Πολυτεχνική Σχολή

Τμήμα Μηχανικών Η/Υ & Πληροφορικής

ΑΛΓΟΡΙΘΜΟΙ ΕΠΙΣΤΗΜΗΣ ΔΕΔΟΜΕΝΩΝ

ΥΔΑ

Collaborative Filtering: Matrix Factorization vs. Neural Networks

---

Πετράκης Κωνσταντίνος

Α.Μ. 1041589

2021

1. Εισαγωγή

Τα συστήματα συστάσεων (recommendation systems) χρησιμοποιούνται για την πρόβλεψη της απόκρισης ενός χρήστη απέναντι σε μια πληθώρα επιλογών. Παραδείγματα εφαρμογών όπου έχουν ευρεία χρήση τα συστήματα συστάσεων αποτελούν η σύσταση ταινιών που πιθανόν θα άρεσαν στους χρήστες κάποιας streaming υπηρεσίας ταινιών (π.χ. Netflix), η σύσταση στους πελάτες ενός ηλεκτρονικού καταστήματος (π.χ. Amazon), προϊόντων τα οποία είναι πολύ πιθανό να ενδιαφέρονται να αγοράσουν με βάση το ιστορικό αγορών τους ή τις πρόσφατες αναζητήσεις του, η σύσταση βίντεο (π.χ. Youtube) ή ακόμη η πρόταση άρθρων στους διαδικτυακούς αναγνώστες κάποιας ηλεκτρονικής εφημερίδας, των οποίων η θεματολογία συμπίπτει με τα ενδιαφέροντα του εκάστοτε αναγνώστη. Είναι επομένως αυτονόητη η διεισδυτικότητα των συστημάτων συστάσεων σε πολλές εκφάνσεις της σημερινής πραγματικότητας. Παλαιότερα ίσως η σημαντικότερη πρόκληση για την υλοποίηση τέτοιων συστημάτων ήταν η συλλογή επαρκών δεδομένων. Σήμερα όμως όπου κατακλυζόμαστε από δεδομένα και λίγο έως πολλοί όλοι μας είμαστε χρήστες κάποιας υπηρεσίας η οποία κάνει χρήση των δεδομένων μας για την παροχή κατάλληλα προσωποποιημένων λειτουργιών, τη μεγαλύτερη πρόκληση αποτελεί η επιλογή της καλύτερης τεχνικής. Η επιλογή λοιπόν των αποτελεσματικότερων αλλά και καταλληλότερων, για το κάθε σενάριο εφαρμογής, μεθόδων αποτελεί υψίστης σημασίας ζήτημα στα συστήματα συστάσεων.

Με βάση τις τεχνικές που χρησιμοποιούν τα συστήματα συστάσεων κατηγοριοποιούνται σε 2 κατηγορίες, σε αυτά που βασίζονται στο περιεχόμενο (content-based) στα οποία οι συστάσεις γίνονται με βάση κάποιες ιδιότητες των αντικειμένων και στα συστήματα συνεργατικού φιλτραρίσματος (Collaborative Filtering) στα οποία και θα επικεντρωθούμε. Ο όρος συνεργατικό φιλτράρισμα αναφέρεται στη σύσταση αντικειμένων με βάση ένα μέτρο ομοιότητας μεταξύ των χρηστών ή των αντικειμένων. Εμείς θα εστιάσουμε σε προσεγγίσεις που αφορούν την ομοιότητα μεταξύ χρηστών. Το συνεργατικό φιλτράρισμα λοιπόν έγκειται στην εύρεση ομάδων όμοιων χρηστών (με παρόμοια γούστα) και στην σύσταση αντικειμένων σε κάποιο χρήστη με βάση τα γούστα των όμοιων με αυτόν χρηστών.

Η παρούσα εργασία αφορά στην σύγκριση μεταξύ δύο εκ των πιο διαδεδομένων τεχνικών για την αποδοτική υλοποίηση του συνεργατικού φιλτραρίσματος, της Παραγοντοποίηση Μητρώου και των Τεχνητών Νευρωνικών Δικτύων (ΤΝΔ). Στην ενότητα 2 παρουσιάζεται η βασική ορολογία που θα χρησιμοποιηθεί, διατυπώνεται πιο αυστηρά το προς επίλυση πρόβλημα και σκιαγραφούνται οι δυο τεχνικές προσέγγισης του προβλήματος που θα συγκριθούν. Στην ενότητα 3 αναλύονται οι επιλεγμένες εργασίες που αφορούν. Αρχικά στην ενότητα 3.1 παρουσιάζεται η πρώτη εργασία που εισήγαγε την χρήση πολυεπίπεδων ΤΝΔ εμπρόσθιας τροφοδότησης (MLP) σαν προσέγγιση για το συνεργατικό φιλτράρισμα. Στην ενότητα 3.2 οι συγγραφείς της εργασίας επανεξέτασαν τα αποτελέσματα της πρώτης και δείχνουν πως η χρήση Παραγοντοποίησης Μητρώου έχει καλύτερα αποτελέσματα από τα MLP. Στην εργασία της ενότητας 3.3 επιχειρείται μια εμπειρική εξήγηση για τους λόγους που οι τεχνικές παραγοντοποίησης μητρώου φαίνεται να είναι καταλληλότερες για το πρόβλημα του συνεργατικού φιλτραρίσματος.

2. *Βασικές Έννοιες και Πρόβλημα Μελέτης*

Αρχικά ας διατυπώσουμε το πρόβλημα που αντιμετωπίζουν τα συστήματα συστάσεων πιο φορμαλιστικά. Αυτό που έχουμε είναι ένα σύνολο χρηστών και ένα σύνολο αντικειμένων. Για να είμαστε σε θέση να προτείνουμε κάποια από τα αντικείμενα στους χρήστες πρέπει να 'μάθουμε' μια συνάρτηση η οποία θα αποτελεί ένα μέτρο αλληλεπίδρασης μεταξύ χρηστών-αντικειμένων, θα προβλέπει δηλαδή κατά πόσο ένα αντικείμενο είναι κατάλληλο για σύσταση σε κάποιο χρήστη. Οι προτιμήσεις των χρηστών στα διάφορα αντικείμενα αποθηκεύονται σαν τα στοιχεία ενός μητρώου, κάθε γραμμή του οποίου αντιστοιχεί σε έναν χρήστη και κάθε στήλη σε ένα ξεχωριστό αντικείμενο. Ένα κλασικό παράδειγμα είναι αυτό της σύστασης ταινιών, των οποίων η αξιολόγηση γίνεται συνήθως με έναν αριθμό από 1 έως 5, και το στοιχείο (i,j) του μητρώου περιέχει την βαθμολογία που έδωσε ο χρήστης i στην ταινία j. Θα αναφερόμαστε σε αυτά τα μητρώα σαν μητρώα αλληλεπίδρασης (Utility matrix). Τα μητρώα αλληλεπίδρασης είναι συνήθως αρκετά αραιά, λόγου χάρη στο παραπάνω παράδειγμα οι χρήστες έχουν δει μονό κάποιες από τις ταινίες, και το πρόβλημα έγκειται στο να εκτιμηθούν όσο γίνεται ακριβέστερα τα ελλιπή στοιχεία του μητρώου. Αυτές οι τιμές αποτελούν την αναμενόμενη αρέσκεια του κάθε χρήστη στο εκάστοτε αντικείμενο, στο παράδειγμα των ταινιών θα αποτελούν την βαθμολογία που εκτιμούμε ότι θα έδινε ένας χρήστης σε κάποια ταινία.

