτά που έχει αλληλεπιδράσει θετικά ο χρήστης. Οι υπολογισμοί αυτής της μεθόδου είναι σημαντικά ταχύτεροι από ότι στο συνεργατικό φιλτράρισμα με βάση τη μνήμη.

Σε αυτή τη μέθοδο υπολογίζεται ο συντελεστής συσχέτισης Pearson ως μέτρο ομοιότητας μεταξύ δύο αντικειμένων i και j, όπως φαίνεται στην παρακάτω σχέση:

$$w_{i,j} = \frac{\Sigma_{u \in U} \left(r_{u,i} - \bar{r}_i\right)\left(r_{u,j} - \bar{r}_j\right)}{\sqrt{\Sigma_{u \in U}\left(r_{u,i} - \bar{r}_i\right)^2}\sqrt{\Sigma_{u \in U}\left(r_{u,j} - \bar{r}_j\right)^2}} \quad (4)$$

όπου U είναι το σύνολο όλων των χρηστών που έχουν αξιολογήσει και τα δύο αντικείμενα i και j, $r_{u,i}$ είναι η αξιολόγηση του αντικειμένου i από τον χρήστη u και $\bar{r}_i$ είναι η μέση αξιολόγηση του αντικειμένου i από όλους τους χρήστες του συνόλου U.

Η πρόβλεψη της αξιολόγησης του αντικειμένου i από έναν χρήστη a υπολογίζεται από την παρακάτω εξίσωση:

$$p_{a,i} = \frac{\Sigma_{j \in K} r_{a,j} w_{i,j}}{\Sigma_{j \in K} |w_{i,j}|} \quad (5)$$

όπου K είναι το σύνολο των k αντικειμένων που έχουν αξιολογηθεί από τον χρήστη a και παρουσιάζουν μεγαλύτερη ομοιότητα με το αντικείμενο i.

Παρακάτω παρουσιάζονται ορισμένες τεχνικές για την αντιμετώπιση πιθανών προκλήσεων κατά την διαδικασία του συνεργατικού φιλτραρίσματος με βάση τη μνήμη.

Συντελεστής Σημασίας Βάρους. Είναι πιθανό για έναν ενεργό χρήστη να υπολογιστούν ως αρκετά κοντινοί ‘γείτονές’ του, χρήστες με τους οποίους έχει πολύ λίγα κοινά αντικείμενα που έχουν αξιολογηθεί. Το γεγονός αυτό μπορεί να οδηγήσει σε λανθασμένες προβλέψεις. Ο συντελεστής σημασίας βάρους μπορεί να επιλύσει αυτό το πρόβλημα όταν πολλαπλασιαστεί με το βάρος ομοιότητας, καθώς δίνει αρκετά μικρή τιμή όταν τα κοινά αντικείμενα μεταξύ των δύο χρηστών είναι πολύ λίγα.

Προεπιλεγμένη Αξιολόγηση. Ένας διαφορετικός τρόπος επίλυσης του προβλήματος όπου τα κοινά αντικείμενα που έχουν αξιολογηθεί είναι πολύ λίγα, είναι να τεθεί μία προεπιλεγμένη τιμή στα αντικείμενα που δεν έχουν αξιολογηθεί. Έτσι, μπορεί να υπολογιστεί το μέτρο ομοιότητας $w_{a,u}$ μεταξύ δύο χρηστών χρησιμοποιώντας την

ένωση των αντικειμένων που έχουν αξιολογηθεί από τον κάθε χρήστη αντί για την τομή τους.

Αντίστροφη Συχνότητα Χρήστη. Όταν υπολογίζεται η ομοιότητα μεταξύ χρηστών κάποια αντικείμενα που πιθανώς να έχουν αξιολογηθεί από όλους τους χρήστες (ομόφωνα είτε θετικά ή αρνητικά) δεν είναι τόσο σημαντικά όσο αντικείμενα που έχουν αξιολογηθεί από λιγότερους χρήστες. Για την επίλυση αυτού του θέματος στο [9] παρουσιάστηκε η αντίστροφη συχνότητα χρήστη που ορίζεται παρακάτω:

$$f_i = log \frac{n}{n_i} \quad (6)$$

όπου n είναι ο συνολικός αριθμός των χρηστών και n_i ο αριθμός των χρηστών που έχουν αξιολογήσει το αντικείμενο i. Συνεπώς, όταν ένα αντικείμενο i έχει αξιολογηθεί σχεδόν από όλους τους χρήστες πολλαπλασιάζεται με την αντίστροφη συχνότητα χρήστη f_i.

Περίπτωση ενίσχυσης. Οι χρήστες που είναι αρκετά όμοιοι με τον ενεργό χρήστη θα πρέπει να λαμβάνονται περισσότερο υπόψη. Για το λόγο αυτό στο [9] παρουσιάστηκε η περίπτωση ενίσχυσης όπου τα αρχικά βάρη του $p_{a,i}$ τροποποιούνται ως εξής:

$$w'_{a,u} = w_{a,u} \cdot \left|w_{a,u}\right|^{\rho-1} \quad (7)$$

όπου ρ είναι ο παράγοντας ενίσχυσης και ισχύει ότι $\rho \geq 1$.

Συνεργατικό Φιλτράρισμα με Βάση το Μοντέλο

Οι μέθοδοι συνεργατικού φιλτραρίσματος με βάση το μοντέλο λειτουργούν υπολογίζοντας τις παραμέτρους στατιστικών μοντέλων για την πρόβλεψη των αξιολογήσεων των χρηστών. Συγκεκριμένα, όπως περιγράφεται στο [10] το συνεργατικό φιλτράρισμα με βάση το μοντέλο μπορεί να αντιμετωπιστεί ως ένα πρόβλημα ταξινόμησης, όπου τα αντικείμενα αντιστοιχούν στο διάνυσμα των χαρακτηριστικών του κάθε χρήστη και οι υπάρχουσες αξιολογήσεις στις ετικέτες. Το αποτέλεσμα είναι οι προβλεπόμενες αξιολογήσεις του κάθε ενεργού χρήστη.

Κάποιες από τις πιο δημοφιλείς τεχνικές συνεργατικού φιλτραρίσματος με βάση το μοντέλο είναι τα μοντέλα παραγοντοποίησης πίνακα και τα μοντέλα λανθάνουσας μεταβλητής [11]. Στα μοντέλα λανθάνουσας μεταβλητής υπολογίζεται και αναπαρίσταται η ομοιότητα των χρηστών και των αντικειμένων από κάποια απλούστερη δομή δεδομένων, ενώ στα μοντέλα συνεργατικού φιλτραρίσματος με

βάση τη μνήμη υπολογίζεται η ομοιότητα είτε μεταξύ των χρηστών ή μεταξύ των αντικειμένων. Η μέθοδος παραγοντοποίησης πίνακα αποτελεί μία κατηγορία της μεθόδου λανθάνουσας μεταβλητής, όπου οι χρήστες και τα αντικείμενα αναπαρίστανται από διανύσματα χαρακτηριστικών w_u και h_i αντίστοιχα, διάστασης k. Η εκπαίδευση γίνεται με σκοπό το εσωτερικό γινόμενο αυτών των διανυσμάτων $w_u^T h_i$ να προσεγγίζει τις υπάρχουσες αξιολογήσεις $r_{u,i}$ με κάποια συνάρτηση απώλειας. Μία επιλογή συνάρτησης απώλειας είναι το τετραγωνικό σφάλμα, όπου ελαχιστοποιείται η παρακάτω συνάρτηση:

$$J(W, H, \{b_u\}_{u=1}^{n}, \{b_i\}_{i=1}^{m}) = \sum_{(u,i)\in L}\left(r_{u,i} - w_u^T h_i\right)^2 \quad (8)$$

όπου ο $W = [w_1 \dots w_n]^T$ είναι ένας πίνακας διαστάσεων $n \times k$, ο $H = [h_1 \dots h_m]$ είναι ένας πίνακας $k \times m$ και L είναι το σύνολο όλων των σημείων (u, i) όπου ο χρήστης u έχει αξιολογήσει το αντικείμενο i. Στην περίπτωση που όλοι οι χρήστες έχουν αξιολογήσει όλα τα αντικείμενα η παραπάνω συνάρτηση γίνεται:

$$J(W, H) = \|R - WH\|_{fro}^2 \quad (9)$$

όπου R είναι ο πίνακας $n \times m$, n ο αριθμός των χρηστών, m ο αριθμός των αντικειμένων και τα στοιχεία του πίνακα που εκφράζουν τις αντίστοιχες αξιολογήσεις είναι όλα γνωστά. Σε αυτή την περίπτωση μπορεί να εφαρμοστεί ο αλγόριθμος SVD στον πίνακα R με:

$$R = UDV^T, \;\; W = U_k D_k^{\frac{1}{2}}, \;\; H = D_k^{\frac{1}{2}} V_k^T \quad (10)$$

όπου U_k, D_k, V_k εμπεριέχουν τις μεγαλύτερες τιμές στον πίνακα R.

Όμως, στις περισσότερες περιπτώσεις οι πιο πολλές αξιολογήσεις είναι άγνωστες και οι προσεγγίσεις με βάρη είναι προτιμότερες. Στην περίπτωση αυτή η συνάρτηση απώλειας εμπεριέχει βάρη και ορίζεται ως εξής:

$$J(W, H) = \|S \odot (R - WH)\|_{fro}^2 \quad (11)$$

όπου το σύμβολο ⊙ ορίζει την πράξη του πολλαπλασιασμού κατά στοιχείο και ο S είναι ένας δυαδικός πίνακας που παίρνει την τιμή 1 όταν η αξιολόγηση είναι γνωστή και 0 διαφορετικά. Οι συνήθεις τεχνικές βελτιστοποίησης είναι gradient-based, όπως για παράδειγμα η εναλλαγή ελαχίστων τετραγώνων, όπου λύνεται ως προς το διάνυσμα H, θεωρώντας το διάνυσμα W σταθερό και αντίστροφα μέχρι να ικανοποιηθεί κάποιο κριτήριο σύγκλισης. Καθώς υπολογίζεται το ένα διάνυσμα W ή H διατηρώντας το άλλο σταθερό, η διαδικασία είναι πλέον γραμμική παλινδρόμηση με βάρη.

Μία τεχνική για να μην υπερπροσαρμοστεί ένα μοντέλο στα δεδομένα, είναι η ελαχιστοποίηση της συνάρτησης $J(W,H)$ με κανονικοποίηση όπως φαίνεται παρακάτω:

$$J(W,H) + \gamma\|W\|^2 + \lambda\|H\|^2 \quad (12)$$

όπου γ και λ είναι παράμετροι κανονικοποίησης που καθορίζονται μέσω Διασταυρωμένης Επικύρωσης (Cross Validation). Όταν ολοκληρωθεί η εκπαίδευση του μοντέλου με τα διανύσματα W και H, το γινόμενο WH επιστρέφει ένα νέο πίνακα που περιέχει τις προβλεπόμενες αξιολογήσεις ώστε να γίνουν οι κατάλληλες συστάσεις.

2.3.5 Σύσταση βάσει Περιεχομένου

Στις συστάσεις συνεργατικού φιλτραρίσματος χρησιμοποιείται μόνο ο πίνακας αξιολογήσεων των χρηστών. Αυτές οι προσεγγίσεις αντιμετωπίζουν τους χρήστες και τα αντικείμενα ως ατομικές μονάδες, όπου γίνονται προβλέψεις χωρίς να λαμβάνονται υπόψη οι ιδιαιτερότητες των μεμονωμένων χρηστών ή αντικειμένων. Ωστόσο, είναι εφικτή μία καλύτερη εξατομικευμένη πρόταση αν είναι γνωστές περισσότερες πληροφορίες για ένα χρήστη όπως τα δημογραφικά στοιχεία [3] ή για ένα αντικείμενο όπως ο σκηνοθέτης και το είδος μίας ταινίας [4]. Οι συστάσεις βάσει περιεχομένου αναφέρονται σε τέτοιες προσεγγίσεις που παρέχουν συστάσεις συγκρίνοντας τις αναπαραστάσεις του περιεχομένου ενός αντικειμένου με τις αναπαραστάσεις περιεχομένου που ενδιαφέρουν το χρήστη.

Αρκετή από την έρευνα σε αυτόν τον τομέα έχει επικεντρωθεί στη σύσταση αντικειμένων με περιεχόμενο κειμένου, όπως ιστοσελίδες, βιβλία και ταινίες όπου οι ιστοσελίδες ή κάποιο σχετικό περιεχόμενο για παράδειγμα κάποια περιγραφή είναι διαθέσιμα μαζί με τις κριτικές των χρηστών. Επομένως, αρκετές προσεγγίσεις αντιμετώπισαν αυτό το πρόβλημα ως εργασία ανάκτησης πληροφοριών, όπου το

περιεχόμενο που σχετίζεται με τις προτιμήσεις του χρήστη αντιμετωπίζεται ως ένα ερώτημα (query) και τα μη αξιολογημένα έγγραφα αξιολογούνται βάσει αυτού του ερωτήματος [5].

2.3.6 Υβριδικές προσεγγίσεις

Προκειμένου να αξιοποιηθούν τα προτερήματα του συνεργατικού φιλτραρίσματος και των συστάσεων με βάση το περιεχόμενο έχουν προταθεί αρκετές υβριδικές προσεγγίσεις που συνδυάζουν και τα δύο. Μία απλή μέθοδος είναι να δημιουργηθούν ξεχωριστές λίστες συστάσεων από τη μέθοδο συνεργατικού φιλτραρίσματος και την μέθοδο συστάσεων βάσει περιεχομένου και να συγχωνευθούν δημιουργώντας μία τελική λίστα συστάσεων [6].

Αρκετές άλλες υβριδικές προσεγγίσεις βασίζονται στο συνεργατικό φιλτράρισμα αλλά διατηρούν επίσης ένα προφίλ βασισμένο στο περιεχόμενο για κάθε χρήστη. Αυτά τα προφίλ βάσει του περιεχομένου χρησιμοποιούνται για την εύρεση παρόμοιων χρηστών, αντί για την αξιολόγηση αντικειμένων. Στο [3] κάθε προφίλ χρήστη αναπαριστάται από ένα διάνυσμα σταθμισμένων λέξεων που προέρχονται από θετικά παραδείγματα εκπαίδευσης χρησιμοποιώντας τον αλγόριθμο Winnow. Οι προβλέψεις γίνονται εφαρμόζοντας το συνεργατικό φιλτράρισμα κατευθείαν στον πίνακα των προφίλ των χρηστών (σε αντίθεση με τον πίνακα των αξιολογήσεων των χρηστών). Ορισμένες υβριδικές προσεγγίσεις προσπαθούν να συνδυάσουν άμεσα το περιεχόμενο και τα συνεργατικά δεδομένα κάτω από ένα πιθανοτικό πλαίσιο. Στο [7] επεκτάθηκε το μοντέλο πτυχών του Hofmann [8] ώστε να ενσωματώνονται τρισδιάστατα δεδομένα όπου συνυπάρχουν οι χρήστες, τα αντικείμενα και το περιεχόμενο. Το μοντέλο δημιουργίας τους προϋποθέτει ότι οι χρήστες επιλέγουν λανθάνοντα θέματα και τα έγγραφα και οι λέξεις περιεχομένου τους δημιουργούνται από αυτά τα θέματα.

Κεφάλαιο 3: Μοντέλο Συστήματος Προσωποποιημένων Συστάσεων

3.1 Εισαγωγή

Σε αυτή την ενότητα παρουσιάζεται ο ορισμός του προβλήματος που επιλύεται στην παρούσα διπλωματική εργασία καθώς επίσης και οι τεχνικές που χρησιμοποιήθηκαν για αυτό.

