ΑΡΘΡΟ ΑΝΑΣΚΟΠΗΣΗΣ

Εξατομικευμένη Ιατρική και μαζικά βιοϊατρικά δεδομένα

Μαργαρίτα-Ιωάννα Κουφάκη,*, Ιωάννης Γ. Χατζής[2], Γεώργιος Π. Πατρινός[1]

[1]Πανεπιστήμιο Πατρών, Σχολή Επιστημών Υγείας, Τμήμα Φαρμακευτικής, Πάτρα

[2]Δημόσιο Ινστιτούτο Επαγγελματικής Κατάρτισης Μεσολογγίου, Μεσολόγγι

**Πανεπιστήμιο Πατρών, Σχολή Επιστημών Υγείας, Τμήμα Φαρμακευτικής, Εργαστήριο Φαρμακογονιδιωματικής και Εξατομικευμένης Θεραπείας, Πανεπιστημιούπολη, Ρίο, 265 04, Πάτρα. Τηλέφωνο: 2610-962339, Email: gpatrinos@upatras.gr*

ΠΕΡΙΛΗΨΗ

Η εξατομικευμένη ιατρική αναπτύχθηκε τα τελευταία χρόνια ως μια νέα θεραπευτική προσέγγιση που αφορά τη θεραπεία και περίθαλψη ενός ασθενή με γνώμονα τις βιολογικές και περιβαλλοντικές πληροφορίες που έχουν συγκεντρωθεί για αυτόν. Κάθε μέρα ένας μεγάλος όγκος βιοϊατρικών δεδομένων συγκεντρώνεται και αποθηκεύεται σε ηλεκτρονικούς φακέλους υγείας, δημιουργώντας έναν τεράστιο όγκο δεδομένων, τα οποία παρουσιάζουν διάφορες προκλήσεις για τη διαχείριση και την ανάλυσή τους. Διάφοροι τομείς γνώσεων είναι απαραίτητοι για την πρόσκτηση των δεξιοτήτων και ικανοτήτων που απαιτούνται για να αξιοποιούνται πλήρως οι δυνατότητες που προσφέρει αυτός ο μεγάλος όγκος των βιοϊατρικών δεδομένων. Αρχίζουμε με μια εισαγωγή στην έννοια «μαζικά δεδομένα» καθώς και στην αποθήκευση και στη διαχείρισή τους. Εξετάζουμε τις στατιστικές μεθόδους και την επιστήμη των δεδομένων, που αποτελούν σημαντικά θεμέλια για την τεχνητή νοημοσύνη, τη μηχανική εκμάθηση και την επεξεργασία φυσικής γλώσσας, απαραίτητες προϋποθέσεις για την ανάπτυξη προγνωστικών μοντέλων με σκοπό τη λήψη κλινικών αποφάσεων. Εν κατακλείδι, προτείνουμε ειδική κατάρτιση για την προετοιμασία της νέας γενιάς επιστημόνων για τον χειρισμό των μαζικών βιοϊατρικών δεδομένων.

ΛΕΞΕΙΣ ΕΥΡΕΤΗΡΙΟΥ: Εξατομικευμένη Ιατρική, μαζικά βιοϊατρικά δεδομένα, τεχνητή νοημοσύνη, μηχανική εκμάθηση, ηλεκτρονικός φάκελος υγείας

Εξατομικευμένη ιατρική και Ιατρική Ακριβείας

Ιατρική ακριβείας ονομάζεται η διαδικασία προσαρμογής της ιατρικής πράξης στα ατομικά χαρακτηριστικά (γενετικά και μη) ενός συγκεκριμένου ασθενούς [1-3]. Κάτι τέτοιο δεν συνεπάγεται τη δημιουργία στοχευμένων φαρμακευτικών προϊόντων προσαρμοσμένων στο γονότυπο ενός συγκεκριμένου ασθενούς, μιας και αυτός είναι στόχος της εξατομικευμένης ιατρικής. Στην πραγματικότητα, η διάκριση μεταξύ της ιατρικής ακρι-

*** Αντεπιστέλλων Συγγραφέας**

Πανεπιστήμιο Πατρών, Σχολή Επιστημών Υγείας, Τμήμα Φαρμακευτικής, Πανεπιστημιούπολη, Ρίο, 265 04, Πάτρα.