Για την υλοποίηση του συνεργατικού φιλτραρίσματος υπάρχουν διάφορες τεχνικές. Μια από τις πιο δημοφιλής είναι η τεχνική της Παραγοντοποίησης Μητρώου (Matrix Factorization) η οποία ανήκει σε μια γενικότερη οικογένεια τεχνικών γνωστές σαν τεχνικές latent (λανθάνων) παραγόντων, οι οποίες όπως προδίδει και το όνομα τους, βασίζονται σε latent παράγοντες και έγιναν αντικείμενο εκτενέστερης μελέτης ιδιαίτερα μετά από την χρήση τους στο διαγωνισμό Netflix prize [12]. Αν έχουμε M χρήστες και N αντικείμενα, δηλαδή το μητρώο αλληλεπίδρασης είναι διαστάσεων MxN, τότε η παραγοντοποίηση μητρώου έγκειται στην εύρεση δυο παραγόντων, δύο μικρότερης διάστασης μητρώων δηλαδή, Mxd και Nxd το γινόμενο των οποίων θα προσεγγίζει όσο το δυνατόν καλύτερα το αρχικό μητρώο αλληλεπίδρασης. Η διάσταση d<min(M,N) αποτελεί υπέρ-παράμετρο, δηλαδή καθορίζεται από εμάς σαν μέρος του μοντέλου, και αντιστοιχεί στην διάσταση του latent χώρου (στο πλήθος των latent παραγόντων). Ο παράγοντας διάσταση Mxd αντιστοιχεί στο μητρώο των χρηστών με τις γραμμές του να αντιστοιχούν στους χρήστες και τις στήλες στους latent παράγοντες. Ο δεύτερος παράγοντας διάστασης Nxd αντιστοιχεί στο μητρώο των αντικειμένων όπου οι γραμμές αντιστοιχούν στα αντικείμενα και οι στήλες στους latent παράγοντες, και προφανώς χρησιμοποιείται ο ανάστροφος του για γινόμενο των δυο μητρώων.

Οι latent παράγοντες αντιστοιχούν σε χαρακτηριστικά του προβλήματος τα οποία ανακαλύπτονται κατά την παραγοντοποίηση. Χρησιμοποιώντας και πάλι το παράδειγμα της σύστασης ταινιών, τα χαρακτηριστικά που ανακαλύπτονται θα μπορούσαν να είναι η ταινία να ανήκει σε κάποιο συγκεκριμένο είδος (π.χ. δράσης), η πλοκή της, ή ακόμη το γεγονός κάποιος συγκεκριμένος ηθοποιός (π.χ. Anthony Hopkins) να παίζει στην ταινία. Το μητρώο χρηστών τότε θα περιείχε σε κάθε γραμμή, δηλ. για κάθε χρήστη, την βαθμολογία που θα έδινε ο χρήστης αν η ταινία ήταν Δράσης, αν εμφανιζόταν σε αυτήν ο Anthony Hopkins και ούτω καθ' εξής. Το μητρώο αντικειμένων θα περιείχε για κάθε στήλη, δηλ. για κάθε αντικείμενο, μια τιμή που θα έδειχνε κατά πόσο μπορεί να θεωρηθεί η συγκεκριμένη ταινία σαν ταινία δράσης, μια δυαδική τιμή 1 ή 0 που θα έδειχνε αν εμφανίζεται σε αυτήν ο Anthony Hopkins κλπ. Μετά λοιπόν από την παραγοντοποίηση του μητρώου αλληλεπίδρασης, η εκτίμηση της βαθμολογίας που θα έδινε κάποιος χρήστη u_i σε κάποια ταινία v_j δίνεται από το εσωτερικό γινόμενο των latent διανυσμάτων της γραμμής i του πίνακα χρηστών και της στήλης j του ανάστροφου πίνακα αντικειμένων. Αξίζει να σημειωθεί πως στις περισσότερες περιπτώσεις η διαισθητική ερμηνεία των χαρακτηριστικών είναι αδύνατη. Η παραπάνω ανάλυση με το παράδειγμα της σύστασης ταινιών έγινε καθαρά για επεξηγηματικούς λόγους. Το γεγονός βέβαια πως δεν μπορούμε να αντιληφθούμε την διαισθητική ερμηνεία των χαρακτηριστικών δεν αποτελεί σε καμία περίπτωση μειονέκτημα της μεθόδου. Κάθε άλλο μάλιστα, αφού φαίνεται πως η παραγοντοποίηση μητρώου ανιχνεύει αυτόματα τα χαρακτηριστικά εκείνα που περιέχουν την περισσότερη πληροφορία και είναι χρησιμότερα για την όσο το δυνατόν ακριβέστερη εκτίμηση των ελλιπών στοιχείων του μητρώου αλληλεπίδρασης.

Η παραγοντοποίηση μητρώου δηλαδή αντιστοιχίζει το αρχικό πρόβλημα σε ένα χαμηλότερης διάστασης latent χώρο ή χώρο χαρακτηριστικών (feature space). Σε αυτόν τον κοινό latent χώρο αντιστοιχίζονται οι χρήστες και τα αντικείμενα και κάθε χρήστης και αντικείμενο αναπαρίσταται από ένα latent διάνυσμα, ενώ μέτρο αλληλεπίδρασης μεταξύ ενός χρήστη και κάποιου αντικειμένου αποτελεί το εσωτερικό γινόμενο των αναπαραστάσεων τους, δηλαδή των latent διανυσμάτων τους. Με αυτόν τον τρόπο η εφαρμογή της παραγοντοποίησης μητρώου έχει σαν αποτέλεσμα σε έναν χρήστη να προτείνονται τα αντικείμενα εκείνα των οποίων η αναπαράσταση στον latent χώρο είναι εγγύτερα στην αναπαράσταση του χρήστη στον ίδιο latent χώρο.

Το ερώτημα που εγείρεται φυσικά είναι πως θα βρεθεί η παραγοντοποίηση για ένα μητρώο αλληλεπίδρασης, δηλαδή πως θα ανακτηθούν τα latent διανύσματα χρηστών και αντικειμένων. Η εύρεση των παραγόντων (μητρώα χρηστών και αντικειμένων) που προσεγγίζουν όσο το δυνατόν καλύτερα το μητρώο αλληλεπίδρασης γίνεται ορίζοντας σαν συνάρτηση σφάλματος τη ρίζα των μέσων τετραγωνικών σφαλμάτων (RMSE) μεταξύ των υπάρχοντων στοιχείων του μητρώου αλληλεπίδρασης και της προσέγγισης τους από το εσωτερικό γινόμενο των αντίστοιχων διανυσμάτων των μητρώων χρηστών και αντικειμένων και βελτιστοποιώντας αυτή την συνάρτηση σφάλματος συνήθως με την χρήση του αλγορίθμου Στοχαστικής Κατάβασης Κλίσης (SGD) ή κάποιας παραλλαγής του που να περιλαμβάνει και ομαλοποίηση (regularization).