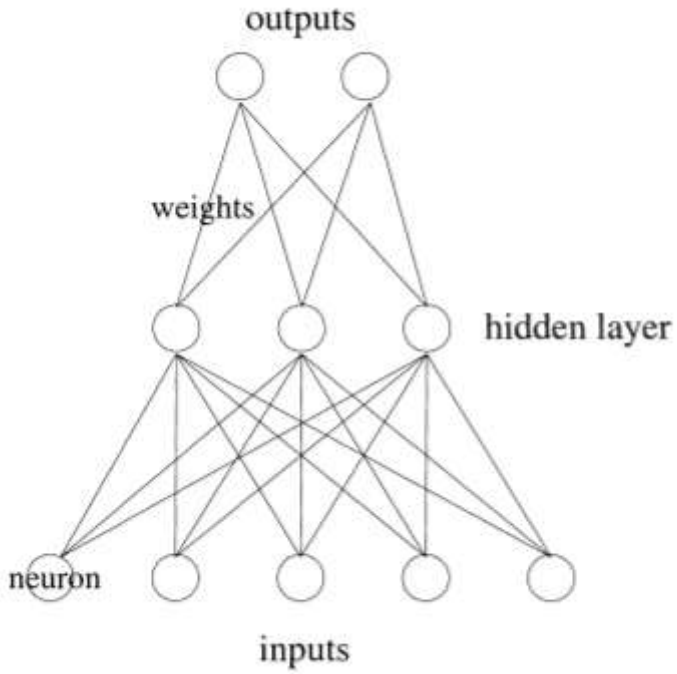
3.2 Ορισμός Προβλήματος

Οι Συνεργατικές Μέθοδοι Συστάσεων ορίζονται ως η δημιουργία ενός μοντέλου συστάσεων το οποίο εκπαιδεύεται με δεδομένα που βρίσκονται κατανεμημένα σε απομακρυσμένες συσκευές χωρίς αυτά να συλλέγονται σε κάποιον κεντρικό διακομιστή. Συγκεκριμένα, ο κεντρικός διακομιστής αποστέλλει το μοντέλο συστάσεων σε κάθε χρήστη, οπού εκεί εκπαιδεύεται τοπικά με τα δεδομένα που βρίσκονται αποθηκευμένα στην κάθε συσκευή. Μόλις πραγματοποιηθεί αυτή η τοπική εκπαίδευση υπολογίζονται οι αντίστοιχες ενημερώσεις του τοπικού μοντέλου και αποστέλλονται πίσω στον διακομιστή. Ο κεντρικός διακομιστής αφού συλλέξει τις ενημερώσεις από όλους τους χρήστες που συμμετέχουν στην εκπαίδευση ενημερώνει τα βάρη του κεντρικού μοντέλου το οποίο είναι έτοιμο για αναδιανομή στους χρήστες για τον επόμενο γύρο εκπαίδευσης.

Σε αυτή την διπλωματική εργασία το μοντέλο συστάσεων είναι ένα νευρωνικό δίκτυο ενσωμάτωσης το οποίο δέχεται σαν είσοδο τα χαρακτηριστικά των χρηστών καθώς και τα χαρακτηριστικά των αντικειμένων και η έξοδος είναι οι προβλέψεις των αξιολογήσεων του κάθε χρήστη προς κάθε αντικείμενο. Στη συνέχεια, χρησιμοποιείται η τεχνική της προσωποποιημένης συνεργατικής μάθησης όπου υπολογίζεται η κατάλληλη μίξη του τοπικού και του κεντρικού μοντέλου ώστε να ελαχιστοποιηθεί το σφάλμα. Έπειτα, από τη συγκέντρωση των τοπικών ενημερώσεων στον κεντρικό διακομιστή υπολογίζεται ο μέσος όρος αυτών από όλους τους χρήστες, με τον οποίο ενημερώνεται εκ νέου το κεντρικό μοντέλο.

3.3 Νευρωνικό Δίκτυο Ενσωμάτωσης

Τα νευρωνικά δίκτυα είναι εμπνευσμένα από την λειτουργικότητα του ανθρώπινου εγκεφάλου όπου εκατοντάδες δισεκατομμύρια νευρώνες επεξεργάζονται παράλληλα πληροφορίες. Όπως αναφέρεται και στο [12], ένα τεχνητό νευρωνικό δίκτυο αποτελείται από ένα στρώμα νευρώνων εισόδου (input layer), ένα, δύο ή τρία κρυφά στρώματα νευρώνων (hidden layers) και στο τέλος από ένα στρώμα νευρώνων εξόδου. Στην εικόνα 4 φαίνεται μία τυπική αρχιτεκτονική ενός νευρωνικού δικτύου όπου εμφανίζονται και οι γραμμές που συνδέουν τους νευρώνες.



Εικόνα 4. Αρχιτεκτονική τεχνητού νευρωνικού δικτύου

Κάθε σύνδεση μεταξύ δύο νευρώνων σχετίζεται με έναν αριθμό που λέγεται βάρος. Η έξοδος h_i του νευρώνα i στο κρυφό στρώμα είναι:

$$h_i = \sigma\left(\sum_{j=1}^{N} V_{ij}x_j + T_i^{h_i d}\right) \quad (13)$$

Όπου $\sigma()$ είναι η συνάρτηση ενεργοποίησης, N ο αριθμός των νευρώνων εισόδου, V_{ij} τα βάρη, x_j οι είσοδοι στους νευρώνες εισόδου και $T_i^{h_i d}$ οι όροι για το κατώφλι των κρυφών νευρώνων. Ο σκοπός της συνάρτησης ενεργοποίησης πέρα από την εισαγωγή της μη γραμμικότητας στο νευρωνικό δίκτυο είναι να δεσμεύει την τιμή του

νευρώνα ώστε το νευρωνικό δίκτυο να μην παραλύσει από τους διαφορετικούς νευρώνες.

Ένα παράδειγμα συνάρτησης ενεργοποίησης είναι η σιγμοειδής (ή λογιστική) συνάρτηση που ορίζεται ως:

$$\sigma(u) = \frac{1}{1+exp(-u)} \quad (14)$$

Άλλες πιθανές συναρτήσεις ενεργοποίησης είναι η εφαπτομένη τόξου και η υπερβολική εφαπτομένη. Έχουν παρόμοια απόκριση στις εισόδους με τη σιγμοειδή συνάρτηση, αλλά διαφέρουν στο εύρος εξόδου.

Είναι αποδεδειγμένο ότι ένα νευρωνικό δίκτυο κατασκευασμένο με τον παραπάνω τρόπο μπορεί να προσεγγίσει οποιαδήποτε υπολογιστική συνάρτηση με αυθαίρετη ακρίβεια. Οι αριθμοί που δόθηκαν στους νευρώνες εισόδους είναι ανεξάρτητες μεταβλητές, ενώ αυτές που επιστρέφονται από τους νευρώνες εξόδου είναι μεταβλητές εξαρτώμενες από τη συνάρτηση που προσεγγίζεται από το νευρωνικό δίκτυο. Οι είσοδοι και έξοδοι ενός νευρωνικού δικτύου μπορεί να είναι δυαδικοί (όπως ναι ή όχι) ή σύμβολα (όπως κόκκινο, μπλε κ.λπ.) όταν τα δεδομένα είναι κατάλληλα κωδικοποιημένα. Αυτή η δυνατότητα παρέχει ένα ευρύ φάσμα εφαρμογών στα νευρωνικά δίκτυα.

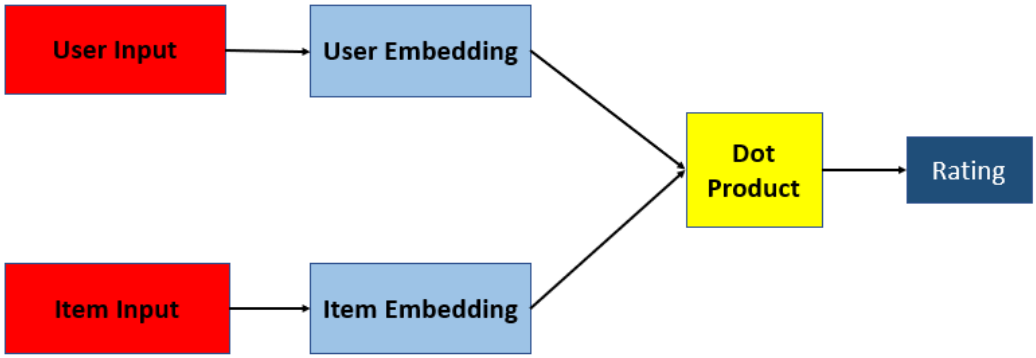
Στρώμα ενσωμάτωσης

Στο στρώμα ενσωμάτωσης (embedding layer) δημιουργούνται τα διανύσματα ενσωμάτωσης (embeddings). Όπως αναφέρεται στο [13], έστω p_u το διάνυσμα για τον χρήστη u και q_i το διάνυσμα για το αντικείμενο i, όπου περιέχουν τα χαρακτηριστικά του χρήστη και του αντικειμένου (πχ ID, ηλικία του χρήστη, περιγραφή του αντικειμένου). Τα δύο διανύσματα είναι και τα δύο διάστασης K. Η αξιολόγηση $\hat{R}_{ui}$ του αντικειμένου i από τον χρήστη u υπολογίζεται από το εσωτερικό γινόμενο των διανυσμάτων p_u, q_i όπως φαίνεται παρακάτω:

$$\hat{R}_{ui} = < p_u, q_i > = p_u^T q_i \quad (15)$$

Στην παρακάτω εικόνα φαίνεται η αρχιτεκτονική του νευρωνικού δικτύου ενσωμάτωσης. Αρχικά, βρίσκεται το στρώμα εισόδου όπου γίνεται παράλληλη είσοδος των δεδομένων των χρηστών και των αντικειμένων. Στο επόμενο στρώμα δημιουργούνται τα διανύσματα ενσωμάτωσης (embeddings). Στο επόμενο στρώμα

υπολογίζεται το εσωτερικό γινόμενο των δύο διανυσμάτων και τέλος στο στρώμα εξόδου φαίνεται η προβλεπόμενη αξιολόγηση.



Εικόνα 5. Αρχιτεκτονική νευρωνικού δικτύου ενσωμάτωσης

3.4 Συνεργατική Μάθηση

Η μέθοδος της συνεργατικής μάθησης αποτελείται από την εκπαίδευση ενός γενικού μοντέλου με δεδομένα τα οποία είναι αποθηκευμένα σε απομακρυσμένες συσκευές. Τα δεδομένα που παράγονται από την κάθε συσκευή αποθηκεύονται και επεξεργάζονται τοπικά, αντί να συλλέγονται ακατέργαστα σε έναν κεντρικό διακομιστή, όπως γίνεται στις παραδοσιακές τεχνικές μηχανικής μάθησης. Στον κεντρικό διακομιστή συλλέγονται μόνο ενδιάμεσες ενημερώσεις από τις τοπικές εκπαιδεύσεις των συσκευών.

Όπως αναφέρεται στο [14], η επικοινωνία μέσω τη μεταφοράς των ενημερώσεων μεταξύ του τοπικού και του γενικού μοντέλου επιτυγχάνεται μέσω των γύρων επικοινωνίας. Υπάρχει ένα σταθερό σύνολο χρηστών K με ένα σταθερό τοπικό σύνολο δεδομένων. Στην αρχή κάθε γύρου επιλέγεται ένα τυχαίο κλάσμα C των χρηστών και ο διακομιστής στέλνει την τρέχουσα κατάσταση του κεντρικού αλγορίθμου σε κάθε έναν από αυτούς τους χρήστες (π.χ. τις τρέχουσες παραμέτρους του μοντέλου). Επιλέγεται μόνο ένα ποσοστό των χρηστών για την βελτίωση της αποδοτικότητας καθώς οι πειραματικές μελέτες δείχνουν ότι η προσθήκη περισσότερων χρηστών πέρα από ένα συγκεκριμένο σημείο μπορεί να μειώσει την απόδοση του μοντέλου. Στη συνέχεια, κάθε χρήστης που έχει επιλεγεί εκτελεί έναν τοπικό υπολογισμό με βάση την τρέχουσα κατάσταση του κεντρικού μοντέλου και το τοπικό σύνολο δεδομένων του και στέλνει μία ενημέρωση στον διακομιστή. Έπειτα,

ο διακομιστής εφαρμόζει αυτές τις ενημερώσεις στο κεντρικό μοντέλο και η διαδικασία επαναλαμβάνεται.

Σε ένα παράδειγμα συνεργατικής μάθησης ορίζεται ως n ο συνολικός αριθμός των χρηστών που είναι συνδεδεμένοι σε έναν κεντρικό διακομιστή. Ο κάθε χρήστης έχει πρόσβαση μόνο στα δικά του δεδομένα και όχι στα δεδομένα των άλλων χρηστών. Ορίζεται ως $f_i: \mathbb{R}^d \to \mathbb{R}$ η απώλεια του τοπικού μοντέλου του χρήστη i και ο στόχος είναι η παρακάτω ελαχιστοποίηση:

$$\min_{w \in \mathbb{R}^d} f(w) := \frac{1}{n} \sum_{i=1}^{n} f_i(w) \qquad (16)$$

Για ένα πρόβλημα μηχανικής μάθησης συνήθως επιλέγεται η συνάρτηση:

$$f_i(w) = l(x_i, y_i, w) \qquad (17)$$

όπου είναι η απώλεια της πρόβλεψης για το παράδειγμα (x_i, y_i) που έγινε με τις παραμέτρους του μοντέλου w.

Υποθέτοντας ότι υπάρχουν K χρήστες στους οποίους είναι κατανεμημένα τα δεδομένα, ορίζεται ως P_k το σύνολο των δεικτών των σημείων δεδομένων του χρήστη k με $n_k = |P_k|$. Έτσι η προηγούμενη συνάρτηση ξαναγράφεται ως:

$$f(w) = \sum_{k=1}^{K} \frac{n_k}{n} F_k(w) \quad \text{όπου} \quad F_k(w) = \frac{1}{n_k} \sum_{i \in P_k} f_i(w) \qquad (18)$$

Ο αλγόριθμος *FedAvg* (Federated Averaging) που παρουσιάστηκε στο [15], στοχεύει στην ελαχιστοποίηση της αντικειμενικής συνάρτησης υποθέτοντας ένα σχήμα σύγχρονης ενημέρωσης και μία συνάρτηση απώλειας. Όπως αναφέρεται στο [16], το μοντέλο αρχικοποιείται στον κεντρικό διακομιστή με μία δεδομένη αρχιτεκτονική με βάρη w_0. Μετά την αρχικοποίηση ξεκινούν να αλληλεπιδρούν ταυτόχρονα μεταξύ τους ο διακομιστής και οι συσκευές των χρηστών κατά τη διάρκεια των γύρων επικοινωνίας. Παρακάτω, περιγράφεται ένας γύρος επικοινωνίας την στιγμή $t \in [1, \dots, T]$:

1. Το κεντρικό μοντέλο w_{t-1} μοιράζεται σε ένα υποσύνολο των χρηστών S_t που επιλέγονται τυχαία από το σύνολο των K χρηστών δεδομένου ενός ποσοστού συμμετοχής C.

2. Κάθε χρήστης $k \in S_t$ εκτελεί ένα ή περισσότερα βήματα εκπαίδευσης στα τοπικά δεδομένα του με βάση την ελαχιστοποίηση της τοπικής αντικειμενικής συνάρτησής του F_k με χρήση της τεχνικής Stochastic Gradient Descent (SGD) και με τοπικό ρυθμό μάθησης (learning rate) η_{local}. Ο αριθμός των βημάτων που εκτελούνται τοπικά είναι:

$$E \times max\ (ceil\left(\frac{n_k}{B}\ , 1\right)) \qquad (19)$$

όπου n_k ο αριθμός των σημείων δεδομένων που είναι διαθέσιμα τοπικά, E ο αριθμός των τοπικών επαναλήψεων και B το τοπικό μέγεθος του batch.

3. Οι χρήστες από το σύνολο S_t στέλνουν πίσω στον διακομιστή τις ενημερώσεις του μοντέλου τους $w_{t,k}$, $k \in S_t$ μόλις τελειώσει η τοπική εκπαίδευση.
4. Ο διακομιστής υπολογίζει ένα μέσο μοντέλο w_t με βάση τις ατομικές ενημερώσεις του χρήστη $w_{t,k}$, $k \in S_t$, όπου στην ενημέρωση κάθε χρήστη αντιστοιχεί ένα βάρος $\frac{n_k}{n_r}$ όπου $n_r = \sum_{k \in S_t} n_k \approx C \times \sum_{k=1}^{K} n_k$.

Όταν $B = \infty$ (δηλαδή το μέγεθος του batch είναι ίσο με το μέγεθος του τοπικού συνόλου δεδομένων) και $E = 1$, τότε εκτελείται μία gradient ενημέρωση στα δεδομένα κάθε χρήστη. Αυτό είναι αυστηρά ισοδύναμο με τον gradient υπολογισμό σε ένα batch που περιλαμβάνει όλα τα επιλεγμένα σημεία δεδομένων του χρήστη. Αυτή η συγκεκριμένη περίπτωση ονομάζεται *FedSGD*. Η τεχνική *FedAvg* (Federated Averaging) είναι η γενική περίπτωση όταν περισσότερες από μία ενημερώσεις πραγματοποιούνται τοπικά για κάθε χρήστη.