Τα τελευταία χρόνια με την εκτενή χρήση ΤΝΔ σε διάφορους τομείς και τα εξαιρετικά αποτελέσματα που επιφέρουν αναπόφευκτα δοκιμάστηκε η χρήση τους και στο συνεργατικό φιλτράρισμα. Πιο συγκεκριμένα η χρήση των ΤΝΔ επιστρατεύτηκε με το επιχείρημα πως το εσωτερικό γινόμενο μεταξύ των latent διανυσμάτων απλά αποτελεί τον γραμμικό συνδυασμό των latent χαρακτηριστικών, μια πράξη η οποία η οποία ίσως να μην είναι αρκετή για να συλλάβει την σύνθετη αλληλεπίδραση μεταξύ χρηστών και αντικειμένων. Για την εύρεση μιας πιο περίπλοκης συνάρτησης ομοιότητας η οποία θα ανακαλύπτει αυτή την σύνθετη αλληλεπίδραση προτάθηκαν τα ΤΝΔ. Υπάρχουν πολλές διαφορετικές αρχιτεκτονικές ΤΝΔ και αρκετές από αυτές έχουν δοκιμαστεί στο συνεργατικό φιλτράρισμα. Εδώ όμως θα εστιάσουμε στη χρήση των απλών πολύ-επίπεδων ΤΝΔ εμπρόσθιας τροφοδότησης (Multi-Layer Perceptrons ή MLP) τα οποία χρησιμοποιήθηκαν για πρώτη φορά στο πεδίο του συνεργατικού φιλτραρίσματος στην εργασία [2]. (χάριν συντομίας από εδώ και στο εξής όπου αναφέρεται ο όρος MLP εννοείται πολύ-επίπεδο ΤΝΔ εμπρόσθιας τροφοδότησης)

Είναι σκόπιμο σε αυτό το σημείο να ορίσουμε και ένα βασικό διαχωρισμό που υπάρχει μεταξύ των προβλημάτων συνεργατικού φιλτραρίσματος. Σε περιπτώσεις στις οποίες τα δεδομένα που έχουμε στην διάθεση μας αποτελούνται από αριθμητικές τιμές (π.χ. αξιολογήσεις ταινιών) για τα αντικείμενα με τα οποία είχαν αλληλεπίδραση οι χρήστες κάνουμε λόγο για explicit feedback σύνολα δεδομένων. Σε αυτές τις περιπτώσεις το μητρώο αλληλεπίδρασης είναι αραιό και περιέχει μόνο τις αριθμητικές τιμές αλληλεπίδρασης. Σε περιπτώσεις όπου έχουμε μόνο πληροφορία π.χ. για το ιστορικό κάθε χρήστη ή για τα clicks που έκανε σε κάποια ιστοσελίδα, τότε το μητρώο αλληλεπίδρασης αποτελείται από τιμές 0 και 1, όπου η τιμή 1 δηλώνει αλληλεπίδραση του χρήστη με κάποιο αντικείμενο (π.χ. κάποιο click) και η τιμή 0 την απουσία αλληλεπίδρασης. Σε αυτές τις περιπτώσεις κάνουμε λόγο για implicit feedback σύνολα δεδομένων Αξίζει να σημειωθεί πως σε implicit feedback δεδομένα το πρόβλημα του συνεργατικού φιλτραρίσματος γίνεται ακόμη πιο δύσκολο καθώς ενώ η τιμή 1 δηλώνει το ενδιαφέρον του χρήστη σε κάποιο αντικείμενο, η τιμή 0 μπορεί να σημαίνει ελλιπή δεδομένα ή το μη ενδιαφέρον του χρήστη στο αντικείμενο είτε γιατί δεν γνώριζε την ύπαρξη του είτε διότι πραγματικά δεν ενδιαφέρεται για αυτό. Σε τέτοια σύνολα δεδομένων επομένως παρέχεται ελάχιστη πληροφορία για το τι δεν αρέσει στο χρήστη (negative feedback). Οι εργασίες που αναλύονται στην ενότητα 3 εστιάζουν σε implicit feedback σύνολα δεδομένων.

Πληθώρα εργασιών και συγκριτικών μελετών τα τελευταία χρόνια επιχειρούν να αποσαφηνίσουν αν η χρήση MLP μπορεί να επιφέρει σημαντικά καλύτερα αποτελέσματα στο συνεργατικό φιλτράρισμα από ότι τεχνικές που βασίζονται στην Παραγοντοποίηση Μητρώου και στο εσωτερικό γινόμενο μεταξύ των latent διανυσμάτων, καθώς και τους λόγους για τους οποίους αυτό μπορεί να μην είναι εφικτό. Στην συνέχεια παρουσιάζονται 3 εργασίες που αφορούν στην σύγκριση μεταξύ αυτών των 2 τεχνικών.

3. *Εργασίες*

3.1. Neural Collaborative Filtering

Η παρούσα εργασία [2] ήταν η πρώτη στην οποία προτάθηκε η χρήση MLP σαν προσέγγιση στο πρόβλημα του συνεργατικού φιλτραρίσματος σε σύνολα δεδομένων implicit feedback. Σε γενικές γραμμές αυτό που προτάθηκε από τους συγγραφείς ήταν αρχικά η προβολή χρηστών και αντικειμένων σε έναν κοινό latent χώρο χαρακτηριστικών χαμηλότερης διάστασης, με τη χρήση των embedding επιπέδων, και στη συνέχεια η χρήση MLP για την μοντελοποίηση της αλληλεπίδρασης μεταξύ αυτών των latent χαρακτηριστικών χρηστών και αντικειμένων με την ελπίδα πως η μη γραμμική αντιστοίχιση που προσφέρουν τα MLP θα επέτρεπε την μάθηση πιο πολύπλοκων, ευέλικτων και αποτελεσματικότερων συναρτήσεων αλληλεπίδρασης μεταξύ χρηστών και αντικειμένων από ότι το εσωτερικό γινόμενο, με απώτερο στόχο πάντα την ακριβέστερη σύσταση αντικειμένων στους χρήστες. Σημειώνεται πως καθώς, σύμφωνα με το Θεώρημα καθολικής προσέγγισης, τα MLP μπορούν να προσεγγίσουν οποιαδήποτε συνεχή συνάρτηση, αναμένεται να είναι σε θέση να προσεγγίσουν το εσωτερικό γινόμενο.

Αρχικά γίνεται αναφορά στις αδυναμίες που μπορεί να έχει η χρήση του εσωτερικού γινομένου μεταξύ των latent διανυσμάτων χρηστών-αντικειμένων, όπως το γεγονός πως το εσωτερικό γινόμενο αποτελεί τον γραμμικό συνδυασμό των διαστάσεων του latent χώρου υποθέτοντας πως αυτές είναι ανεξάρτητες μεταξύ τους η ότι η επιλογή μεγαλύτερης διάστασης για τον latent χώρο μπορεί να οδηγήσει σε υπέρ-προσαρμογή (overfitting). Στη συνέχεια παρουσιάζουν τα γενικό πλαίσιο της εργασίας τους το οποίο αντιστοιχεί σε μια πολύ-επίπεδη αναπαράσταση για την μοντελοποίηση της αλληλεπίδρασης χρηστών-αντικειμένων. Είσοδο σε αυτήν αποτελούν ζεύγη διανυσμάτων one-hot αναπαραστάσεων χρηστών και αντικειμένων τα οποία μετασχηματίζονται στα αντίστοιχα embedding διανύσματα χρηστών και αντικειμένων με τη χρήση του embedding επιπέδου. Αυτά τα embedding διανύσματα έχουν τον ρόλο των latent διανυσμάτων και είναι αυτά τα οποία τροφοδοτούνται στα κρυφά επίπεδα του MLP, το επίπεδο εξόδου του οποίου αποτελείται από έναν νευρώνα η τιμή του οποίου αντιστοιχεί στη εκτιμώμενη αλληλεπίδραση η μη του χρήστη με το αντικείμενο.

Παρατηρώντας πως οι τιμές που απαιτείται να προβλεφθούν σε implicit feedback δεδομένα είναι 0 ή 1 οι συγγραφείς αντιμετώπισαν το πρόβλημα της εκπαίδευσης του MLP σαν ένα πρόβλημα δυαδική ταξινόμησης, όπου η τιμή εξόδου του τελικού νευρώνα αντιπροσωπεύει την πιθανότητα το αντικείμενο που δόθηκε στην εισόδου να είναι σχετικό με τον χρήστη εισόδου. Αυτό τους επέτρεψε να εκπαιδεύσουν το MLP χρησιμοποιώντας σαν συνάρτηση σφάλματος την δυαδική εντροπία (binary crossentropy) και σαν αλγόριθμο βελτιστοποίησης τη Στοχαστική Κατάβαση Κλίσης (SGD), ορίζοντας φυσικά την σιγμοειδή σαν συνάρτηση ενεργοποίησης του νευρώνα εξόδου ώστε να περιοριστεί η έξοδος του στο [0,1].

Εν συνεχεία οι συγγραφείς επιχειρηματολογούν πως το εσωτερικό γινόμενο μεταξύ των latent διανυσμάτων στο μοντέλο παραγοντοποίησης μητρώου αποτελεί ειδική περίπτωση του γενικότερου πλαισίου που έχουν ορίσει, το οποίο εκμεταλλεύεται τα MLP. Αυτό γιατί εάν χρησιμοποιηθεί ένα μόνο κρυφό επίπεδο μετά το επίπεδο embedding το οποίο να εκτελεί τον στοιχείο προς στοιχείο πολλαπλασιασμό των embedding διανυσμάτων, και εάν τα συναπτικά βάρη μεταξύ αυτού του κρυφού επιπέδου και του νευρώνα εξόδου γίνουν όλα ίσα με 1 και σαν συνάρτηση ενεργοποίησης στο νευρώνα εξόδου χρησιμοποιηθεί η γραμμική τότε το μοντέλο τους θα έχει ακριβώς τα ίδια αποτελέσματα με το μοντέλο παραγοντοποίησης μητρώου. Ονομάζουν αυτήν την παραλλαγή ‘Γενικευμένη Παραγοντοποίηση Μητρώου’ (GMF). Ακόμη δοκιμάζουν την απλή χρήση ενός MLP, στο επίπεδο εισόδου του οποίου τροφοδοτούν την συνένωση των διανυσμάτων χρήστη και αντικειμένου, ενώ το πλήθος των νευρώνων των κρυφών επιπέδων μειώνεται για κάθε διαδοχικό επίπεδο. Σαν συνάρτηση ενεργοποίησης στους νευρώνες των κρυφών επιπέδων επιλέχθηκε η συνάρτηση ράμπας (ReLu).

Επιπλέον προκειμένου να εκμεταλλευτούν από την μία την γραμμικότητα της γενικευμένης παραγοντοποίησης μητρώου και από την άλλη τη μη-γραμμικότητα των MLP, ορίζουν την συγχώνευση αυτών των δύο μοντέλων υπό το όνομα ‘Νευρωνική Παραγοντοποίηση Μητρώου’ (NeuMF) με την ελπίδα πως θα επωφεληθούν των πλεονεκτημάτων και των δύο, για την ακόμη καλύτερη μοντελοποίηση της αλληλεπίδρασης

χρηστών-αντικειμένων. Η συγχώνευση εκτελείται επιτρέποντας στα δύο μοντέλα να εξάγουν διαφορετικά embeddings χρηστών και αντικειμένων το κάθε ένα και συνενώνοντας τα τελευταία τους κρυφά επίπεδα σε ένα κοινό επίπεδο πριν τον κόμβο εξόδου. Το NeuMF δοκιμάζεται εκπαιδεύοντας εξ' αρχής και τα δύο μοντέλα από κοινού, αλλά και προ-εκπαιδεύοντας αρχικά το κάθε ένα από τα δύο μοντέλα ξεχωριστά και στην συνέχεια εκτελώντας fine-tuning στο συγχωνευμένο μοντέλο, εκκινώντας από τα βάρη στα οποία σταμάτησε η εκπαίδευση κάθε μοντέλου.

Οι συγγραφείς εξετάζουν πειραματικά τα 3 μοντέλα που όρισαν (GMF,MLP,NeuMF) σε δύο σύνολα δεδομένων, το Pinterest, το οποίο αντιστοιχεί σε implicit feedback σύνολο δεδομένων και το MovieLens, αφού πρώτα έθεσαν κάθε δοσμένη βαθμολογία ταινίας ίση με τη τιμή 1 και όλες τις ελλιπείς βαθμολογίες ίσες με 0 μετατρέποντας το έτσι σε implicit feedback. Το σύνολο ελέγχου (test set) αποτελείται από το τελευταίο αντικείμενο με το οποίο έχει αλληλοεπιδράσει κάθε χρήστης και όλες οι υπόλοιπες αλληλεπιδράσεις χρησιμοποιήθηκαν σαν δείγματα εκπαίδευσης. Όσον αφορά, τα απαραίτητα για την διαδικασία της εκπαίδευσης δείγματα αρνητικής ανάδρασης, για κάθε δείγμα αλληλεπίδρασης χρήστη-αντικειμένου δειγματοληπτούνται τυχαία ομοιόμορφα 4 δείγματα από τις μη παρατηρημένες αλληλεπιδράσεις σε κάθε κύκλο εκπαίδευσης. Για την αξιολόγηση δειγματοληπτούνται τυχαία 100 αντικείμενα για κάθε χρήστη, με τα οποία δεν έχουν αλληλοεπιδράσει και ελέγχεται, για κάθε χρήστη, η θέση του αντικειμένου του συνόλου ελέγχου στη ταξινομημένη λίστα των 101 αντικειμένων που αντιστοιχεί στον κάθε έναν με τις μετρικές Hit Ratio (HR) και Normalized Discounted Cumulative Gain (NDCG). Τα τελικά αποτελέσματα προκύπτουν από τον μέσο όρο των δυο μετρικών στο σύνολο των χρηστών. Το HR είναι ίσο με 1 αν το αντικείμενο του συνόλου ελέγχου είναι στα ανάμεσα στα πρώτα 10 αντικείμενα και 0 αλλιώς ενώ το NDCG είναι ίσο με $1/\log(r+1)$, όπου r θέση του αντικειμένου του συνόλου ελέγχου στην τελική ταξινομημένη λίστα.

Συγκρίνοντας τα αποτελέσματα των τριών προτεινόμενων μοντέλων με τεχνικές τεχνολογικής στάθμης που αντιμετωπίζουν το ίδιο πρόβλημα παρατηρήθηκε πως το NeuMF έχει σημαντικά καλύτερη απόδοση όσον αφορά και τις δύο μετρικές και για τα δύο επιλεγμένα σύνολα δεδομένων, ενώ τα MLP και το GMF έχουν επίσης καλή απόδοση, με το GMF να φαίνεται γενικά πιο αποτελεσματικό από τα απλά MLP, με την απόδοση των τελευταίων όμως να ευνοείται από την προσθήκη επιπλέον κρυφών επιπέδων και την απόδοση του GMF να μειώνεται, λόγω υπέρ-προσαρμογής, όσο αυξάνονται οι νευρώνες του κρυφού του επίπεδου.