Το κεντρικό βήμα υπολογισμού του μέσου όρου μπορεί να γραφτεί χρησιμοποιώντας ένα ρυθμό κεντρικής ενημέρωσης η_{global} ως εξής:

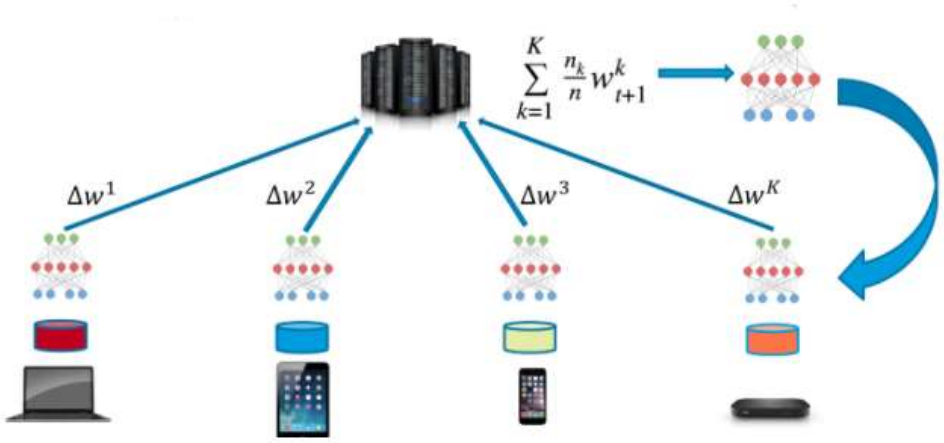
$$w_t \leftarrow w_{t-1} - \eta_{global} \sum_{k \in S_t} \frac{n_k}{n} (w_{t-1} - w_{t,k}) \qquad (20)$$

Θέτοντας το ρυθμό κεντρικής ενημέρωσης η_{global} ίσο με 1 ισοδυναμεί με την περίπτωση υπολογισμού του μέσου όρου με βάρη. Η παραπάνω εξίσωση υπογραμμίζει τον παράλληλο υπολογισμό μεταξύ του κεντρικού μέσου όρου και της gradient ενημέρωσης:

$$G_t = \sum_{k \in S_t} \frac{n_k}{n} (w_{t-1} - w_{t,k}) \qquad (21)$$

Αυτός ο παράλληλος υπολογισμός ενθαρρύνει τη χρήση προσαρμοστικών ενημερώσεων ανά συντεταγμένη για την G_t που έχουν αποδειχθεί επιτυχημένες για την βελτιστοποίηση κεντρικοποιημένων βαθιών νευρωνικών δικτύων όπως ο Adam [17]. Ο υπολογισμός του μέσου όρου με βάση τη στιγμή επιτρέπει την εξομάλυνση του μοντέλου λαμβάνοντας υπόψιν τους προηγούμενους γύρους ενημερώσεων που υπολογίστηκαν από διαφορετικά υποσύνολα χρηστών.

Στην παρακάτω εικόνα παρουσιάζεται η αρχιτεκτονική ενός μοντέλου συνεργατικής μάθησης. Το κεντρικό μοντέλο αποστέλλεται από τον κεντρικό διακομιστή στις απομακρυσμένες συσκευές. Από εκεί υπολογίζονται τα τοπικά μοντέλα και οι τοπικές ενημερώσεις αποστέλλονται από την κάθε συσκευή στον κεντρικό διακομιστή. Εκεί με την τεχνική Federated Averaging υπολογίζεται ο μέσος όρος των ενημερώσεων που θα αποτελέσει τα νέα βάρη του κεντρικού μοντέλου για την συγκεκριμένη επανάληψη. Η διαδικασία αυτή επαναλαμβάνεται έως ότου ολοκληρωθεί ο προκαθορισμένος αριθμός των γύρων επαναλήψεων.



Εικόνα 6. Αρχιτεκτονική μοντέλου συνεργατικής μάθησης

3.5 Προσωποποιημένη Συνεργατική Μάθηση

Όπως αναφέρεται στο [18], σε ένα τυπικό παράδειγμα συνεργατικής μάθησης όπου ο στόχος είναι η εκμάθηση ενός κεντρικού μοντέλου για όλες τις συσκευές συνεργατικά, το εκπαιδευμένο κεντρικό μοντέλο που προέκυψε από την ελαχιστοποίηση της κοινής εμπειρικής κατανομής με την κατάλληλη τοποθέτηση βαρών, ενδέχεται να μη γενικεύεται τέλεια στα τοπικά δεδομένα των χρηστών. Αυτό συμβαίνει κυρίως όταν η ετερογένεια μεταξύ των δεδομένων είναι υψηλή (δηλαδή

όταν το βέλτιστο κεντρικό και το βέλτιστο τοπικό μοντέλο ενδέχεται να διαφέρουν σημαντικά). Ωστόσο, υποθέτοντας ότι όλα τα δεδομένα των χρηστών προέρχονται από την (κατά προσέγγιση) παρόμοια κατανομή, είναι αναμενόμενο ότι το κεντρικό μοντέλο έχει καλύτερη ακρίβεια γενίκευσης σε οποιαδήποτε κατανομή χρήστη από ότι το τοπικό μοντέλο του χρήστη.

Από την οπτική του τοπικού χρήστη το βασικό κίνητρο για τη συμμετοχή στην συνεργατική μάθηση είναι η προσπάθεια μείωσης του τοπικού σφάλματος γενίκευσης με τη βοήθεια των δεδομένων των άλλων χρηστών. Σε αυτή την περίπτωση η ιδανική κατάσταση θα ήταν να μπορεί ο χρήστης να χρησιμοποιήσει πληροφορίες από το κεντρικό μοντέλο για να αντισταθμίσει τον μικρό αριθμό των τοπικών δεδομένων εκπαίδευσης, ελαχιστοποιώντας το πρόβλημα που προκαλείται από την ετερογένεια μεταξύ των τοπικών δεδομένων του κάθε χρήστη και των δεδομένων που μοιράζονται από τις άλλες συσκευές.

Είναι προφανές πως όταν η τοπική κατανομή συσχετίζεται σε μεγάλο βαθμό με την κεντρική κατανομή είναι προτιμότερο το κεντρικό μοντέλο. Διαφορετικά, το κεντρικό μοντέλο μπορεί να μην είναι αποτελεσματικό για να χρησιμοποιηθεί ως τοπικό μοντέλο. Αυτό αποτελεί βασικό κίνητρο για να συνδυαστεί το κεντρικό και το τοπικό μοντέλο με ένα βάρος, ως κοινό μοντέλο πρόβλεψης δηλαδή το προσωποποιημένο μοντέλο.

Στην προσωποποιημένη συνεργατική μάθηση ο στόχος είναι να βρεθεί ο βέλτιστος συνδυασμός του κεντρικού και του τοπικού μοντέλου, προκειμένου να επιτευχθεί ένα καλύτερο μοντέλο εξατομικευμένο στο χρήστη. Στην παρούσα διπλωματική εργασία κάθε χρήστης εκπαιδεύει ένα τοπικό μοντέλο και ενσωματώνει ένα μέρος του κεντρικού μοντέλου με κάποιο βάρος ανάμιξης a_i. Συγκεκριμένα:

$$h_{a_i} = a_i \hat{h}_i^* + (1 - a_i)\bar{h}* \quad (22)$$

όπου $\bar{h}* = argmin_{h \in H} \hat{\mathcal{L}}_{\bar{D}}(h)$ είναι ο κεντρικός εμπειρικός ελαχιστοποιητής ρίσκου και $\hat{h}_i^* = argmin_{h \in H} \hat{\mathcal{L}}_{D_i}\left(a_i h + (1 - a_i)\bar{h}*\right)$ είναι το αναμεμειγμένο μοντέλο που ελαχιστοποιεί την εμπειρική απώλεια στον χρήστη i.

Η ισορροπία μεταξύ των δύο μοντέλων καθορίζεται από μία παράμετρο a_i, η οποία σχετίζεται με την διαφορετικότητα του τοπικού και του κεντρικού μοντέλου. Συγκεκριμένα, όταν η κατανομές των κεντρικών και των τοπικών δεδομένων είναι αρκετά παρόμοιες μεταξύ τους είναι αναμενόμενο ότι η βέλτιστη επιλογή για την παράμετρο ανάμιξης είναι μία χαμηλή τιμή ώστε να επωφεληθούν τα τοπικά μοντέλα από τα δεδομένα άλλων συσκευών. Από την άλλη μεριά, όταν η τοπική και η κεντρική κατανομή δεδομένων διαφέρουν σημαντικά μεταξύ τους, τότε η παράμετρος ανάμιξης πρέπει να είναι κοντά στη μονάδα ώστε να μειωθεί η συνεισφορά από τα δεδομένα των άλλων συσκευών στο βέλτιστο τοπικό μοντέλο.

Πιο συγκεκριμένα, σε κάθε γύρο επανάληψης ο διακομιστής επιλέγει τυχαία K χρήστες ως ένα σύνολο U_t. Κάθε επιλεγμένος χρήστης διατηρεί τρία μοντέλα στην επανάληψη t:

1. Την τοπική έκδοση του κεντρικού μοντέλου $w_i^{(t)}$.
2. Το τοπικό μοντέλο $u_i^{(t)}$.
3. Το αναμεμειγμένο μοντέλο $\bar{u}_i^{(t)} = a_i u_i^{(t)} + (1 - a_i) w_i^{(t)}$.

Στο τέλος κάθε γύρου επανάληψης υπολογίζονται τα βάρη των προσωποποιημένων μοντέλων όλων των χρηστών και στέλνονται στον κεντρικό διακομιστή. Εκεί υπολογίζεται ο μέσος όρος των βαρών όπου θα διαμορφώσει τα νέα βάρη του κεντρικού μοντέλου στο τέλος της επανάληψης t. Στη συνέχεια η διαδικασία επαναλαμβάνεται με το νέο ενημερωμένο κεντρικό μοντέλο.

Κεφάλαιο 4: Πειραματική Μελέτη και Αξιολόγηση

4.1 Εισαγωγή

Σε αυτό το κεφάλαιο παρουσιάζεται η πειραματική μελέτη που έγινε για την υλοποίηση αυτής της διπλωματικής εργασίας. Συγκεκριμένα, αναπτύχθηκε ένας αλγόριθμος συνεργατικού συστήματος συστάσεων σε ένα σύνολο δεδομένων αναφορικά με αξιολογήσεις χρηστών για ταινίες. Αρχικά, παρουσιάζονται τα τεχνικά χαρακτηριστικά που επιλέχθηκαν (π.χ. η γλώσσα προγραμματισμού), έπειτα το σύνολο δεδομένων που επιλέχθηκε, στη συνέχεια τα αποτελέσματα που προέκυψαν και τέλος γίνεται η αξιολόγηση αυτών των αποτελεσμάτων.

4.2 Η Γλώσσα Προγραμματισμού Python

Η γλώσσα προγραμματισμού που επιλέχθηκε για την πειραματική μελέτη της διπλωματικής εργασίας είναι η Python (Python 3). Όπως αναφέρεται στο [19], η Python είναι μία διαδραστική αντικειμενοστραφής γλώσσα προγραμματισμού. Παρέχει δομές δεδομένων υψηλού επιπέδου όπως λίστες, λεξικά, modules, classes, αυτόματη διαχείριση μνήμης κλπ. Έχει μία εξαιρετικά απλή σύνταξη, ωστόσο είναι μία ισχυρή και γενικής χρήσης γλώσσα προγραμματισμού. Σχεδιάστηκε το 1990 από τον Guido van Rossum. Όπως και άλλες γλώσσες προγραμματισμού είναι δωρεάν, ακόμα και για εμπορικούς σκοπούς και μπορεί να εκτελεστεί σε οποιονδήποτε σύγχρονο υπολογιστή. Ένα πρόγραμμα Python μεταγλωττίζεται αυτόματα από τον διερμηνευτή (interpreter) σε ανεξάρτητη πλατφόρμα κώδικα byte που στη συνέχεια ερμηνεύεται.

4.3 Βιβλιοθήκες

Παρακάτω παρουσιάζονται οι βασικότερες βιβλιοθήκες της γλώσσας προγραμματισμού Python που χρησιμοποιήθηκαν για την υλοποίηση της παρούσας διπλωματικής εργασίας, όπως αναφέρονται στο Python Package Index:

- **Pandas.** Η βιβλιοθήκη Pandas παρέχει γρήγορες και ευέλικτες δομές δεδομένων που έχουν σχεδιαστεί για να λειτουργούν με δομημένα δεδομένα

(πολυδιάστατα, δυνητικά ετερογενή, πίνακες) και δεδομένα χρονοσειρών. Στόχος είναι το θεμελιώδες δομικό στοιχείο υψηλού επιπέδου για την πρακτική και πραγματική ανάλυση δεδομένων στην Python.

- **Numpy.** Η βιβλιοθήκη Numpy παρέχει υποστήριξη για μεγάλους πολυδιάστατους πίνακες καθώς και μαθηματικές συναρτήσεις υψηλού επιπέδου.
- **Matplotlib.** Η βιβλιοθήκη Matplotlib είναι μία ολοκληρωμένη βιβλιοθήκη για τη δημιουργία στατικών, κινούμενων και διαδραστικών απεικονίσεων.
- **Tensorflow.** Η βιβλιοθήκη Tensorflow, όπως αναφέρεται στο Python Package Index είναι μία βιβλιοθήκη ανοιχτού λογισμικού κώδικα για αριθμητικούς υπολογισμούς υψηλής απόδοσης. Η ευέλικτη αρχιτεκτονική επιτρέπει την εύκολη ανάπτυξη υπολογισμού σε διάφορες πλατφόρμες (CPUs, GPUs, TPUs) και από υπολογιστές σε συστάδες διακομιστών σε κινητές συσκευές και απομακρυσμένες συσκευές. Αρχικά, αναπτύχθηκε από μηχανικούς από την ομάδα του Google Brain στον οργανισμό AI όπως Google, συνοδεύεται από ισχυρή υποστήριξη για μηχανική μάθηση και βαθιά μάθηση και ο ευέλικτος αριθμητικός πυρήνας υπολογισμού χρησιμοποιείται σε πολλούς επιστημονικούς τομείς.
- **Keras.** Το Keras είναι ένα API (Application Programming Interface) νευρωνικών δικτύων υψηλού επιπέδου για την γλώσσα προγραμματισμού Python. Όπως αναφέρεται στο keras.io το Keras ακολουθεί βέλτιστες πρακτικές για τη μείωση του γνωστικού φορτίου. Συγκεκριμένα, προσφέρει σταθερά και απλά API, ελαχιστοποιεί τον αριθμό των ενεργειών του χρήστη που απαιτούνται για συχνές περιπτώσεις χρήστη και παρέχει σαφή ανατροφοδότηση σχετικά με το σφάλμα χρήστη.
- **Scikit-learn.** Το Scikit-learn είναι ένα module της Python για μηχανική μάθηση.

4.4 Το Σύνολο Δεδομένων

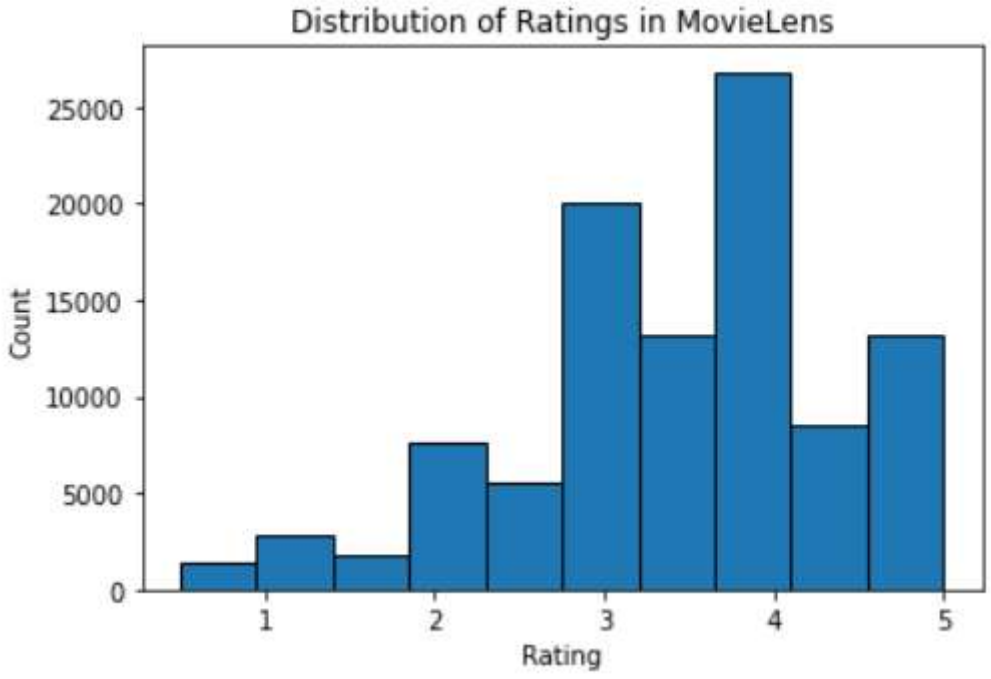
Το σύνολο δεδομένων που επιλέχθηκε για την συγκεκριμένη διπλωματική εργασία είναι το MovieLens (όπου από το link https://grouplens.org/datasets/movielens/latest/ επιλέχθηκε το ml-latest-small.zip), το οποίο χρησιμοποιείται ευρέως για έρευνα σε συστήματα συστάσεων. Από αυτό το σύνολο δεδομένων χρησιμοποιήθηκε το αρχείο ‘ratings’ το οποίο περιέχει 100836 εγγραφές, οι οποίες είναι οι αξιολογήσεις 610 διαφορετικών χρηστών για 9724 διαφορετικές ταινίες. Το αρχείο αυτό αποτελείται από τέσσερις στήλες οι οποίες είναι οι ‘userId’, ‘movieId’, ‘rating’, ‘timestamp’. Η ελάχιστη τιμή αξιολόγησης που μπορεί να δώσει όπως χρήστης σε μία ταινία είναι 0.5 και η μέγιστη 5.0.