Ακόμη όσον αφορά την προ-εκπαίδευση στο NeuMF, εκτός της περίπτωσης στο MovieLens όπου στο τελικό κρυφό επίπεδο χρησιμοποιούνται 8 νευρώνες, η χρήση της έχει ευεργετικά αποτελέσματα στην απόδοση του μοντέλου, κατά μέσο όρο 2.2% για το MovieLens και 1.1% για το Pinterest. Επιπλέον για να επιβεβαιώσουν την καταλληλότητα επιλογής της δυαδικής εντροπίας ως συνάρτηση σφάλματος δείχνουν πως αυτή μειώνεται, ειδικά κατά τους 10 πρώτους κύκλους εκπαίδευσης, και για τα τρία μοντέλα, με το NeuMF να έχει καλύτερη απόδοση από το MLP το οποίο με τη σειρά του έχει καλύτερη απόδοση από το GMF, όσο ο αριθμός των κύκλων εκπαίδευσης αυξάνεται. Επίσης η καταλληλότητα της δυαδικής εντροπίας επιβεβαιώνεται και από το γεγονός πως επιτρέπει τον καθορισμό του πλήθος των δειγμάτων αρνητικής ανάδρασης που θα χρησιμοποιηθούν για κάθε δείγμα αλληλεπίδρασης χρήστη-αντικειμένου σε σχέση με άλλες συναρτήσεις σφάλματος που επιτρέπουν μόνο αντιστοιχία ένα προς ένα. Τέλος επισημαίνεται πως η αύξηση του πλήθους των μη-γραμμικών κρυφών επιπέδων στα MLP επιφέρει καλύτερα αποτελέσματα και ως προς τις δύο μετρικές αξιολόγησης. Οι ερευνητές με βάση τα εμπειρικά αποτελέσματα που έλαβαν καταλήγουν στο συμπέρασμα πως η χρήση των MLP επιφέρει καλύτερα αποτελέσματα από ότι τεχνική της παραγοντοποίησης μητρώου στο συνεργατικό φιλτράρισμα.

3.2. Neural Collaborative Filtering vs. Matrix Factorization Revisited

Σε αυτήν την εργασία [1] επανεξετάζονται τα αποτελέσματα της προηγούμενης εργασίας [2] ως προς την υπεροχή των MLP έναντι μεθόδων Παραγοντοποίησης Μητρώου στο συνεργατικό φιλτράρισμα. Οι συγγραφείς αρχικά επιβεβαιώνουν πειραματικά πως η χρήση του εσωτερικού γινομένου, με την κατάλληλη επιλογή υπέρ-παραμέτρων, μεταξύ των embedding διανυσμάτων οδηγεί σε καλύτερα αποτελέσματα από ότι συναρτήσεις αλληλεπίδρασης που μαθαίνονται με τη χρήση MLP. Επίσης παρέχουν μια εμπειρική απόδειξη πως είναι δύσκολο για ένα MLP να 'μάθει' το εσωτερικό γινόμενο ενώ αναφέρονται και σε ζητήματα υλοποίησης και πρακτικής εφαρμογής των μεθόδων παραγοντοποίησης μητρώου και των MLP.

Οι συγγραφείς χρησιμοποιούν μια παραλλαγή του εσωτερικού γινομένου μεταξύ των embedding διανυσμάτων με την προσθήκη explicit biases. Εκπαιδεύουν το μοντέλο παραγοντοποίησης μητρώου με τη χρήση της συνάρτησης δυαδικής εντροπίας σαν συνάρτηση σφάλματος με L2 ομαλοποίηση, βελτιστοποιώντας την με τον αλγόριθμο SGD και χρησιμοποιώντας αρνητική δειγματοληψία m αντικειμένων με τον ίδιο τρόπο που χρησιμοποιήθηκε και στην αρχική εργασία [2]. Είναι αυτονόητο πως τα πειράματα εκτελούνται στα ίδια σύνολα δεδομένων που χρησιμοποιήθηκαν και στην αρχική εργασία. Για την σύγκριση μεταξύ των δυο μεθόδων συναρτήσει της διάστασης των embedding διανυσμάτων, μετατράπηκε το πλήθος των νευρώνων του τελευταίου κρυφού επιπέδου των MLP (όρος 'predictive factor') συναρτήσει του οποίου λαμβάνονταν τα αποτελέσματα στην εργασία [2] στην αντίστοιχη διάσταση των embedding διανυσμάτων. Τελικά δοκιμάζονται embedding διαστάσεων {16, 32, 64, 96, 128, 192}.

Η αξιολόγηση των πειραμάτων γίνεται ακριβώς με τον ίδιο τρόπο που αξιολογήθηκαν και τα αποτελέσματα της πρώτης εργασίας [2]. Τα αποτελέσματα εδώ δείχνουν πως η χρήση εσωτερικού γινομένου μεταξύ των embeddings χρηστών-αντικειμένων, για την μοντελοποίηση της μεταξύ τους αλληλεπίδρασης έχει καλύτερη απόδοση από την χρήση των μοντέλων MLP και NeuMF και ως προς τις 2 μετρικές αξιολόγησης και για όλες τις επιλεγμένες διαστάσεις embeddings πλην μίας (για διάσταση 192 το προ-εκπαιδευμένο NeuMF έχει ελαφρώς καλύτερο HR στο MovieLens). Επίσης καταρρίπτεται ο ισχυρισμός πως η τροφοδότηση ενός μέρους των embedding διαμέσου ενός MLP θα μπορούσε να βελτιώσει τα αποτελέσματα του εσωτερικού γινομένου, όπως είχε υποστηριχθεί στην πρώτη εργασία με την χρήση του NeuMF. Όσον αφορά το γεγονός πως η χρήση του GMF, το οποίο αποτελεί γενίκευση του εσωτερικού γινομένου, επιφέρει αρκετά κατώτερα αποτελέσματα σε σύγκριση με το εσωτερικό γινόμενο ενώ θεωρητικά θα περίμενε κανείς να έχει καλύτερη απόδοση, οφείλεται σύμφωνα με τους συγγραφείς από τη μία στην εγγενή δυσκολία εκπαίδευσης τέτοιων μεθόδων, και κυρίως στην προσθήκη νέων παραμέτρων στο κατά τα άλλα απλό μοντέλο του εσωτερικού γινομένου. Για παράδειγμα η χρήση ομαλοποίησης δεν έχει καμία επίδραση εάν δεν εκτελεστεί ομαλοποίηση και στο διάνυσμα w των νέων παραμέτρων που εισάγει το GMF.

Ακόμη επισημαίνεται, και έχει σχολιαστεί και σε άλλες εργασίες, π.χ. [6], πως τα αποτελέσματα που αναφέρονται στην [2] είναι εκείνα που αντιστοιχούν στο καλύτερο τρέξιμο των μοντέλων στο σύνολο ελέγχου, υπερεκτιμώντας πιθανόν την γενικευτική ικανότητα των μοντέλων MLP και NeuMF. Το γεγονός αυτό σε συνδυασμό με το γεγονός πως η επιλογή όλων των υπέρ-παραμέτρων στην παρούσα εργασία έγινε σε ένα ξεχωριστό σύνολο επικύρωσης (και όχι στο σύνολο ελέγχου όπως στην πρώτη) και τα αποτελέσματα που λαμβάνονται αντιστοιχούν στον μέσο όρο εκτέλεσης της διαδικασίας αξιολόγησης 8 φορές, αποτελούν ισχυρές ενδείξεις πως η απόδοση της μεθόδου παραγοντοποίησης μητρώου υπερτερεί σημαντικά σε σχέση με τα MLP και NeuMF.