Στην εικόνα 7 φαίνονται οι πρώτες πέντε εγγραφές του συνόλου δεδομένων:

	userId	movieId	rating	timestamp
0	279	4816	4.0	1506395821
1	333	3564	1.0	965671775
2	606	1571	4.0	1171824516
3	372	1055	2.0	874414588
4	254	3949	5.0	1180563891

Εικόνα 7. Πρώτες εγγραφές του συνόλου δεδομένων

Στη εικόνα 8, απεικονίζεται η κατανομή των μετρήσεων των διαφορετικών δυνατών τιμών των αξιολογήσεων.



Εικόνα 8. Κατανομή των αξιολογήσεων του συνόλου δεδομένων

Όπως προκύπτει από την κατανομή των αξιολογήσεων οι περισσότεροι χρήστες έχουν αξιολογήσει ταινίες με ‘3’ και ‘4’, ενώ οι αξιολογήσεις με ‘0.5’, ‘1.5’ και ‘1.5’ είναι οι λιγότερες.

4.5 Περιγραφή Πειραματικής Μελέτης

Η εκτέλεση της εργασίας έγινε σε υπολογιστή με τα παρακάτω τεχνικά χαρακτηριστικά:

- Επεξεργαστής: Intell Corel i7-7500U CPU @ 2.70GHz 2.90GHz
- Εγκατεστημένη μνήμη (RAM): 6,00 GB
- Χωρητικότητα: 222GB

Παρακάτω περιγράφονται τα βήματα που ακολουθήθηκαν για την υλοποίηση της πειραματικής μελέτης της διπλωματικής εργασίας.

Αρχικά, εισάγονται οι απαραίτητες βιβλιοθήκες οι οποίες είναι:

- Pandas
- Numpy
- Matplotlib
- Random
- Tensorflow
- Keras
- Scikit-learn

Στη συνέχεια αφού διαβαστεί το αρχείο 'ratings.csv' δημιουργείται ένα dataframe με 100836 εγγραφές και 4 στήλες.

Προεπεξεργασία Δεδομένων

Έπειτα, επιλέγονται από το dataframe οι στήλες 'userId' και 'movieId' ως τα χαρακτηριστικά των δεδομένων (ως x) και η στήλη 'rating' ως η ετικέτα των δεδομένων (ως y). Τα δεδομένα y που αναφέρονται ως αξιολογήσεις των χρηστών για ταινίες κανονικοποιούνται ώστε να παίρνουν τιμές μεταξύ 0 και 1. Η κανονικοποίηση των δεδομένων είναι μία τεχνική που χρησιμοποιείται στη μηχανική μάθηση ώστε να κάνει την εκπαίδευση του μοντέλου λιγότερη ευαίσθητη στη διαφοροποίηση των τιμών των δεδομένων. Αυτό επιτρέπει στο μοντέλο να συγκλίνει σε καλύτερα βάρη και να δημιουργείται ένα πιο ακριβές μοντέλο. Παρακάτω φαίνεται ο τύπος όπως κανονικοποίησης:

$$y_{norm} = \frac{y - y_{min}}{y_{max} - y_{min}} \quad (23)$$

Διαχωρισμός των Δεδομένων σε Train και Test

Στη συνέχεια, τα δεδομένα x και y χωρίζονται σε δεδομένα εκπαίδευσης που αποτελούνται από το 90% των αρχικών δεδομένων και δεδομένα ελέγχου που αποτελούνται από το 10% των αρχικών δεδομένων. Συνεπώς, τα νέα σύνολα που προκύπτουν είναι τα σύνολα x_train και y_train που θα χρησιμοποιηθούν για την εκπαίδευση του μοντέλου και τα x_test και y_test που θα χρησιμοποιηθούν για την αξιολόγηση του μοντέλου, κατά την οποία τα δεδομένα x_test θα αποτελέσουν την είσοδο του μοντέλου και τα αποτελέσματα που θα προκύψουν θα είναι οι προβλεπόμενες αξιολογήσεις. Τέλος, αυτές οι προβλεπόμενες αξιολογήσεις συγκρίνονται με τις πραγματικές τιμές από το σύνολο y_test και έτσι αξιολογείται η αποτελεσματικότητα του μοντέλου που δημιουργήθηκε.

Δημιουργία Χρηστών

Σε μία πραγματική εφαρμογή συνεργατικής μάθησης, κάθε συσκευή διατηρεί τα δεδομένα της αποθηκευμένα και τα επεξεργάζεται τοπικά ώστε να μην έχει πρόσβαση σε αυτά οποιοσδήποτε άλλος. Για την υλοποίηση αυτής της πειραματικής μελέτης δημιουργήθηκαν τρία σενάρια. Στην πρώτη περίπτωση δημιουργήθηκαν 10 clients, στην δεύτερη περίπτωση 20 clients και στην τρίτη περίπτωση 30 clients.

Για τον διαχωρισμό των δεδομένων σε 10, 20 και 30 clients δημιουργήθηκε ο πίνακας user-item, ο οποίος έχει ως γραμμές τους χρήστες, ως στήλες τις ταινίες και ως στοιχεία την αξιολόγηση που έχει δώσει ο συγκεκριμένος χρήστης στη συγκεκριμένη ταινία. Στη συνέχεια, έγινε συσταδοποίηση σε αυτόν τον πίνακα ώστε κάθε συστάδα να εμπεριέχει χρήστες που είναι παρόμοιοι μεταξύ τους. Τα δεδομένα των χρηστών κάθε συστάδας αντιπροσωπεύουν τα δεδομένα κάθε client αντίστοιχα. Ο αλγόριθμος συσταδοποίησης που επιλέχθηκε είναι ο K-means. Όπως αναφέρεται στο [22], ο αλγόριθμος K-means είναι μία ευρέως γνωστή μέθοδος συσταδοποίησης. Τα σημεία δεδομένων ταξινομούνται σε μία από τις k συστάδες, όπου το k έχει καθοριστεί εκ των προτέρων. Η ένταξη των σημείων δεδομένων στις συστάδες γίνεται αφού υπολογιστούν τα κέντρα της κάθε συστάδας (όπου είναι ο μέσος όρος). Στη συνέχεια κάθε αντικείμενο συμπεριλαμβάνεται στη συστάδα με το πλησιέστερο κέντρο. Αυτή η προσέγγιση ελαχιστοποιεί τη συνολική διασπορά εντός των συστάδων με επαναληπτική ανακατανομή των μελών των συστάδων. Συνεπώς, ο αλγόριθμος K-means δέχεται ως ορίσματα ένα σύνολο δεδομένων S και έναν ακέραιο k και έχει σαν έξοδο μία διαμέριση του συνόλου S στα υποσύνολα $S_1, \dots, S_k$.

Στην συγκεκριμένη υλοποίηση τα ορίσματα εισόδου του αλγορίθμου K-means ήταν ο πίνακας user-item και ο αριθμός k που πήρε τις τιμές 10, 20 και 30. Επομένως, στην πρώτη περίπτωση δημιουργήθηκαν 10 συστάδες όπου η κάθε μία αντιπροσωπεύει τα δεδομένα του καθενός από τους 10 clients. Στην δεύτερη περίπτωση δημιουργήθηκαν 20 συστάδες όπου η κάθε μία αντιπροσωπεύει τα δεδομένα του καθενός από τους 20 clients και αντίστοιχα στην τρίτη περίπτωση δημιουργήθηκαν

30 συστάδες όπου η κάθε μία αντιπροσωπεύει τα δεδομένα του καθενός από τους 30 clients.

Δημιουργία Μοντέλου Συστήματος Συστάσεων

Το μοντέλο που δημιουργήθηκε για την υλοποίηση της εργασίας είναι ένα νευρωνικό δίκτυο το οποίο αποτελείται από τέσσερα στρώματα (layers). Το μοντέλο δέχεται σαν είσοδο τον αριθμό των χρηστών, τον αριθμό των ταινιών και το embedding_size, το οποίο ορίζει το μέγεθος των διανυσμάτων που θα δημιουργηθούν στο μοντέλο. Παρακάτω παρουσιάζονται τα layers του μοντέλου:

1. **Input layer.** Το πρώτο στρώμα στο νευρωνικό δίκτυο είναι το στρώμα εισόδου όπου γίνεται η είσοδος των δεδομένων.
2. **Embedding layer.** Σε αυτό το στρώμα δημιουργούνται παράλληλα τα embeddings, δηλαδή τα διανύσματα για τους χρήστες και τις ταινίες. Τα διανύσματα αυτά είναι τα user_embedding και movie_embedding, τα οποία έχουν το καθένα διάσταση ίση με το embedding_size, που στη συγκεκριμένη εργασία έχει τιμή 50.
3. **Dot layer.** Σε αυτό το στρώμα τα δύο διανύσματα των χρηστών και των ταινιών συγχωνεύονται σε ένα μέσω του υπολογισμού του εσωτερικού γινομένου και έτσι δημιουργείται το διάνυσμα merged.
4. **Reshape layer.** Το στρώμα αυτό διορθώνει τη δομή (shape) του συγχωνευμένου διανύσματος.

Στην εικόνα 9 φαίνεται ο τύπος όλων των στρωμάτων (layer type) από τα οποία αποτελείται το νευρωνικό δίκτυο και η δομή της εξόδου (output shape) του κάθε στρώματος. Στη συνέχεια αναγράφεται ο αριθμός των παραμέτρων των στρωμάτων. Στο στρώμα Embedding για το διάνυσμα user_embedding ο αριθμός των παραμέτρων είναι 30500. Αυτό οφείλεται στο γεγονός ότι ο αριθμός των χρηστών, δηλαδή η είσοδος num_users του μοντέλου είναι 610 και εφόσον η διάσταση του διανύσματος που θα δημιουργηθεί για τον κάθε χρήστη (embedding_size) είναι 50, όπως οι παράμετροι είναι $50 * 610 = 30500$. Αντίστοιχα, στο ίδιο στρώμα για τη δημιουργία του διανύσματος movie_embedding ο αριθμός των παραμέτρων είναι 486200, το οποίο οφείλεται στο γεγονός ότι ο αριθμός των ταινιών (num_movies) είναι ίσος με 9724. Συνεπώς, εφόσον το κάθε διάνυσμα θα έχει μέγεθος 50 και θα δημιουργηθούν 9724 διανύσματα, ένα για κάθε ταινία, ο συνολικός αριθμός των παραμέτρων θα είναι $50 * 9724 = 486200$. Επιπλέον, στην εικόνα φαίνεται το στρώμα με το οποίο συνδέεται κάθε νέο στρώμα (Connected to) όπου σε όλες τις περιπτώσεις εκτός του στρώματος εισόδου, είναι η έξοδος του προηγούμενου.

```
model.summary()
```

```
Model: "model"
```

Layer (type)	Output Shape	Param #	Connected to
user (InputLayer)	[(None, 1)]	0	
movie (InputLayer)	[(None, 1)]	0	
user_embedding (Embedding)	(None, 1, 50)	30500	user[0][0]
movie_embedding (Embedding)	(None, 1, 50)	486200	movie[0][0]
dot_product (Dot)	(None, 1, 1)	0	user_embedding[0][0] movie_embedding[0][0]
reshape (Reshape)	(None, 1)	0	dot_product[0][0]

```
Total params: 516,700
Trainable params: 516,700
Non-trainable params: 0
```

Εικόνα 9. Περιγραφή των στρωμάτων του μοντέλου

Παρακάτω αναφέρονται οι τιμές που επιλέχθηκαν για ορισμένες παραμέτρους αναφορικά με το μοντέλο:

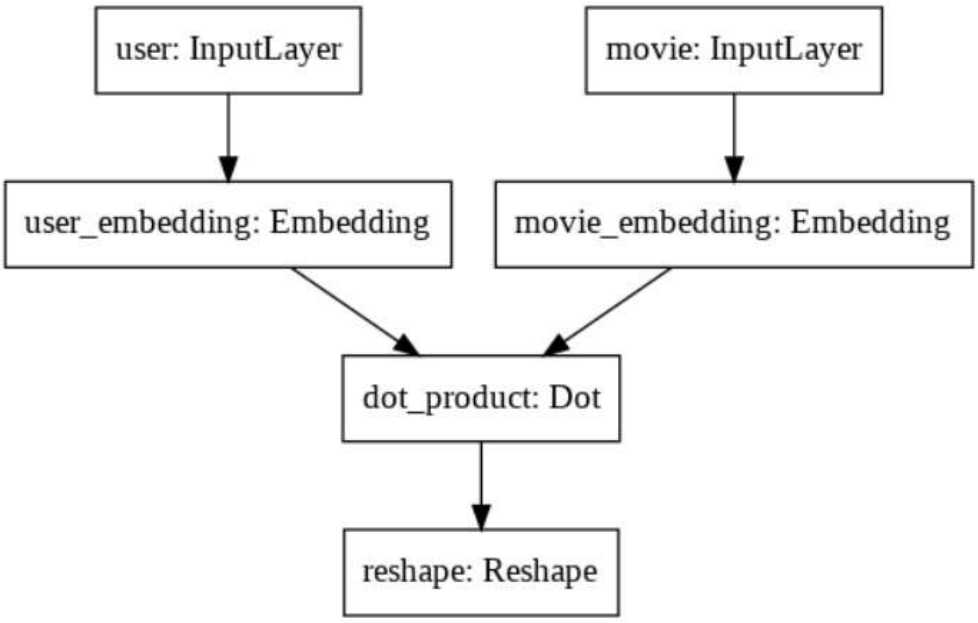
Ρυθμός εκμάθησης. Ο Ρυθμός εκμάθησης (learning rate) είναι μία παράμετρος η οποία ελέγχει το ρυθμό με τον οποίο ενημερώνονται οι παράμετροι του μοντέλου σε κάθε επανάληψη. Η τιμή του ρυθμού εκμάθησης σε αυτή την εργασία είναι 0.001.

Epochs. Αυτή η παράμετρος περιγράφει τον αριθμό των εποχών και η τιμή του είναι 1.

Steps. Με αυτή την παράμετρο δηλώνονται τα βήματα που θα γίνονται ανά epoch και τα οποία είναι 10.

Adam optimizer. Όπως αναφέρεται στο [20], ο Adam optimizer (Adaptive moment estimator) είναι μία από τις πιο δημοφιλείς μεθόδους βελτιστοποίησης στον τομέα των νευρωνικών δικτύων. Στο [21] φαίνεται μέσα από πειραματικές μελέτες ότι ο Adam optimizer συγκλίνει πολύ πιο γρήγορα για πολυεπίπεδα ή συνελικτικά νευρωνικά δίκτυα από οποιοδήποτε άλλο εργαλείο βελτιστοποίησης. Για αυτούς τους λόγους επιλέχθηκε και σε αυτήν την πειραματική μελέτη.

Στην εικόνα 10 φαίνεται η αρχιτεκτονική του νευρωνικού δικτύου που δημιουργήθηκε. Αρχικά, γίνεται παράλληλη είσοδος των χρηστών και των ταινιών, έπειτα δημιουργούνται παράλληλα τα embeddings για τους χρήστες και για τις ταινίες. Στη συνέχεια, συγχωνεύονται και τέλος διορθώνεται η δομή του τελικού διανύσματος.



Εικόνα 10. Αρχιτεκτονική Νευρωνικού Δικτύου

Μετρικές Αξιολόγησης

Η μετρική που επιλέχθηκε για την αξιολόγηση του συγκεκριμένου νευρωνικού δικτύου είναι το Mean Squared Error (MSE). Το Mean Squared Error υπολογίζει το μέσο τετραγωνικό σφάλμα, δηλαδή τον μέσο όρο του τετραγώνου της διαφοράς μεταξύ των προβλεπόμενων και των πραγματικών τιμών. Η τιμή του MSE δίνεται από τον παρακάτω τύπο:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (24)$$

όπου n είναι ο συνολικός αριθμός των όρων από τους οποίους υπολογίζεται το MSE, y_i είναι οι πραγματική τιμή και $\hat{y}_i$ είναι η τιμή που πρόβλεψε το μοντέλο. Το μέσο τετραγωνικό σφάλμα παίρνει μη αρνητικές τιμές και όσο πιο κοντά στο μηδέν βρίσκεται τόσο πιο αποδοτικό είναι το μοντέλο.

Συνεργατική Εκπαίδευση Μοντέλου

Αφού έχει δημιουργηθεί το μοντέλο μηχανικής μάθησης και έχουν διαμοιραστεί τα δεδομένα στους δέκα χρήστες μπορεί να ξεκινήσει η διαδικασία της συνεργατικής εκπαίδευσης του μοντέλου. Αρχικά, ορίζεται το νευρωνικό δίκτυο που δημιουργήθηκε ως το κεντρικό μοντέλο.

Οι γύροι επικοινωνίας μεταξύ του κεντρικού μοντέλου και των τοπικών μοντέλων έχουν οριστεί να είναι 55.

Στον πρώτο γύρο επικοινωνίας ορίζονται τα αρχικά βάρη του κεντρικού μοντέλου. Στις περισσότερες πειραματικές μελέτες συνεργατικής μάθησης τα βάρη ορίζονται ως μηδενικά. Σε αυτή την εργασία προκειμένου τα αρχικά βάρη να έχουν την απαιτούμενη δομή εκπαιδεύτηκε το κεντρικό μοντέλο με τις πρώτες δέκα εγγραφές από τα δεδομένα εκπαίδευσης. Αυτά τα βάρη του κεντρικού μοντέλου θα αποτελέσουν τα αρχικά βάρη για όλα τα τοπικά μοντέλα.

Στη συνέχεια, ξεκινάει μία νέα δομή επανάληψης στον αλγόριθμο η οποία βρίσκεται εσωτερικά της δομής επανάληψης για τον κάθε γύρο επικοινωνίας. Αυτή η εσωτερική δομή επανάληψης αφορά την εκπαίδευση των τοπικών μοντέλων και πραγματοποιείται για τα δεδομένα του κάθε χρήστη ξεχωριστά.

Σε κάθε επανάληψη για κάθε χρήστη ορίζεται ως τοπικό μοντέλο το νευρωνικό δίκτυο που έχει δημιουργηθεί με αρχικά βάρη τα βάρη που έχει το κεντρικό μοντέλο τη δεδομένη χρονική στιγμή. Έπειτα, το τοπικό μοντέλο που δημιουργήθηκε και με την συγκεκριμένη αρχικοποίηση των βαρών, εκπαιδεύεται με τα τοπικά δεδομένα του συγκεκριμένου χρήστη. Αφού έχει ολοκληρωθεί η εκπαίδευση συγκεντρώνονται τα νέα βάρη του τοπικού μοντέλου σε μία λίστα. Μόλις ολοκληρωθεί η εκπαίδευση των τοπικών μοντέλων όλων των χρηστών η λίστα αυτή θα περιέχει τα νέα τοπικά βάρη όλων των χρηστών για τον συγκεκριμένο γύρο επικοινωνίας.

Σε αυτή τη διπλωματική εργασία μελετήθηκε η δημιουργία προσωποποιημένων μοντέλων, τα οποία αποτελούν έναν κατάλληλο συνδυασμό του κεντρικού και του τοπικού μοντέλου. Συνεπώς, μετά τον υπολογισμό των βαρών του κάθε τοπικού μοντέλου υπολογίζονται τα προσωποποιημένα του βάρη με την παρακάτω σχέση:

$$Personalized\ weights = a * local\_weights + (1 - a) * global\_weights$$

όπου a είναι η παράμετρος μίξης των δύο μοντέλων. Όπως αναφέρεται στο [7], η βέλτιστη επιλογή για την τιμή του a εξαρτάται από την διαφορετικότητα μεταξύ των κατανομών των δεδομένων του κεντρικού και των τοπικών μοντέλων. Συγκεκριμένα, όταν η κατανομή του τοπικού μοντέλου είναι παρόμοια με το κεντρικό μία χαμηλή τιμή του a θα είναι περισσότερο αποδοτική καθώς το τοπικό μοντέλο επωφελείται και από τα δεδομένα των υπόλοιπων χρηστών. Αντίθετα, όταν η κατανομή του

τοπικού μοντέλου διαφέρει αρκετά από το κεντρικό τότε μία τιμή του a κοντά στη μονάδα θα είναι αποδοτικότερη καθώς δεν θα λαμβάνονται τόσο υπόψη τα δεδομένα των άλλων χρηστών.

Κατά την εκτέλεση αυτής της πειραματικής μελέτης δοκιμάστηκαν αρκετές τιμές για την παράμετρο a μεταξύ των τιμών 0 και 1. Η αποδοτικότερη τιμή βρέθηκε να είναι για $a = 0.1$.

Αφού υπολογιστούν τα προσωποποιημένα βάρη από κάθε τοπικό μοντέλο αποστέλλονται στον κεντρικό διακομιστή. Στην προσομοίωση που έγινε σε αυτή την εργασία, αυτά τα βάρη συγκεντρώθηκαν σε μία λίστα ώστε να επεξεργαστούν έξω από την δομή επανάληψης για τον κάθε χρήστη. Για την επεξεργασία των νέων βαρών που έχουν δημιουργηθεί υπολογίζεται ο μέσος όρος αυτών, ο οποίος στη συνέχεια τίθεται ως τα νέα βάρη του κεντρικού μοντέλου σύμφωνα με την τεχνική Federated Averaging όπως παρουσιάστηκε στο Κεφάλαιο 2.

Έχοντας φτάσει στο πέρας του συγκεκριμένου γύρου επικοινωνίας όλη η διαδικασία που περιγράφηκε παραπάνω επαναλαμβάνεται έως ότου να ολοκληρωθεί ο αριθμός των γύρων επικοινωνίας που έχει οριστεί και στην εργασία αυτή είναι 55.

Αποτελέσματα

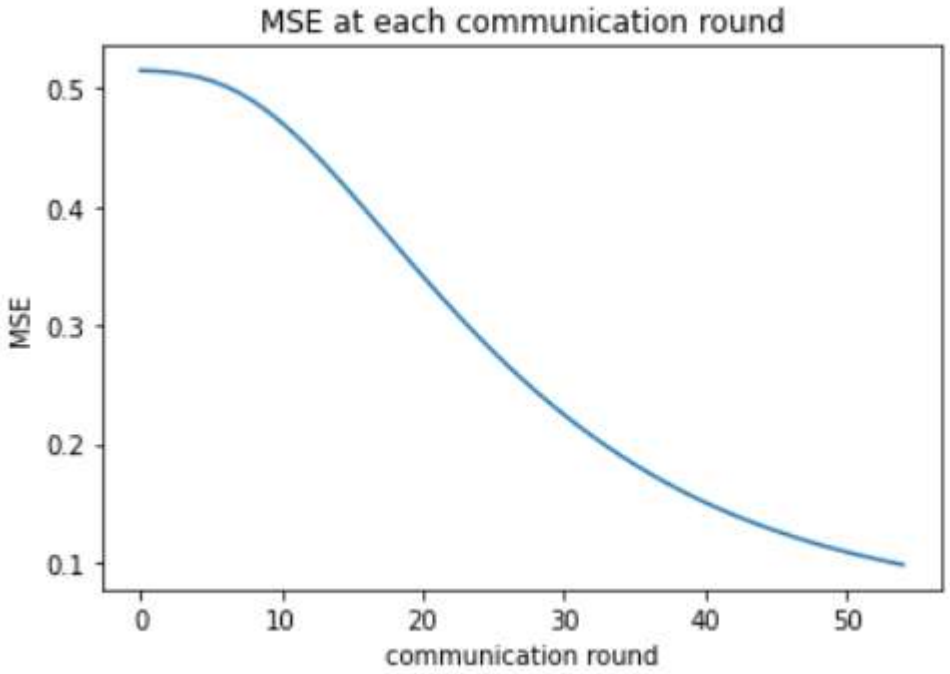
Στο τέλος κάθε γύρου επικοινωνίας υπολογίζεται το Mean Squared Error (MSE) του μοντέλου εκείνη τη χρονική στιγμή, καθώς έχει οριστεί ως η μετρική αξιολόγησης. Με αυτό τον τρόπο φαίνεται ο τρόπος με τον οποίο μεταβάλλεται το μέσο τετραγωνικό σφάλμα κατά τη διάρκεια της συνεργατικής εκπαίδευσης του μοντέλου σε κάθε γύρο επικοινωνίας.

Παρακάτω παρουσιάζονται τα αποτελέσματα που προέκυψαν για τις τρεις περιπτώσεις που υλοποιήθηκαν σε αυτή την πειραματική μελέτη και είναι οι εξής:

- Υλοποίηση με 10 clients
- Υλοποίηση με 20 clients
- Υλοποίηση με 30 clients

Υλοποίηση με 10 clients

Σε αυτή την υλοποίηση δημιουργήθηκαν 10 clients στους οποίους έγινε συνεργατική εκπαίδευση. Στο παρακάτω διάγραμμα φαίνεται η μεταβολή του Mean Squared Error κατά τους γύρους επικοινωνίας:



Εικόνα 11. Το Mean Squared Error κατά τους γύρους επικοινωνίας για 10 clients

Σε αυτή την υλοποίηση το Μέσο Τετραγωνικό Σφάλμα του κεντρικού μοντέλου μετά από την συνεργατική εκπαίδευση που πραγματοποιήθηκε στους δέκα clients είναι:

$$MSE = 0.09312889$$

Αυτή η τιμή του MSE βρίσκεται αρκετά κοντά στο 0, συνεπώς το αποτέλεσμα αυτό σημαίνει πως ο αλγόριθμος που δημιουργήθηκε και εκπαιδεύτηκε με τα συγκεκριμένα δεδομένα είναι αρκετά αποδοτικός. Συνεπώς, το κεντρικό μοντέλο που δημιουργήθηκε αποτελεί ένα σύστημα συστάσεων που μπορεί να προβλέψει την αξιολόγηση που θα έκανε ένας χρήστης για μία ταινία. Έτσι θα προτείνονται σε κάθε

χρήστη οι ταινίες που η προβλεπόμενη αξιολόγηση είναι υψηλή και άρα είναι πιθανότερο να παρακολουθήσει και θα αποφεύγονται οι προτάσεις για ταινίες που η προβλεπόμενη αξιολόγηση είναι χαμηλή και ίσως να μην επιθυμεί να δει ο χρήστης.

Στον πίνακα 1 παρουσιάζεται ο τρόπος με τον οποίο επηρεάζονται οι τιμές του Μέσου Τετραγωνικού Σφάλματος για τις διάφορες τιμές της παραμέτρου a που δοκιμάστηκαν:

a	MSE
0.9	0.11858336
0.8	0.1106482
0.7	0.1074194
0.6	0.10610023
0.5	0.09883619
0.4	0.09777329
0.3	0.09686816
0.2	0.09413089
0.1	0.09312889

Πίνακας 1. Οι τιμές του MSE για τις διάφορες τιμές του a

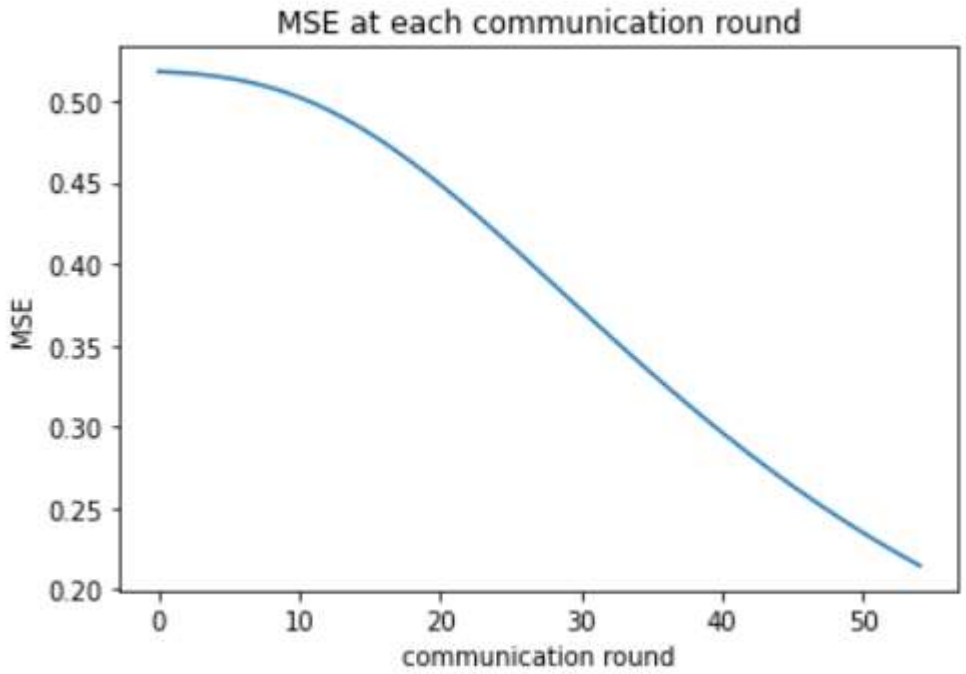
Όπως είναι φανερό καθώς μικραίνει η τιμή του a μειώνεται η τιμή του MSE και η βέλτιστη τιμή του είναι όταν $a = 0.1$.

Τέλος, ο χρόνος που απαιτήθηκε για να ολοκληρωθεί η εκπαίδευση του μοντέλου με 10 clients σε 55 γύρους επικοινωνίας είναι:

$$time \rightarrow 00:13:49$$

Υλοποίηση με 20 clients

Σε αυτή την υλοποίηση δημιουργήθηκαν 20 clients στους οποίους έγινε συνεργατική εκπαίδευση. Στο παρακάτω διάγραμμα φαίνεται η μεταβολή του Mean Squared Error κατά τους γύρους επικοινωνίας:



Εικόνα 12. Το Mean Squared Error κατά τους γύρους επικοινωνίας για 20 clients

Σε αυτή την υλοποίηση το Μέσο Τετραγωνικό Σφάλμα του κεντρικού μοντέλου μετά από την συνεργατική εκπαίδευση που πραγματοποιήθηκε στους είκοσι clients είναι:

$$MSE = 0.22593987$$

Αυτή η τιμή του MSE είναι αρκετά μεγαλύτερη από την τιμή του MSE στην υλοποίηση με 10 clients. Όπως είναι φανερό και από την εικόνα 12 το MSE δεν έχει αρχίσει να συγκλίνει σε μία τελική τιμή στους 55 γύρους επικοινωνίας.

Στον πίνακα 2 παρουσιάζεται ο τρόπος με τον οποίο επηρεάζονται οι τιμές του Μέσου Τετραγωνικού Σφάλματος για τις διάφορες τιμές της παραμέτρου a που δοκιμάστηκαν:

a	MSE
0.9	0.28639311
0.8	0.27956065
0.7	0.27408181
0.6	0.26105941
0.5	0.25352181
0.4	0.24343089
0.3	0.23678841
0.2	0.23166914
0.1	0.22593987

Πίνακας 2. Οι τιμές του MSE για τις διάφορες τιμές του a

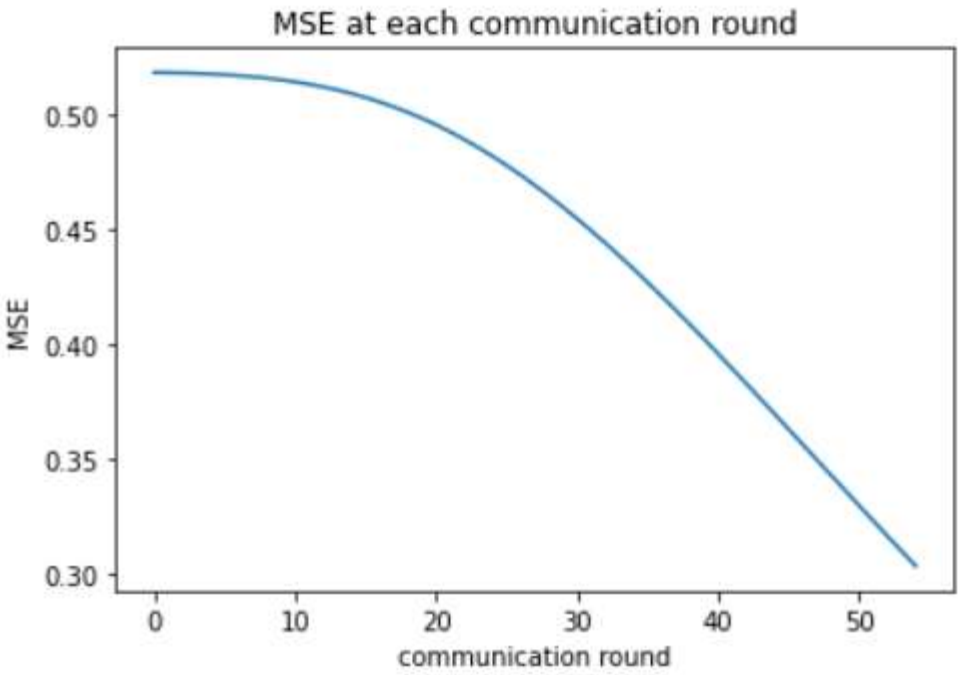
Όπως είναι φανερό καθώς μικραίνει η τιμή του a μειώνεται η τιμή του MSE και η βέλτιστη τιμή του είναι όταν $a = 0.1$.

Τέλος, ο χρόνος που απαιτήθηκε για να ολοκληρωθεί η εκπαίδευση του μοντέλου με 20 clients σε 55 γύρους επικοινωνίας είναι:

$$time \rightarrow 00:27:05$$

Υλοποίηση με 30 clients

Σε αυτή την υλοποίηση δημιουργήθηκαν 30 clients στους οποίους έγινε συνεργατική εκπαίδευση. Στο παρακάτω διάγραμμα φαίνεται η μεταβολή του Mean Squared Error κατά τους γύρους επικοινωνίας:



Εικόνα 13. Το Mean Squared Error κατά τους γύρους επικοινωνίας για 30 clients

Σε αυτή την υλοποίηση το Μέσο Τετραγωνικό Σφάλμα του κεντρικού μοντέλου μετά από την συνεργατική εκπαίδευση που πραγματοποιήθηκε στους είκοσι clients είναι:

$$MSE = 0.30974033$$

Αυτή η τιμή του MSE είναι αρκετά μεγαλύτερη από τις τιμές του MSE στις υλοποιήσεις με 10 και με 20 clients. Όπως είναι φανερό και από την εικόνα 13 το MSE δεν έχει αρχίσει να συγκλίνει σε μία τελική τιμή στους 55 γύρους επικοινωνίας.