Οι συγγραφείς πάνε ένα βήμα πάρα πέρα και με σκοπό να καταδείξουν την δυσκολία προσέγγιση της πράξης του εσωτερικού γινομένου μεταξύ embedding διανυσμάτων από ένα MLP, κατασκευάζουν ένα τεχνητό πείραμα στο οποίο συγκρίνουν τη διαφορά στο RMSE (Ρίζα Μέσου Τετραγωνικού Σφάλματος) μεταξύ του εσωτερικού γινομένου και των MLP. Για το σκοπό αυτό κατασκευάζουν διανύσματα embeddings χρηστών και αντικειμένων από μία Κανονική (Gaussian) κατανομή και θέτουν τις πραγματικές τιμές αλληλεπίδρασης χρηστών-αντικειμένων ίσες με το εσωτερικό γινόμενο αυτών των Gaussian διανυσμάτων συν μια παράμετρο Gaussian θορύβου. Για να είναι σε θέση οι ερευνητές να ερμηνεύσουν την διαφορά

στο RMSE μεταξύ των 2 τεχνικών, οι παράμετροι των κατανομών επιλέγονται έτσι ώστε αυτό το τεχνητό πείραμα να ευθυγραμμίζεται με το πρόβλημα Netflix Prize, στο οποίο είναι γνωστό ποιες διαφορές στο RMSE θεωρούνται σημαντικές (0.01: πολύ σημαντική,0.001: σημαντική). Τα αποτελέσματα δείχνουν πως ένα MLP μπορεί να προσεγγίσει την απόδοση του εσωτερικού γινομένου υπό την προϋπόθεση τα κρυφά του επίπεδα να περιέχουν αρκετούς κρυφούς νευρώνες και δοθέντων αρκετών δεδομένων εκπαίδευσης. Ωστόσο το πλήθος αυτών των απαιτούμενων δεδομένων αυξάνεται πολυωνυμικά ως προς τις διαστάσεις των embedding διανυσμάτων και το αντίστροφο του επιθυμητού σφάλματος προσέγγισης. Αυτό σε συνδυασμό με την δυσκολία προσέγγισης της απόδοσης του εσωτερικού γινομένου από ένα MLP όσο αυξάνεται το πλήθος των διαστάσεων d, καθιστούν την διαφορά του RMSE μεταξύ MLP και εσωτερικού γινομένου ίση σχεδόν με 0.02, το διπλάσιο της τιμής η οποία δείχνει πολύ σημαντική διαφορά.

Επιπλέον άξιο σχολιασμού είναι το γεγονός πως η μέθοδος της παραγοντοποίησης μητρώου έχει σημαντικά πλεονεκτήματα όσον αφορά στην χρήση της σε εφαρμογές πραγματικού κόσμου. Χαρακτηριστικό παράδειγμα τα context-aware συστήματα στα οποία υπάρχει η ανάγκη σύστασης κάποιων αντικείμενων σε κάποιο χρήστη με βάση τα τελευταία αντικείμενα με τα οποία είχε αλληλεπίδραση κατά την περιήγηση του σε κάποιο ιστότοπο. Εδώ η ανάγκη εκπαίδευσης των MLP τα καθιστά απαγορευτικά για χρήση. Επιπρόσθετα για ένα πρόβλημα σύστασης όπου πρέπει να επιλεγεί ένα κατάλληλο υποσύνολο μεταξύ n αντικειμένων, η χρονική πολυπλοκότητα των MLP είναι O(d^2n) ενώ του εσωτερικού γινομένου O(dn). Η χρονική πολυπλοκότητα είναι απαγορευτική και στις δυο περιπτώσεις, αλλά η ύπαρξη αλγορίθμων που προσεγγίζουν τη λύση που παρέχει το εσωτερικό γινόμενο σε υπό-γραμμικό χρόνο το καθιστούν καταλληλότερο σε πρακτικές εφαρμογές.

Τέλος επισημαίνεται πως κάθε άλλο παρά αποθαρρύνεται η χρήση ΤΝΔ και γίνεται ειδική μνεία σε αρχιτεκτονικές ΤΝΔ, οι οποίες περιέχουν τα MLP, των οποίων η λειτουργία είναι ίσως πιο κατάλληλη για το πρόβλημα του συνεργατικού φιλτραρίσματος. Για παράδειγμα προτείνονται ΤΝΔ τα οποία χρησιμοποιούνται συχνά σε προβλήματα κατηγοριοποίησης, τα οποία θα αντιστοιχίζουν την αναπαράσταση ενός χρήστη σε ένα embedding διάνυσμα στο τελευταίο κρυφό τους επίπεδο και το επίπεδο εξόδους τους θα αποτελείται από πλήθος νευρώνων ίσο με το πλήθος αντικειμένων. Με αυτόν τον τρόπο για έναν δοσμένο χρήστη παράγεται μια αναμενόμενη τιμή αλληλεπίδρασης του με κάθε αντικείμενο, με την πράξη πριν το επίπεδο εξόδου που το πετυχαίνει αυτό να αντιστοιχεί στο εσωτερικό γινόμενο των embedding διανυσμάτων χρήστη-αντικειμένων. Επίσης ενδιαφέρον παρουσιάζει και η ιδέα στο [13] να λάβουμε το εξωτερικό γινόμενο μεταξύ των embedding διανυσμάτων και το δισδιάστατο μητρώο που θα προκύψει να τροφοδοτηθεί σε ένα Συνελικτικό Νευρωνικό Δίκτυο (CNN), των οποίων η χρήση παρουσιάζει πολλά πλεονεκτήματα, όπως ο διαμοιρασμός παραμέτρων και η τοπική εξαγωγή χαρακτηριστικών, μεταξύ άλλων.

3.3. On the Effectiveness of Low Rank Approximations for Collaborative Filtering compared to Neural Networks

Στην παρούσα εργασία [3] οι συγγραφείς επιχειρούν μέσα από κάποια πειράματα να εξετάσουν τους λόγους για τους οποίους μοντέλα MLP δεν φαίνεται να έχουν το ίδιο καλή απόδοση με μεθόδους που βασίζονται στην παραγοντοποίηση μητρώου και γιατί η τελευταία φαίνεται να ταιριάζει διαισθητικά με το πρόβλημα του συνεργατικού φιλτραρίσματος. Αρχικά παρατηρούν πως παρά την κυριαρχία και τα εξαιρετικά αποτελέσματα των ΤΝΔ σε πολλούς τομείς δεν φαίνεται να συμβαίνει το ίδιο στο συνεργατικό φιλτράρισμα. Αυτό ίσως να οφείλεται στην αραιότητα των αλληλεπιδράσεων χρηστών-αντικειμένων η οποία δεν επιτρέπει την αποδοτική εξαγωγή χαρακτηριστικών από τα MLP, γεγονός που οδήγησε στο συνδυασμό τους με μεθόδους παραγοντοποίησης μητρώου.

Αυτό που προτείνεται είναι ένας τρόπος σύνδεσης των μεθόδων παραγοντοποίησης μητρώου με την συνδιασπορά μεταξύ latent διανυσμάτων και αλληλεπιδράσεων χρηστών-αντικειμένων. Πρώτα από όλα γίνεται η θεώρηση πως κάθε αντικείμενο περιγράφεται πλήρως από ένα σύνολο d χαρακτηριστικών και υπό αυτή τη θεώρηση τα στοιχεία των d-διαστάτων latent διανυσμάτων των αντικειμένων ερμηνεύονται σαν την ισχύ του αντίστοιχου χαρακτηριστικού στο εκάστοτε αντικείμενο. Με την ίδια λογική τα στοιχεία των d-διάστατων latent διανυσμάτων των χρηστών αντιστοιχούν στην προτίμηση του κάθε χρήστη για το εκάστοτε χαρακτηριστικό. Αν τώρα κάποιος χρήστης έχει μια ιδιαίτερη προτίμηση σε κάποιο χαρακτηριστικό, αυτή η προτίμηση αναμένεται να αντανακλάται και στα αντικείμενα με τα οποία έχει αλληλεπίδραση αυτός ο χρήστης, δηλαδή θα παρατηρείται ισχυρή παρουσία του χαρακτηριστικού αυτού στα latent διανύσματα αυτών των αντικειμένων. Αναλυτικότερα αυτό σημαίνει πως θεωρώντας τα στοιχεία του μητρώου αλληλεπίδρασης σαν τις πραγματοποιήσεις μιας τυχαίας μεταβλητής R, τα στοιχεία των latent διανυσμάτων των χρηστών σαν τις πραγματοποιήσεις μιας τυχαίας μεταβλητής E_k^u και τα στοιχεία των latent διανυσμάτων των αντικειμένων σαν τις πραγματοποιήσεις μιας τυχαίας μεταβλητής E_k^v για ένα μόνο χαρακτηριστικό k, τότε για το σύνολο των χαρακτηριστικών ισχύει cov(R,$\sum_{k=1}^{d} a_k E_k^u E_k^v$)>0 όπου a_k συντελεστές βαρύτητας για το κάθε χαρακτηριστικό. Πιο συγκεκριμένα παρατηρείται πως πολλές φορές οι γραμμές και οι στήλες του μητρώου αλληλεπίδρασης είναι ισχυρά συσχετισμένες εξαιτίας, λόγου χάρη χρηστών που συμπεριφέρονται με παρόμοιο τρόπο ή αντικειμένων που αξιολογούνται με παρόμοιο τρόπο από το σύνολο των χρηστών (π.χ. δημοφιλή αντικείμενα). Η διαισθητική ερμηνεία πίσω από αυτό είναι πως μια χαμηλότερης τάξης προσέγγιση του μητρώου αλληλεπίδρασης θα περιέχει αρκετή, αν όχι όλη, την αναγκαία πληροφορία που απαιτείται για να προβλεφθούν οι αλληλεπιδράσεις χρηστών-αντικειμένων, ιδιότητα την οποία εκμεταλλεύονται οι μέθοδοι παραγοντοποίησης μητρώου, οι οποίες ικανοποιούν την παραπάνω σχέση θεωρώντας όλα τα χαρακτηριστικά ίσης σημαντικότητας και εκτελώντας το εσωτερικό γινόμενο μεταξύ των latent διανυσμάτων χρηστών-αντικειμένων.

Οι συγγραφείς για να επιβεβαιώσουν αυτή τους την θεώρηση εκτελούν ένα σύνολο πειραμάτων παραγοντοποίησης μητρώου στο MovieLens 100K. Αντιμετωπίζουν το παρόν σύνολο δεδομένων και σαν explicit άλλα και σαν implicit feedback σύνολο ενώ και για τις 2 περιπτώσεις χρησιμοποιούν διάσταση latent διανυσμάτων ίση με 32, σαν αλγόριθμο βελτιστοποίηση τον Adam και δεν κάνουν χρήση ομαλοποίησης. Υπολογίζουν την συνδιασπορά μεταξύ των αλληλεπιδράσεων ενός χρήστη και των latent διανυσμάτων των αντικειμένων με τα οποία είχε ή δεν είχε αλληλεπίδραση και αυτές οι συνδιασπορές συσχετίζονται με το latent διάνυσμα του χρήστη με την μετρική συσχέτισης Pearson, με την παραπάνω διαδικασία να επαναλαμβάνεται για κάθε χαρακτηριστικό. Με παρόμοιο τρόπο ορίζονται και οι συσχετίσεις μεταξύ των αντικειμένων και πάλι με τη μετρική συσχέτισης Pearson.

Με αυτόν τον τρόπο επιβεβαιώνεται εμπειρικά η θεώρηση πως οι αλληλεπιδράσεις χρηστών-αντικειμένων που προκύπτουν με τη μέθοδο της παραγοντοποίησης μητρώου βασίζονται στη συνδιασπορά μεταξύ των προτιμήσεων των χρηστών και των χαρακτηριστικών των αντικειμένων, αφού για την περίπτωση που το σύνολο δεδομένων αντιμετωπίζεται σαν implicit feedback η μέση συσχέτιση των χρηστών ήταν 0.8246 ενώ η μέση συσχέτιση των

αντικειμένων 0.7256 ενώ στην περίπτωση που αντιμετωπίζεται σαν explicit feedback οι μέσες συσχετίσεις ήταν 0.8987 και 0.8156 για χρήστες και αντικείμενα αντίστοιχα.

Για τη σύγκριση μεταξύ MLP και της παραγοντοποίησης μητρώου χρησιμοποιείται μόνο η implicit feedback εκδοχή του MovieLens 100K. Πιο συγκεκριμένα εξετάζεται κατά πόσο η χρήση latent διανυσμάτων που έχουν εξαχθεί από την παραγοντοποίηση μητρώου στις εισόδους ενός ΤΝΔ επηρεάζει την απόδοση του ως προς τις μετρικές MRR (Mean Reciprocal Rank), MAP@10 (Mean Average Precision) και AUC (Area Under Curve). Χρησιμοποιούνται 3 παραλλαγές. Είτε τα latent διανύσματα αποτελούν ελεύθερες παραμέτρους και προσαρμόζονται κατά την εκπαίδευση μαζί με τα βάρη του MLP, όμοια με τις προηγούμενες εργασίες που εξετάσαμε, είτε αρχικοποιούνται τα latent διανύσματα με τις τιμές που έλαβαν από την μέθοδο της παραγοντοποίησης μητρώου. Στην δεύτερη περίπτωση δοκιμάζεται είτε να μένουν αμετάβλητα καθ' όλη την διάρκεια της εκπαίδευσης ή να προσαρμόζονται μαζί με τα βάρη του MLP. Επίσης οι συγγραφείς δοκιμάζουν δυο διαφορετικούς τρόπους τροφοδότησης των MLP με τα latent διανύσματα. Χρησιμοποιούν την συνένωση των δυο διανυσμάτων και τον στοιχείο προς στοιχείο πολλαπλασιασμό των στοιχείων τους. Όλοι οι παραπάνω συνδυασμοί εκτελούνται για 0, 1, 2 και 3 κρυφά επίπεδα και διάφορες συναρτήσεις ενεργοποίησης.