Στον πίνακα 3 παρουσιάζεται ο τρόπος με τον οποίο επηρεάζονται οι τιμές του Μέσου Τετραγωνικού Σφάλματος για τις διάφορες τιμές της παραμέτρου a που δοκιμάστηκαν:

a	MSE
0.9	0.38939839
0.8	0.38556721
0.7	0.37150682
0.6	0.36750302
0.5	0.35367405
0.4	0.34005443
0.3	0.32660967
0.2	0.31345231
0.1	0.30974033

Πίνακας 3. Οι τιμές του MSE για τις διάφορες τιμές του α

Όπως είναι φανερό καθώς μικραίνει η τιμή του a μειώνεται η τιμή του MSE και η βέλτιστη τιμή του είναι όταν $a = 0.1$.

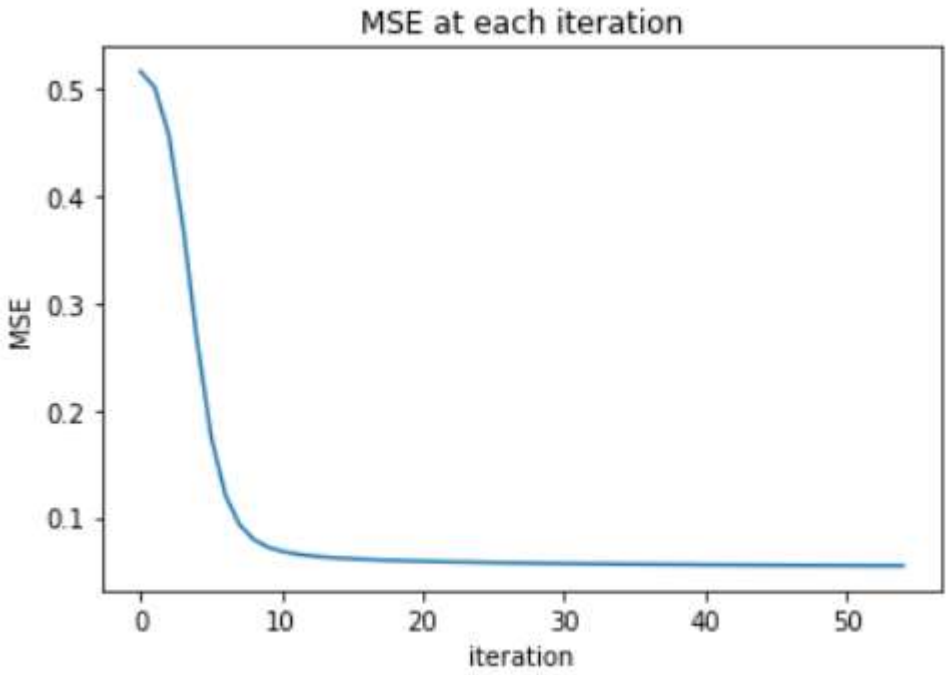
Τέλος, ο χρόνος που απαιτήθηκε για να ολοκληρωθεί η εκπαίδευση του μοντέλου με 30 clients σε 55 γύρους επικοινωνίας είναι:

$$time \rightarrow 00:41:34$$

4.6 Σύγκριση με Μοντέλο μη Συνεργατικής Μάθησης

Σε αυτή την ενότητα παρουσιάζονται τα αποτελέσματα της σύγκρισης του μοντέλου που δημιουργήθηκε σε αυτή την εργασία με μία απλή προσέγγιση μη συνεργατικής μάθησης. Για αυτό το σκοπό δημιουργήθηκε ένα νευρωνικό δίκτυο ίδιο με αυτό που παρουσιάστηκε στο Κεφάλαιο 4.5 με τις ίδιες παραμέτρους. Τα δεδομένα εισαγωγής στον νευρωνικό δίκτυο είναι όλο το σύνολο δεδομένων με την ίδια προεπεξεργασία με τη μόνη διαφορά ότι δεν διαχωρίζονται σε χρήστες. Η εκπαίδευση έγινε σε 55 επαναλήψεις καθώς στην προσέγγιση συνεργατικής μάθησης οι γύροι επικοινωνίας ήταν 55 και συνεπώς κάθε τοπικό μοντέλο εκπαιδευόταν σε 55 επαναλήψεις.

Στο παρακάτω διάγραμμα φαίνεται η μεταβολή του Mean Squared Error κατά τους γύρους επικοινωνίας:



Εικόνα 14. Το Mean Squared Error κατά τους γύρους επικοινωνίας απλού μοντέλου

Σε αυτή την υλοποίηση το Μέσο Τετραγωνικό Σφάλμα του κεντρικού μοντέλου μετά από την συνεργατική εκπαίδευση που πραγματοποιήθηκε στους είκοσι clients είναι:

$$MSE = 0.05672312$$

Αυτή η τιμή του MSE είναι αρκετά μικρότερη από τις τιμές του MSE στις υλοποιήσεις συνεργατικής μάθησης που παρουσιάστηκαν στην προηγούμενη ενότητα. Επιπλέον, από την εικόνα 14 φαίνεται ότι το Μέσο Τετραγωνικό Σφάλμα μειώνεται αρκετά πιο απότομα σε σχέση με τις συνεργατικές προσεγγίσεις και συγκλίνει σε πολύ λιγότερες επαναλήψεις. Συγκεκριμένα, από την όγδοη κιόλας επανάληψη έχει ήδη φτάσει την τιμή του MSE στην υλοποίηση με 10 clients μετά από 55 γύρους επικοινωνίας.

Από την σύγκριση των αποτελεσμάτων των μοντέλων προκύπτει ότι το μοντέλο μη συνεργατικής μάθησης συγκλίνει αρκετά γρηγορότερα από τα μοντέλα συνεργατικής μάθησης. Αυτό συμβαίνει καθώς το απλό μοντέλο εκπαιδεύεται με όλα τα δεδομένα και επομένως χρειάζεται λιγότερες επαναλήψεις για να επιτευχθεί η σύγκλιση. Αντίθετα, στα μοντέλα συνεργατικής μάθησης το κεντρικό μοντέλο δεν έχει πρόσβαση στα δεδομένα και το κάθε τοπικό μοντέλο εκπαιδεύεται με το $\frac{1}{10}$, $\frac{1}{20}$ και το $\frac{1}{30}$ των δεδομένων αντίστοιχα. Στο κεντρικό μοντέλο αποστέλλεται μόνο ο μέσος όρος από τα βάρη όλων των τοπικών μοντέλων. Συνεπώς, είναι αναμενόμενο ότι για την σύγκλιση του μοντέλου απαιτούνται περισσότερες επαναλήψεις.

Σύγκριση χρόνων εκτέλεσης

Όσον αφορά το χρόνο εκτέλεσης του αλγορίθμου της απλής προσέγγισης μετρήθηκε μόνο ο χρόνος που απαιτήθηκε για την εκπαίδευση του μοντέλου ο οποίος ήταν:

$$time = 00:00:31$$

Αυτός ο χρόνος είναι σημαντικά μικρότερος από τους χρόνους εκτέλεσης των προηγούμενων προσεγγίσεων οι οποίοι ήταν:

- Υλοποίηση με 10 clients: $00:14:01$
- Υλοποίηση με 20 clients: $00:27:33$
- Υλοποίηση με 30 clients: $00:41:19$

Από τους χρόνους όλων των υλοποιήσεων εξάγεται το συμπέρασμα πως όσο αυξάνεται ο αριθμός των clients τόσο αυξάνεται ο χρόνος εκτέλεσης. Το γεγονός αυτό οφείλεται στους εξής παράγοντες:

1. Αρχικά, η πειραματική μελέτη που έγινε για αυτή την εργασία αποτελεί μία προσομοίωση συνεργατικής μάθησης. Σε μία πραγματική εφαρμογή τα τοπικά μοντέλα εκπαιδεύονται σε διαφορετική συσκευή το καθένα γεγονός που επιτρέπει την παράλληλη εκπαίδευση τους. Σε αυτή την προσομοίωση όλη η εκτέλεση έγινε στον ίδιο υπολογιστή και τα τοπικά μοντέλα εκπαιδεύονταν το ένα μετά το άλλο. Δηλαδή, για να ξεκινήσει η τοπική εκπαίδευση του ενός τοπικού μοντέλου πρέπει να έχει ολοκληρωθεί πρώτα η εκπαίδευση του προηγούμενου. Κάτι τέτοιο είναι αρκετά ακριβό σε χρόνο σε σχέση με μία πραγματική εφαρμογή συνεργατικής μάθησης. Επίσης, στον αλγόριθμο της απλής προσέγγισης σε κάθε γύρο επανάληψης γίνεται μία εκπαίδευση μοντέλου με όλα τα δεδομένα μαζί, ενώ στον αλγόριθμο της συνεργατικής μάθησης σε κάθε γύρο επικοινωνίας εκπαιδεύονται 10, 20 ή 30 τοπικά μοντέλα και παρά το γεγονός ότι το κάθε μοντέλο εκπαιδεύεται με το $\frac{1}{10}$, $\frac{1}{20}$ ή το $\frac{1}{30}$ των δεδομένων αντίστοιχα είναι αρκετά πιο χρονοβόρα διαδικασία.
2. Επιπλέον, στην συνεργατική μάθηση η επικοινωνία μεταξύ του κεντρικού διακομιστή και των χρηστών είναι μία ακριβή διαδικασία. Στην συγκεκριμένη πειραματική μελέτη όλες οι ενημερώσεις των τοπικών μοντέλων συγκεντρώνονται σε μία λίστα, αποστέλλονται στο κεντρικό μοντέλο όπου εκεί υπολογίζεται ο μέσος όρος αυτών και στη συνέχεια ενημερώνονται εκ νέου τα βάρη του κεντρικού μοντέλου. Συνεπώς, η επικοινωνία αυτή είναι ένας παράγοντας που επηρεάζει το χρόνο εκπαίδευσης και έχει αρκετά μεγαλύτερη τιμή στην συνεργατική εκπαίδευση από ότι στην απλή εκπαίδευση.

Συμπεράσματα

Από τα αποτελέσματα που προέκυψαν με τη σύγκριση των δύο προσεγγίσεων μηχανικής μάθησης προκύπτουν ορισμένα πλεονεκτήματα και μειονεκτήματα και στις δύο περιπτώσεις. Η απλή προσέγγιση μηχανικής μάθησης όπου συγκεντρώνονται όλα τα δεδομένα στον κεντρικό διακομιστή και εκπαιδεύεται το μοντέλο έχει ορισμένα πλεονεκτήματα σε σύγκριση με την προσέγγιση συνεργατικής μάθησης. Αρχικά, το μοντέλο συγκλίνει πολύ ταχύτερα στην επιθυμητή τιμή του MSE και επομένως χρειάζονται αρκετά λιγότερες επαναλήψεις. Επιπλέον, ο χρόνος που απαιτείται για την ολοκλήρωση της εκπαίδευσης του μοντέλου είναι πολύ λιγότερος. Το βασικό μειονέκτημα αυτής της προσέγγισης είναι ότι σε μία πραγματική εφαρμογή δεν προστατεύονται τα δεδομένα των χρηστών, αφού συγκεντρώνονται ακατέργαστα στον κεντρικό διακομιστή. Από την άλλη μεριά το μοντέλο συνεργατικής μάθησης έχει το πλεονέκτημα ότι προστατεύει την ιδιωτικότητα των χρηστών που συμμετέχουν στην εκπαίδευση καθώς τα δεδομένα παραμένουν στην κάθε συσκευή. Το βασικό μειονέκτημα αυτού του μοντέλου είναι ότι χρειάζεται

αρκετά περισσότερος χρόνος για να εκπαιδευτεί το νευρωνικό δίκτυο από ότι στο απλό μοντέλο.

Επιπλέον, όσον αφορά τις τρεις περιπτώσεις συνεργατικής μάθησης που υλοποιήθηκαν προκύπτουν ορισμένα συμπεράσματα. Αρχικά, ο χρόνος εκτέλεσης αυξάνεται καθώς αυξάνεται και ο αριθμός των clients. Ακόμα, όσο αυξάνεται ο αριθμός των client τόσο αυξάνονται οι γύροι επικοινωνίας που απαιτούνται για τη σύγκλιση του μοντέλου. Συγκεκριμένα, σε 55 γύρους επικοινωνίας οι τιμές των MSE των τριών υλοποιήσεων ήταν:

- Υλοποίηση με 10 clients: 0.09312889
- Υλοποίηση με 20 clients: 0.22593987
- Υλοποίηση με 30 clients: 0.30974033

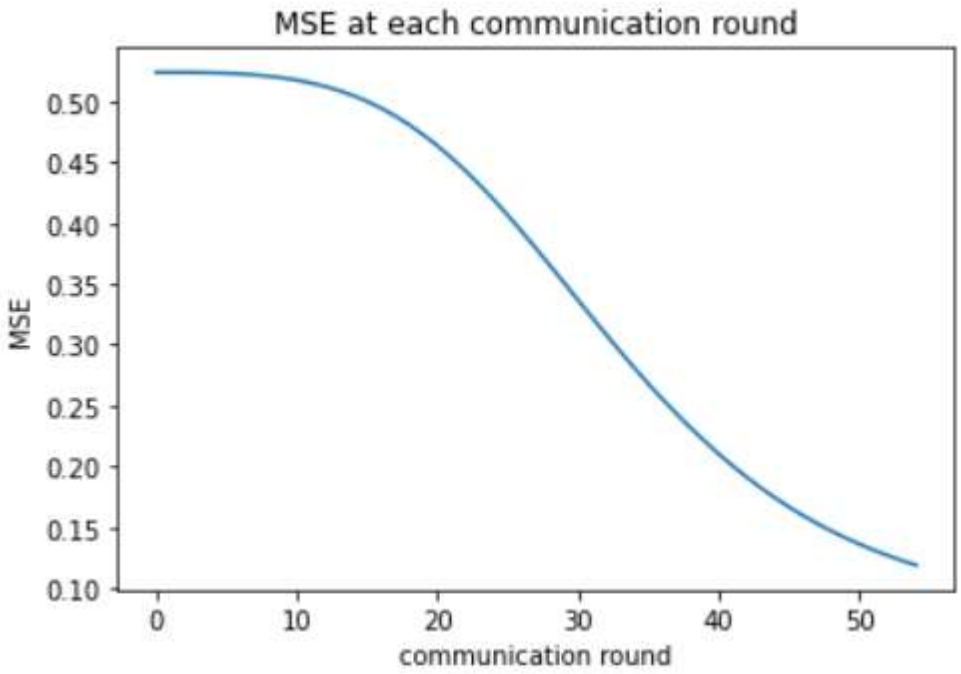
Το αποτέλεσμα της υλοποίησης με 10 clients είναι εμφανώς μικρότερο από τα αποτελέσματα των άλλων δύο υλοποιήσεων και αντίστοιχα το αποτέλεσμα της υλοποίησης με 20 clients είναι εμφανώς μικρότερο από το αποτέλεσμα της υλοποίησης με 30 clients. Το MSE της υλοποίησης με 10 clients σε 55 γύρους επικοινωνίας έχει ήδη αρχίσει να συγκλίνει σε αντίθεση με τα MSE των άλλων δύο υλοποιήσεων. Συνεπώς, για να επιτευχθεί σύγκλιση και στις άλλες δύο περιπτώσεις απαιτούνται παραπάνω γύροι επικοινωνίας.

4.7 Σύγκριση με μη Προσωποποιημένο Μοντέλο

Σε αυτή την ενότητα γίνεται σύγκριση των μοντέλων συνεργατικής μάθησης που παρουσιάστηκαν στην ενότητα 4.5 αυτού του Κεφαλαίου με τρία ίδια μοντέλα χωρίς την προσωποποιημένη προσέγγιση. Για το λόγο αυτό δημιουργήθηκαν τρία ίδια μοντέλα με τις ίδιες ακριβώς παραμέτρους. Το σύνολο δεδομένων είναι το ίδιο, γίνεται η ίδια προεπεξεργασία και υλοποιούνται τρεις περιπτώσεις με 10, 20 και 30 clients. Η μόνη διαφορά είναι ότι μετά την κάθε τοπική εκπαίδευση δεν υπολογίζονται τα προσωποποιημένα βάρη αλλά αποστέλλονται κατευθείαν τα τοπικά βάρη στον κεντρικό διακομιστή χωρίς κάποια επεξεργασία.

Υλοποίηση με 10 clients χωρίς προσωποποιημένη προσέγγιση

Σε αυτή την υλοποίηση δημιουργήθηκαν 10 clients στους οποίους έγινε συνεργατική εκπαίδευση χωρίς την προσωποποιημένη προσέγγιση. Στο παρακάτω διάγραμμα φαίνεται η μεταβολή του Mean Squared Error κατά τους γύρους επικοινωνίας:



Εικόνα 15. Το MSE με 10 clients κατά τους γύρους επικοινωνίας μη προσωποποιημένης προσέγγισης

Σε αυτή την υλοποίηση το Μέσο Τετραγωνικό Σφάλμα του κεντρικού μοντέλου μετά από την μη προσωποποιημένη συνεργατική εκπαίδευση που πραγματοποιήθηκε στους δέκα clients μετά από 55 γύρους επικοινωνίας είναι:

$$MSE = 0.11910178$$

Παρατηρείται ότι η τιμή αυτή είναι ελαφρώς μεγαλύτερη από την τιμή του MSE που προέκυψε στο μοντέλο προσωποποιημένης συνεργατικής μάθησης με 10 clients, η οποία είναι:

$$MSE = 0.09312889$$

Η αύξηση αυτή οφείλεται στην έλλειψη της προσωποποιημένης προσέγγισης στην συνεργατική εκπαίδευση του μοντέλου.

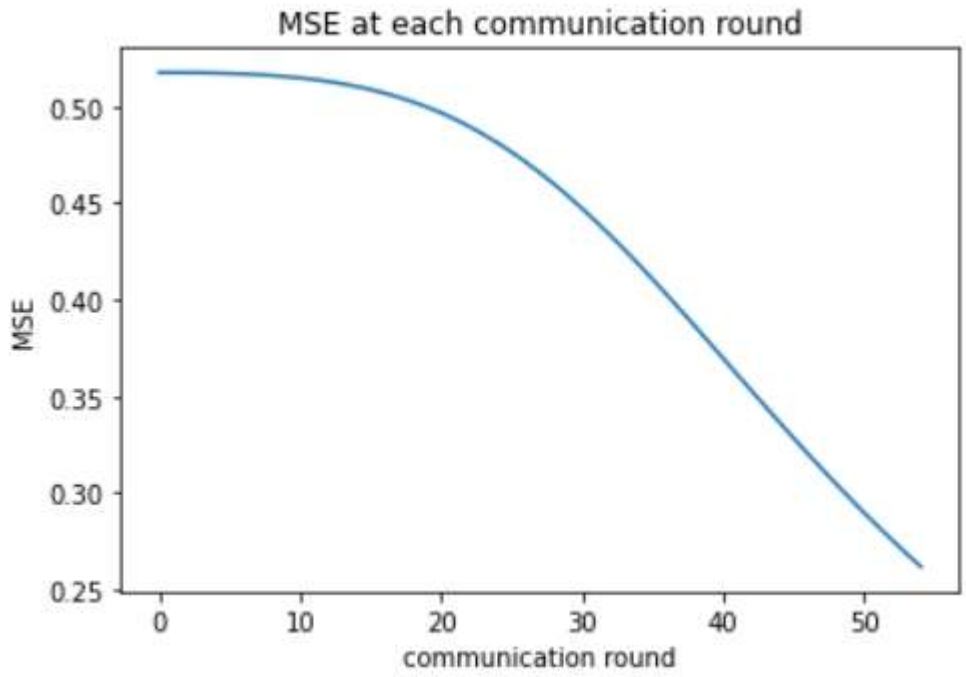
Τέλος, ο χρόνος που απαιτήθηκε για να ολοκληρωθεί η εκπαίδευση του μοντέλου με 10 clients σε 55 γύρους επικοινωνίας είναι:

$$time \rightarrow 00:14:07$$

Ο χρόνος εκτέλεσης της μη προσωποποιημένης προσέγγισης είναι σχεδόν ίδιος με τον χρόνο εκτέλεσης της προσωποποιημένης προσέγγισης.

Υλοποίηση με 20 clients χωρίς προσωποποιημένη προσέγγιση

Σε αυτή την υλοποίηση δημιουργήθηκαν 20 clients στους οποίους έγινε συνεργατική εκπαίδευση χωρίς την προσωποποιημένη προσέγγιση. Στο παρακάτω διάγραμμα φαίνεται η μεταβολή του Mean Squared Error κατά τους γύρους επικοινωνίας:



Εικόνα 16. Το MSE με 20 clients κατά τους γύρους επικοινωνίας μη προσωποποιημένης προσέγγισης

Σε αυτή την υλοποίηση το Μέσο Τετραγωνικό Σφάλμα του κεντρικού μοντέλου μετά από την μη προσωποποιημένη συνεργατική εκπαίδευση που πραγματοποιήθηκε στους είκοσι clients μετά από 55 γύρους επικοινωνίας είναι:

$$MSE = 0.27041891$$

Παρατηρείται ότι η τιμή αυτή είναι αρκετά μεγαλύτερη από την τιμή του MSE που προέκυψε στο μοντέλο προσωποποιημένης συνεργατικής μάθησης με 20 clients, η οποία είναι:

$$MSE = 0.22593987$$

Η αύξηση αυτή οφείλεται στην έλλειψη της προσωποποιημένης προσέγγισης στην συνεργατική εκπαίδευση του μοντέλου.

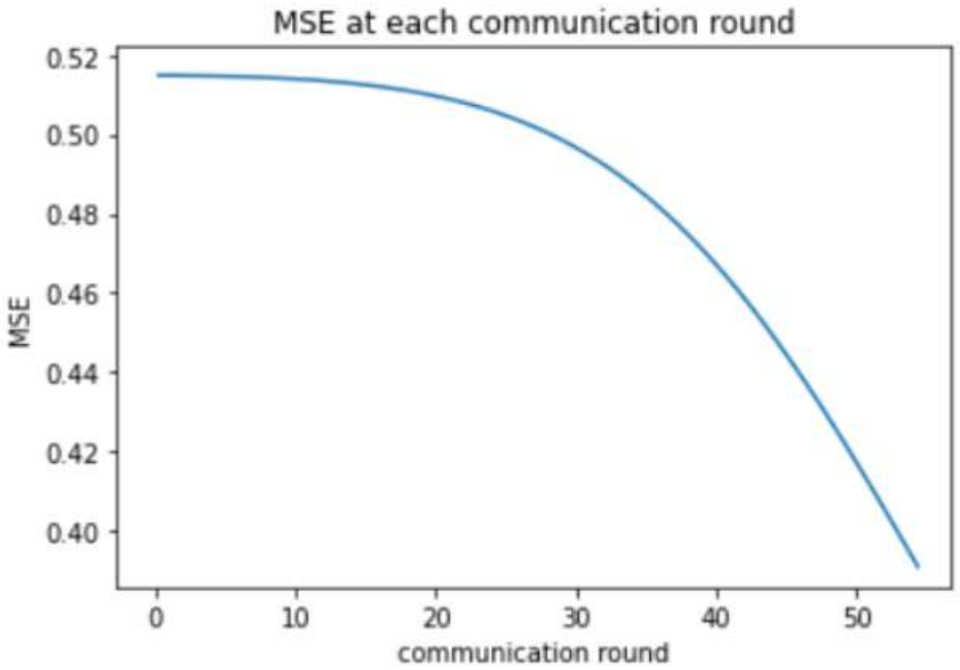
Τέλος, ο χρόνος που απαιτήθηκε για να ολοκληρωθεί η εκπαίδευση του μοντέλου με 20 clients σε 55 γύρους επικοινωνίας είναι:

$$time \rightarrow 00:27:21$$

Ο χρόνος εκτέλεσης της μη προσωποποιημένης προσέγγισης είναι αρκετά παρόμοιος με τον χρόνο εκτέλεσης της προσωποποιημένης προσέγγισης.

Υλοποίηση με 30 clients χωρίς προσωποποιημένη προσέγγιση

Σε αυτή την υλοποίηση δημιουργήθηκαν 30 clients στους οποίους έγινε συνεργατική εκπαίδευση χωρίς την προσωποποιημένη προσέγγιση. Στο παρακάτω διάγραμμα φαίνεται η μεταβολή του Mean Squared Error κατά τους γύρους επικοινωνίας:



Εικόνα 17. Το MSE με 30 clients κατά τους γύρους επικοινωνίας μη προσωποποιημένης προσέγγισης

Σε αυτή την υλοποίηση το Μέσο Τετραγωνικό Σφάλμα του κεντρικού μοντέλου μετά από την μη προσωποποιημένη συνεργατική εκπαίδευση που πραγματοποιήθηκε στους τριάντα clients μετά από 55 γύρους επικοινωνίας είναι:

$$MSE = 0.38939839$$

Παρατηρείται ότι η τιμή αυτή είναι αρκετά μεγαλύτερη από την τιμή του MSE που προέκυψε στο μοντέλο προσωποποιημένης συνεργατικής μάθησης με 30 clients, η οποία είναι:

$$MSE = 0.30974033$$

Η αύξηση αυτή οφείλεται στην έλλειψη της προσωποποιημένης προσέγγισης στην συνεργατική εκπαίδευση του μοντέλου.

Τέλος, ο χρόνος που απαιτήθηκε για να ολοκληρωθεί η εκπαίδευση του μοντέλου με 30 clients σε 55 γύρους επικοινωνίας είναι:

$$time \rightarrow 00:40:56$$

Ο χρόνος εκτέλεσης της μη προσωποποιημένης προσέγγισης είναι σχεδόν ίδιος με τον χρόνο εκτέλεσης της προσωποποιημένης προσέγγισης.

Συμπεράσματα

Από τη σύγκριση των μοντέλων προκύπτουν ορισμένα συμπεράσματα ως προς την αξιολόγησή τους. Στον παρακάτω πίνακα παρουσιάζονται οι τιμές του MSE με 10, 20 και 30 clients:

Αριθμός clients	MSE προσωποποιημένης προσέγγισης	MSE μη προσωποποιημένης προσέγγισης
10	**0.09312889**	**0.11910178**
20	**0.22593987**	**0.28141891**
30	**0.30974033**	**0.38939839**

Πίνακας 4. Σύγκριση των τιμών του MSE

Από τον πίνακα 4 είναι φανερό ότι σε όλες τις περιπτώσεις τα μοντέλα προσωποποιημένης προσέγγισης παρουσιάζουν μεγαλύτερη αποδοτικότητα από ότι τα μοντέλα μη προσωποποιημένης προσέγγισης. Το γεγονός αυτό σημαίνει ότι η προσωποποιημένη προσέγγιση μπορεί να ευνοήσει τη συνεργατική μάθηση και να δημιουργούνται αποδοτικότερα μοντέλα.

Επιπλέον, από την παρατήρηση του πίνακα 4 προκύπτει ότι όσο αυξάνεται ο αριθμός των clients τόσο αυξάνεται και η διαφορά ανάμεσα στις δύο προσεγγίσεις και συγκεκριμένα τόσο περισσότερο βελτιώνει την αποδοτικότητα η προσωποποιημένη προσέγγιση. Αυτό οφείλεται στο γεγονός ότι με την αύξηση του αριθμού των clients η ανομοιογένεια ανάμεσα στις κατανομές των δεδομένων των clients είναι μεγαλύτερη και επομένως ο υπολογισμός των προσωποποιημένων μοντέλων είναι περισσότερο ευνοϊκός.

Στην προσωποποιημένη προσέγγιση αντί να στέλνονται τα βάρη των τοπικών μοντέλων στον κεντρικό διακομιστή, υπολογίζονται τα βάρη των προσωποποιημένων μοντέλων που αποτελούν μία μίξη του εκάστοτε τοπικού μοντέλου και του κεντρικού. Κάτι τέτοιο είναι όλο και περισσότερο ευνοϊκό όσο αυξάνεται η ανομοιογένεια ανάμεσα στους clients. Αυτό συμβαίνει καθώς όταν η τοπική κατανομή συσχετίζεται σε μεγάλο βαθμό με την κεντρική κατανομή είναι προτιμότερο το κεντρικό μοντέλο. Όμως σε αντίθετη περίπτωση το κεντρικό μοντέλο μπορεί να μην είναι αποτελεσματικό για να χρησιμοποιηθεί ως τοπικό μοντέλο. Συνεπώς, ο συνδυασμός του κεντρικού και του τοπικού μοντέλου με μία κατάλληλη παράμετρο μίξης είναι η βέλτιστη επιλογή.

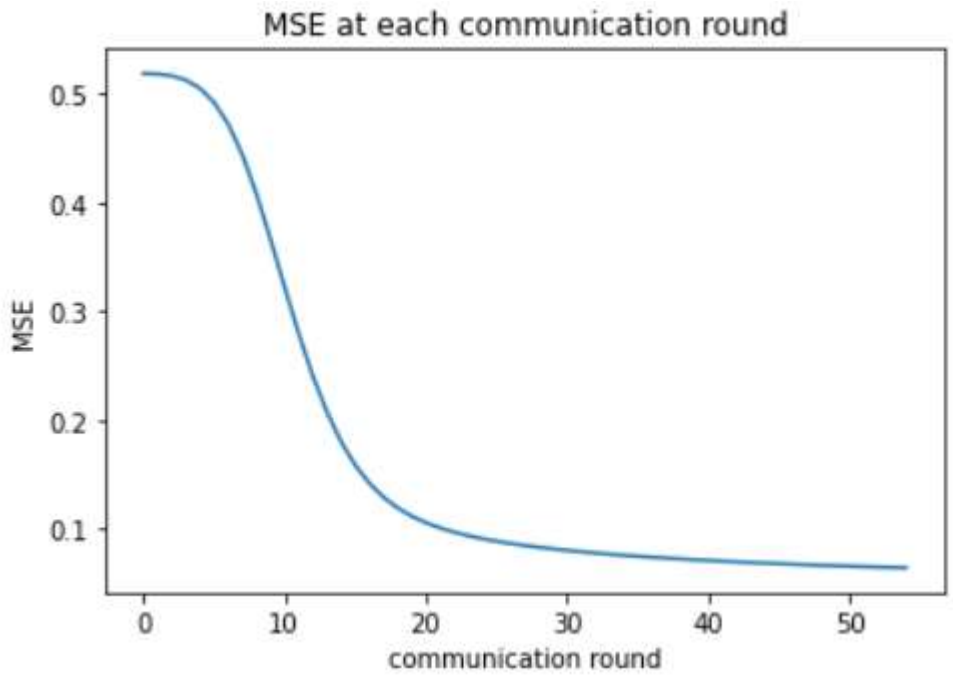
Όσον αφορά το χρόνο εκτέλεσης των δύο προσεγγίσεων δεν παρουσιάζεται σημαντική διαφορά, γεγονός που σημαίνει ότι η προσωποποιημένη προσέγγιση μπορεί να βελτιώσει την αποδοτικότητα της συνεργατικής εκπαίδευσης μοντέλων χωρίς να επιβαρύνει το χρόνο που απαιτείται.

4.8 Σύγκριση με Δημιουργία Χρηστών με Τυχαίο Τρόπο

Σε αυτή την ενότητα γίνεται σύγκριση των μοντέλων συνεργατικής μάθησης που παρουσιάστηκαν στην ενότητα 4.5 αυτού του Κεφαλαίου με τρία ίδια μοντέλα με την μόνη διαφορά να είναι ο τρόπος δημιουργίας των χρηστών. Για το λόγο αυτό τα δεδομένα εκπαίδευσης διαχωρίστηκαν τυχαία και ισόποσα στον κάθε client και δημιουργήθηκαν τρία ίδια μοντέλα με τις ίδιες ακριβώς παραμέτρους όπου και εκπαιδεύτηκαν με τις τεχνικές της προσωποποιημένης συνεργατικής μάθησης. Υλοποιήθηκαν αντίστοιχα τρεις περιπτώσεις με 10, 20 και 30 clients.