Τα πειράματα δείχνουν πως η τροφοδότηση ενός MLP με την συνένωση των latent διανυσμάτων χρηστών και αντικειμένων δεν μπορεί να πλησιάσει την απόδοση της παραγοντοποίησης μητρώου και εικάζεται πως αυτό οφείλεται στη μεγάλη ευελιξία των MLP, στην ικανότητα τους δηλαδή να προσεγγίζουν οποιασδήποτε συνεχή συνάρτηση. Ισχυρή ένδειξη προς αυτήν την κατεύθυνση παρέχει το γεγονός πως εάν το MLP τροφοδοτηθεί με τη συνένωση των latent διανυσμάτων που έχουν προκύψει από την μέθοδο παραγοντοποίησης μητρώου έχει σημαντικά καλύτερη απόδοση, η οποία βελτιώνεται περαιτέρω εάν δεν επιτραπεί η τροποποίηση των latent διανυσμάτων κατά την εκπαίδευση. Ωστόσο τα ΤΝΔ πετυχαίνουν την καλύτερη απόδοση λαμβάνοντας τον στοιχείο προς στοιχείο πολλαπλασιασμό των latent διανυσμάτων στην είσοδο τους. Πιο συγκεκριμένα τροφοδοτώντας ένα MLP με τον στοιχείο προς στοιχείο πολλαπλασιασμό των latent διανυσμάτων που προέκυψαν από το καλύτερο μοντέλο παραγοντοποίησης μητρώου οι συγγραφείς έλαβαν καλύτερα αποτελέσματα κατά 8.58% σε σχέση με την παραγοντοποίηση μητρώου και υποθέτουν πως αυτό οφείλεται στην προσαρμογή των παραμέτρων a_k της παραπάνω σχέσης, δηλαδή στην ζύγιση της σημαντικότητας των χαρακτηριστικών.

4. Συμπεράσματα και Προοπτικές

Παρά το γεγονός πως οι παραπάνω εργασίες που αναλύθηκαν επικεντρώνονται στην χρήση MLP, υπάρχουν κι άλλες πολλές διαφορετικές αρχιτεκτονικές ΤΝΔ οι οποίες μπορούν να χρησιμοποιηθούν στο συνεργατικό φιλτράρισμα. Αναφερθήκαμε ήδη σε έναν τρόπο χρήσης των CNN [13]. Άλλες επιλογές αποτελούν η χρήση Autoencoders και παραλλαγών τους [11] τα οποία αντιστοιχούν σε μοντέλα μη-επιβλεπόμενης μάθησης που επιχειρούν να ανακατασκευάσουν την είσοδο τους στο επίπεδο εξόδου τους, η χρήση Recurrent Neural Networks (RNN) τα οποία εφαρμόζονται ευρέως σε context-aware εφαρμογές [9], η χρήση Restricted Boltzman Machines (RBM) [8] τα οποία αποτελούνται από ένα ορατό και ένα κρυφό επίπεδο με τους νευρώνες του κάθε επιπέδου να επικοινωνούν μόνο με τους νευρώνες του άλλου επιπέδου και ποτέ μεταξύ τους, ή ακόμα και η χρήση Generative Adversial Networks (GAN) [7] τα οποία κάνουν χρήση ανταγωνιστικής μάθησης μεταξύ ενός παραγωγού νέων τεχνητών δειγμάτων και ενός διαχωριστή μεταξύ τεχνητών και πραγματικών δειγμάτων. Τέλος με βάση τα αποτελέσματα της $3^{ης}$ εργασίας που εξετάσαμε, τεχνικές που συνδυάζουν την παραγοντοποίηση μητρώου με τα MLP θα μπορούσαν να επιφέρουν βελτιωμένα αποτελέσματα όπως στην εργασία [10].

Εν κατακλείδι παρά το γεγονός πως η επιστράτευση των ΤΝΔ στο πεδίο των συστημάτων συστάσεων αποτελεί πεδίο έντονης ερευνητικής δραστηριότητας [4], πολλές φορές μπορεί η καλή τους απόδοση σε κάποιο σύνολο δεδομένων να μας ξεγελά σχετικά με την γενικότερη ικανότητα τους σε παρόμοια προβλήματα και την θεώρηση τους ως την καλύτερη επιλογή. Αυτό αποτελεί γενικότερο πρόβλημα στο πεδίο των συστημάτων συστάσεων εξαιτίας μεταξύ άλλων της έλλειψης ενός κοινώς αποδεκτού συνόλου δεδομένων (όπως π.χ. το ImageNet στην αναγνώριση εικόνων) για την αξιολόγηση των προτεινόμενων τεχνικών σε κοινή βάση ή της γενικότερης δυσκολίας που υπάρχει στην αξιολόγηση των μεθόδων στην συγκεκριμένη επιστημονική περιοχή [5].

Σημαντικό πρόβλημα, το οποίο καθιστά ακόμα πιο δύσκολη την σύγκριση μεταξύ των προτεινόμενων μοντέλων, αποτελεί το γεγονός πως στις περισσότερες περιπτώσεις δίνεται περισσότερη έμφαση από τους ερευνητές στη βελτιστοποίηση της τεχνικής που προτείνουν χωρίς όμως να αφιερώνεται η ίδια προσπάθεια και στις τεχνικές σύγκρισης. Αυτό αποτελεί μόνο ένα από τα πολλά προβλήματα που υπάρχουν για την δίκαιη σύγκριση μεταξύ των μεθόδων που προτείνονται για τα συστήματα συστάσεων, όπως παρατηρείται μεταξύ άλλων στην εργασία [6].

Αναφορές

[1] S. Rendle, W. Krichene, L. Zhang και J. Anderson, «Neural Collaborative Filtering vs. Matrix Factorization Revisited,» σε *Fourteenth ACM Conference on Recommender Systems*, 2020.

[2] X. He, L. Liao, H. Zhang, L. Nie, X. Hu και T.-S. Chua, «Neural Collaborative Filtering,» σε *WWW '17 Proceedings of 26th International Conference on World Wide Web*, 2017.

[3] M. Kurovski και F. Wilhelm, «On the Effectiveness of Low-rank approximations for Collaborative Filtering compared to Neural Networks,» arXiv: Information Retrieval, 2019.

[4] S. Zhang, L. Yao, A. Sun και Y. Tay, «Deep Learning Based Reccomender System: A Survey and New Perspectives,» *ACM Computing Surveys,* τόμ. 52, αρ. 1, p. 5, 2019.

[5] S. Rendle, L. Zhang και Y. Koren, «On the Difficulty of Evaluating Baselines,» arXiv: Information Retrieval, 2019.

[6] M. F. Dacrema, S. Boglio, P. Cremonesi και D. Jannach, «A Troubling Analysis of Reproducibility and Progress in Reccomender Systems Research,» arXiv: Information Retrieval, 2019.

[7] X. He, Z. He, X. Du και T.-S. Chua, «Adversial Personalized Ranking for Reccomendation,» σε *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2018.

[8] K. Georgiev και P. Nakov, «A non-IID Framework for Collaborative Filtering with Restricted Boltzman Machines,» σε *International Conference on Machine Learning*, 2013.

[9] A. Beutel, P. Covington, S. Jain, C. Xu, J. Li, V. Gatto και E. H.Chi, «Latent Cross: Making Use of Context in Recurrent Recommender Systems,» σε *Web Search and Data Mining*, 2018.

[10] B. Zheng και M. Mao, «Neural Metric Matrix Factorization,» *IOP Conference Series: Materials Science and Engineering,* τόμ. 768, αρ. 5, p. 52077, 2020.

[11] X. Dong, L. Yu, Z. Wu, Y. Sun, L. Yuan και F. Zhang, «A Hybrid Collaborative Filtering Model with Deep Structure for Recommender Systems,» σε *National Conference on Artificial Intelligence*, 2017.

[12] J. Bennett και S. Lanning, «The Netflix Prize,» 2007.

[13] X. He, X. Du, X. Wang, F. Tiang, J. Tang και T.-S. Chua, «Outer product-based neural collaborative filtering,» σε *International Joint Conference on Artificial Intelligence*, 2018.