Υλοποίηση με 10 clients

Σε αυτή την υλοποίηση δημιουργήθηκαν 10 clients με τυχαίο διαχωρισμό των δεδομένων, στους οποίους έγινε συνεργατική εκπαίδευση. Στο παρακάτω διάγραμμα φαίνεται η μεταβολή του Mean Squared Error κατά τους γύρους επικοινωνίας:



Εικόνα 18. Το MSE με 10 clients κατά τους γύρους επικοινωνίας με τυχαίο διαχωρισμό των δεδομένων

Σε αυτή την υλοποίηση το Μέσο Τετραγωνικό Σφάλμα του κεντρικού μοντέλου μετά από την συνεργατική εκπαίδευση που πραγματοποιήθηκε στους δέκα clients μετά από 55 γύρους επικοινωνίας είναι:

$$MSE = 0.06455368$$

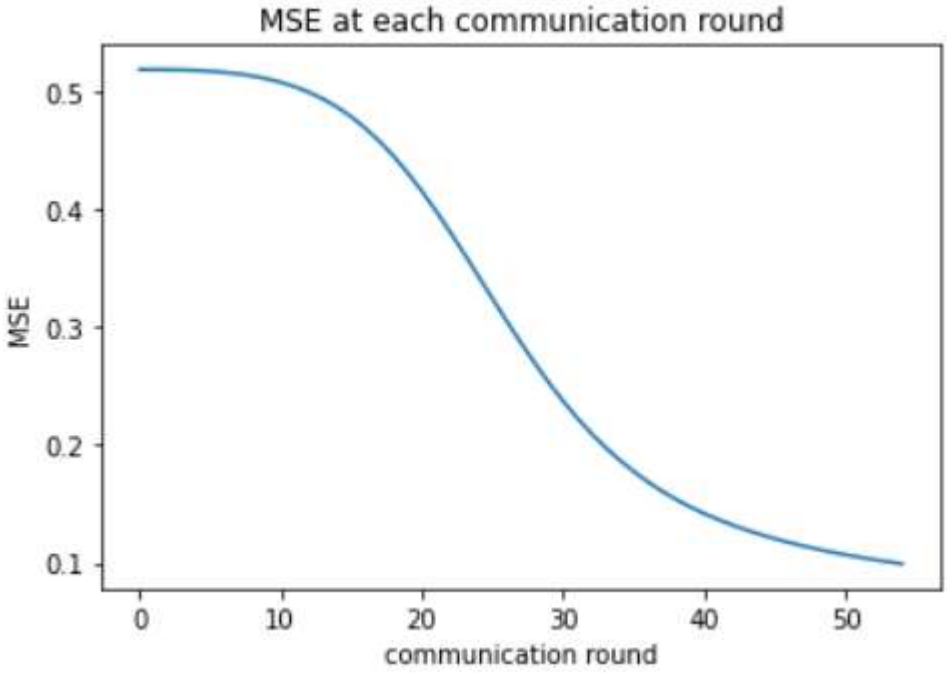
Ο χρόνος που απαιτήθηκε για να ολοκληρωθεί η εκπαίδευση του μοντέλου με 10 clients σε 55 γύρους επικοινωνίας είναι:

$$time \rightarrow 00:13:39$$

Ο χρόνος εκτέλεσης αυτής της προσέγγισης δεν παρουσιάζει διαφορά με τον χρόνο εκτέλεσης της αρχικής προσέγγισης.

Υλοποίηση με 20 clients

Σε αυτή την υλοποίηση δημιουργήθηκαν 20 clients με τυχαίο διαχωρισμό των δεδομένων, στους οποίους έγινε συνεργατική εκπαίδευση. Στο παρακάτω διάγραμμα φαίνεται η μεταβολή του Mean Squared Error κατά τους γύρους επικοινωνίας:



Εικόνα 19. Το MSE με 20 clients κατά τους γύρους επικοινωνίας με τυχαίο διαχωρισμό των δεδομένων

Σε αυτή την υλοποίηση το Μέσο Τετραγωνικό Σφάλμα του κεντρικού μοντέλου μετά από την συνεργατική εκπαίδευση που πραγματοποιήθηκε στους είκοσι clients μετά από 55 γύρους επικοινωνίας είναι:

$$MSE = 0.09943834$$

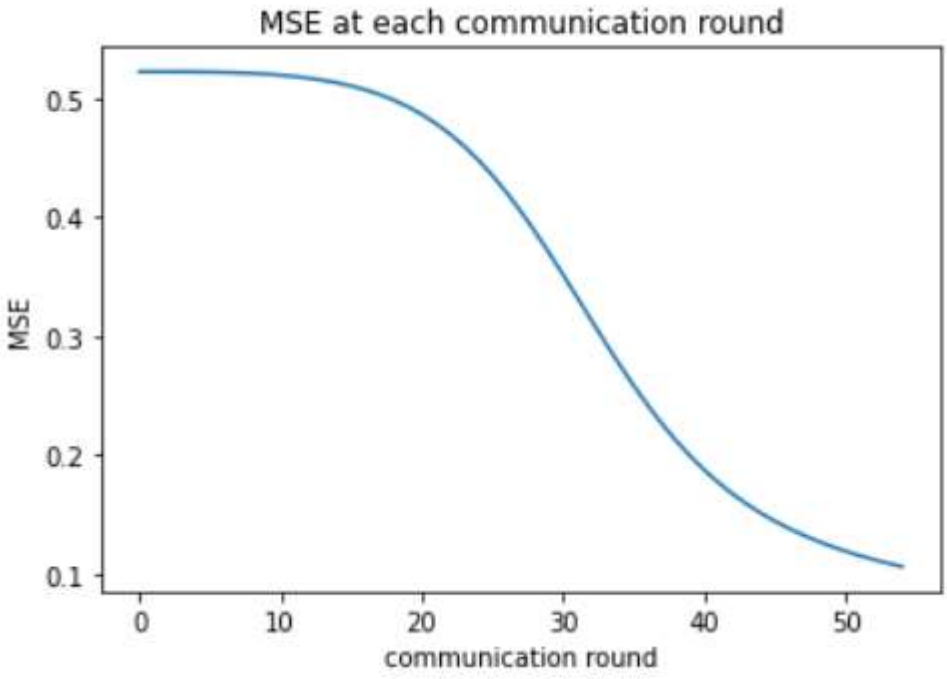
Ο χρόνος που απαιτήθηκε για να ολοκληρωθεί η εκπαίδευση του μοντέλου με 20 clients σε 55 γύρους επικοινωνίας είναι:

$$time \rightarrow 00:28:02$$

Ο χρόνος εκτέλεσης αυτής της προσέγγισης δεν παρουσιάζει διαφορά με τον χρόνο εκτέλεσης της αρχικής προσέγγισης.

Υλοποίηση με 30 clients

Σε αυτή την υλοποίηση δημιουργήθηκαν 30 clients με τυχαίο διαχωρισμό των δεδομένων, στους οποίους έγινε συνεργατική εκπαίδευση. Στο παρακάτω διάγραμμα φαίνεται η μεταβολή του Mean Squared Error κατά τους γύρους επικοινωνίας:



Εικόνα 20. Το MSE με 30 clients κατά τους γύρους επικοινωνίας με τυχαίο διαχωρισμό των δεδομένων

Σε αυτή την υλοποίηση το Μέσο Τετραγωνικό Σφάλμα του κεντρικού μοντέλου μετά από την συνεργατική εκπαίδευση που πραγματοποιήθηκε στους είκοσι clients μετά από 55 γύρους επικοινωνίας είναι:

$$MSE = 0.106587$$

Ο χρόνος που απαιτήθηκε για να ολοκληρωθεί η εκπαίδευση του μοντέλου με 30 clients σε 55 γύρους επικοινωνίας είναι:

$$time \rightarrow 00:41:18$$

Ο χρόνος εκτέλεσης αυτής της προσέγγισης δεν παρουσιάζει διαφορά με τον χρόνο εκτέλεσης της αρχικής προσέγγισης.

Συμπεράσματα

Από τη σύγκριση των μοντέλων προκύπτουν ορισμένα συμπεράσματα ως προς την αξιολόγησή τους. Στις περιπτώσεις που οι clients δημιουργήθηκαν με τυχαίο τρόπο και με ίσο αριθμό δεδομένων ο καθένας τα MSE των μοντέλων χρειάζονταν αρκετά λιγότερους γύρους επικοινωνίας για να συγκλίνουν. Επιπλέον, είχαν φτάσει τις τιμές των MSE των αρχικών περιπτώσεων με τη δημιουργία χρηστών μέσω συσταδοποίησης, ήδη σε 25 γύρους επικοινωνίας. Αυτό οφείλεται στο γεγονός ότι στις αρχικές περιπτώσεις τα δεδομένα των clients παρουσίαζαν μεγαλύτερη ετερογένεια μεταξύ τους και επομένως τα μοντέλα χρειάζονταν περισσότερες επαναλήψεις ώστε να συγκλίνουν.

Κεφάλαιο 5: Συμπεράσματα και Μελλοντική Εργασία

Το κεφάλαιο αυτό περιλαμβάνει τα συμπεράσματα που εξήχθησαν κατά την εκπόνηση αυτής της διπλωματικής εργασίας και παρουσιάζει πιθανές μελλοντικές επεκτάσεις.

5.1 Σύνοψη

Σε αυτή την διπλωματική εργασία εξετάστηκε η προσέγγιση της συνεργατικής μάθησης στα συστήματα συστάσεων. Αρχικά, περιγράφηκε η διαδικασία της συνεργατικής μάθησης, όπου έχει σαν στόχο την εκπαίδευση ενός κεντρικού μοντέλου από δεδομένα τα οποία βρίσκονται κατανεμημένα σε απομακρυσμένες συσκευές χωρίς αυτά να δεσμεύονται και να μεταφέρονται σε κάποιον κεντρικό διακομιστή. Επίσης, παρουσιάστηκε η αρχιτεκτονική ενός τέτοιου μοντέλου και οι βασικές προκλήσεις που εμφανίζονται. Επιπλέον, περιγράφηκαν τα συστήματα συστάσεων, τα οποία επεξεργάζονται τις πληροφορίες που αφορούν τις προτιμήσεις των χρηστών και αφού κάνουν τις σχετικές προβλέψεις, προτείνουν στον κάθε χρήστη προϊόντα και υπηρεσίες που είναι πιο πιθανό να προτιμάει.

Στη συνέχεια, παρουσιάστηκε η υλοποίηση που έγινε σε αυτή την εργασία. Συγκεκριμένα, το μοντέλο που δημιουργήθηκε που ήταν ένα νευρωνικό δίκτυο ενσωμάτωσης, οι τεχνικές συνεργατικής μάθησης που χρησιμοποιήθηκαν και η προσέγγιση της προσωποποιημένης μάθησης όπου υπολογίζονται προσωποποιημένα μοντέλα για κάθε χρήστη.

Τέλος, περιγράφηκε το πλαίσιο της πειραματικής μελέτης που έγινε, τα τεχνικά μέσα που χρησιμοποιήθηκαν και ο αλγόριθμος που αναπτύχθηκε. Η πειραματική μελέτη αφορούσε τη δημιουργία μοντέλου που προβλέπει την αξιολόγηση που θα έκανε ένας χρήστης για ένα σύνολο ταινιών με σκοπό να του προτείνει τις ταινίες για τις οποίες προβλέφθηκε η μεγαλύτερη αξιολόγηση και άρα είναι πιο πιθανό να επιλέξει να δει. Για αυτές τις ανάγκες έγιναν τρεις υλοποιήσεις με διαφοροποίηση στον αριθμό των client όπου οι τιμές ήταν 10, 20 και 30. Αφού εκπαιδεύτηκαν αυτά τα μοντέλα με συνεργατικό τρόπο παρουσιάστηκαν τα αποτελέσματα που προέκυψαν και έγινε η αξιολόγηση.

5.2 Συμπεράσματα

Συμπερασματικά, μέσα από το θεωρητικό υπόβαθρο της διπλωματικής εργασίας, την υλοποίηση και την πειραματική μελέτη αλλά και την τελική αξιολόγηση αποδείχθηκε ότι η συνεργατική μάθηση είναι μία αποτελεσματική τεχνική μηχανικής μάθησης στο πλαίσιο της οποίας μπορούν να εκπαιδευτούν αποτελεσματικά μοντέλα συστημάτων συστάσεων. Κατά τη σύγκριση με ένα μοντέλο μη συνεργατικής μάθησης προκύπτει ότι αυτό συγκλίνει ταχύτερα και χρειάζεται λιγότερους γύρους επανάληψης για να φτάσει τα ίδια αποτελέσματα. Ωστόσο, τα προσωπικά δεδομένα του κάθε χρήστη δεν προστατεύονται όπως στο παράδειγμα της συνεργατικής μάθησης. Επιπλέον, κατά τη σύγκριση με τα μη προσωποποιημένα μοντέλα παρατηρήθηκε ότι η τιμή του Μέσου Τετραγωνικού Σφάλματος ήταν μεγαλύτερη από ότι στα προσωποποιημένα μοντέλα, ενώ ο χρόνος εκτέλεσης ήταν ο ίδιος. Συγκεκριμένα, όσο αυξανόταν ο αριθμός των clients και άρα και η ανομοιογένεια ανάμεσα στους clients, τόσο μεγαλύτερη ήταν η βελτίωση από την προσωποποιημένη προσέγγιση. Επομένως, ο υπολογισμός των προσωποποιημένων μοντέλων μπορεί να βελτιώσει την απόδοση των μοντέλων που εκπαιδεύονται με τεχνικές συνεργατικής μάθησης και όσο μεγαλύτερη είναι η διαφοροποίηση ανάμεσα στις τοπικές και στις κεντρικές κατανομές των μοντέλων τόσο περισσότερο σημαντική είναι η χρήση προσωποποιημένης προσέγγισης.

5.3 Μελλοντικές επεκτάσεις

Μία μελλοντική επέκταση αυτής της πειραματικής μελέτης είναι η υλοποίηση σε περιβάλλον όπου τα δεδομένα του κάθε χρήστη θα βρίσκονται σε ξεχωριστή συσκευή. Με αυτόν τον τρόπο θα μπορεί να γίνεται παράλληλη εκπαίδευση των τοπικών μοντέλων και να εξαχθούν συμπεράσματα σχετικά με την διαφοροποίηση που ενδεχομένως να προκύψει στο χρόνο εκτέλεσης. Μία ακόμα μελλοντική επέκταση αυτής της εργασίας είναι να μελετηθεί πως θα αντιμετωπιστεί η περίπτωση εμφάνισης κακόβουλων χρηστών στα πλαίσια ενός συστήματος συνεργατικής μάθησης.

Βιβλιογραφία

[1]. T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, “Federated learning: Challenges, methods, and future directions,” IEEE Signal Processing Magazine, vol. 37, no. 3, pp. 50–60, 2020

[2]. Melville, P.; Sindhwani, V. Recommender systems. In Encyclopedia of Machine Learning and Data Mining; Springer: Berlin, Germany, 2017; pp. 1056–1066.

[3]. Michael J. Pazzani. A framework for collaborative, content-based and demographic filtering. Artificial Intelligence Review, 13(5-6):393–408, 1999.

[4]. Prem Melville, Raymond J. Mooney, and Ramadass Nagarajan. Contentboosted collaborative filtering for improved recommendations. In Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI02), pages 187–192, Edmonton, Alberta, 2002.

[5]. Marko Balabanovic and Yoav Shoham. Fab: Content-based, collaborative recommendation. Communications of the Association for Computing Machinery, 40(3):66–72, 1997.

[6]. P. Cotter and B. Smyth. PTV: Intelligent personalized TV guides. In Twelfth Conference on Innovative Applications of Artificial Intelligence, pages 957– 964, 2000.

[7]. Alexandrin Popescul, Lyle Ungar, David M. Pennock, and Steve Lawrence. Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments. In Proceedings of the Seventeenth Conference on Uncertainity in Artificial Intelligence, 2001.

[8]. Thomas Hofmann. Probabilistic latent semantic analysis. In Proceedings of Uncertainty in Artificial Intelligence (UAI), 1999.

[9]. John S. Breese, David Heckerman, and Carl Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, Madison, WI, July 1998.

[10]. Daniel Billsus and Michael J. Pazzani. Learning collaborative information filters. In Proceedings of the Fifteenth International Conference on Machine Learning (ICML-98), pages 46–54, Madison, WI, 1998. Morgan Kaufmann.

[11]. Robert Bell Yehuda Koren and Chris Volinsky. Matrix factorization techniques for recommender systems. In IEEE Computer, volume 42 (8), pages 30–37, 2009.

[12]. Sun-Chong Wang. Interdisciplinary Computing in Java Programming. Springer Science & Business Media, 2012.

[13]. Weiyu Cheng, Yanyan Shen, Yanmin Zhu, and Linpeng Huang. 2018. DELF: A Dual-Embedding based Deep Latent Factor Model for Recommendation. In IJCAI. 3329–3335.

[14]. H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, et al. Communication efficient learning of deep networks from decentralized data. arXiv preprint arXiv:1602.05629, 2016.

[15]. H. Brendan McMahan, Eider Moore, Daniel Ramage, and Blaise Aguera y Arcas, "Federated learning of deep networks using model averaging," CoRR, vol. abs/1602.05629, 2016.

[16]. D. Leroy, A. Coucke, T. Lavril, T. Gisselbrecht, and J. Dureau, "Federated learning for keyword spotting," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[17]. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[18]. Y. Deng, M. M. Kamani, and M. Mahdavi, "Adaptive Personalized Federated Learning," arXiv:2003.13461 [cs, stat], Mar. 2020. [Online]. Available: http://arxiv.org/abs/2003.13461

[19]. Sanner MF (1999) Python: a programming language for software integration and development. Journal ofMolecular Graphics& Modelling 17, 57–61.

[20]. S. Bock, M. Weiß, and J. Goppold, An improvement of the convergence proof of the ADAM-Optimizer. Clusterkonferenz 2018, 2018. [Online]. Available: http : / / arxiv.org/abs/1804.10587.

[21]. D. P. Kingma and J. L. Ba, Adam: A Method for stochastic Optimization. San Diego: The International Conference on Learning Representations (ICLR), 2015.

[22]. Osama Abu Abbas, " Comparison between clustering Algorithms", The International Arab Journal of Information Technology, vol.5, No.3, July 2008.