Τηλεφωνο: 2610-962339, Email: koufakimargianna@gmail.com

βείας και της εξατομικευμένης ιατρικής έγκειται στο ότι η πρώτη επικεντρώνεται στη δημιουργία ενός νέου τρόπου ταξινόμησης μιας ασθένειας, ο οποίος λαμβάνει υπόψη του τα μοναδικά χαρακτηριστικά ενός ασθενούς, ενώ η εξατομικευμένη ιατρική επικεντρώνεται στην ανακάλυψη/ανάπτυξη εξειδικευμένων μεθόδων θεραπείας (π.χ. φάρμακα, ανοσοθεραπείες) που στοχεύουν στις ατομικές γονιδιωματικές υπογραφές του ασθενούς [2]. Αυτή η ειδοποιός διαφορά οδήγησε στη συχνότερη εφαρμογή της «εξατομικευμένης ιατρικής» από τη φαρμακοβιομηχανία, ενώ την "ιατρική ακριβείας" μελετούν ιδιαίτερα οι κλινικοί ιατροί που ενδιαφέρονται να σχεδιάσουν κατάλληλες κλινικές οδηγίες για διαφορετικές υποομάδες ασθενών.

Προκειμένου να προσαρμόσουν τις ιατρικές θεραπείες στα χαρακτηριστικά μεμονωμένων ασθενών, οι ερευνητές έχουν επικεντρωθεί στη διάκριση ασθενών σε «υποομάδες ιδιαίτερης σημασίας». Η δημιουργία αυτών των υποομάδων «ιδιαίτερης σημασίας» γίνεται με γνώμονα ένα συγκεκριμένο χαρακτηριστικό του ασθενούς. Για παράδειγμα, αυτό το χαρακτηριστικό θα μπορούσε να είναι η εθνικότητα, το φύλο ή η σεξουαλική ταυτότητα, η κοινωνικοοικονομική ομάδα (χαμηλό ή υψηλό εισόδημα), ή ακόμη ο τύπος της ασθένειας (π.χ. άσθμα).

Οι ασθενείς μπορεί επίσης να συσχετισθούν μεταξύ τους έχοντας ως κριτήριο τον τύπο της αλλεργίας τους. Για παράδειγμα, όλοι οι ασθενείς με αλλεργίες στα αυγά είναι μια υποομάδα ασθενών ιδιαίτερης σημασίας που απαιτεί προσαρμοσμένα θεραπευτικά σχήματα που είναι ιδιαίτερα σημαντικό να γνωρίζουν οι κλινικοί ιατροί. Π.χ. Αυτοί οι ασθενείς θα πρέπει να λάβουν το εμβόλιο κατά της γρίπης σε διαιρεμένη δόση [4] και όχι σε εφάπαξ.

Ενώ κάποια χαρακτηριστικά είναι ήδη γνωστά για τη δημιουργία υποομάδων ασθενών, όπως όσων πάσχουν από τροφικές αλλεργίες, όπως αναφέρθηκε προηγούμενα, ωστόσο, δεν είναι γνωστά χαρακτηριστικά με τα οποία θα μπορούσε να δημιουργηθούν νέες υποομάδες ασθενών. Για το σκοπό αυτό έχουν σχεδιαστεί μέθοδοι πληροφορικής για τη δημιουργία υποομάδων ασθενών, οι οποίες αναζητούν κοινά χαρακτηριστικά α) στα δεδομένα που έχουν καταγραφεί στον ηλεκτρονικό φάκελο υγείας [5], β) στις χρονικές μεταβολές στις καταστάσεις εργαστηριακών τιμών (π.χ. ελεγχόμενος έναντι ανεξέλεγκτου διαβήτη) [6], γ) στις ανεπιθύμητες αντιδράσεις σε φάρμακα που έχουν αναφερθεί λόγω γενετικών παραγόντων όπως μεταλλάξεις του CYP [7] αλλά και δ) στους τύπους καρκίνου [8]. Απαιτούνται εξελιγμένες μέθοδοι πληροφορικής για τον εντοπισμό συγκεκριμένων πληθυσμών ασθενών και τη διαστρωμάτωση σε υποπληθυσμούς ιδιαίτερης κλινικής σημασίας, ώστε να μπορεί να εφαρμοστεί σ' αυτούς ιατρική ακριβείας. Αυτό παραμένει ένα ερευνητικό πεδίο με σημαντικές προκλήσεις, καθώς κάθε μεμονωμένος ασθενής εμφανίζει ένα σύνθετο συνδυασμό φαινοτύπων της νόσου και συμπτωματολογίας [9].

Μαζικά βιοϊατρικά δεδομένα

Με τον όρο «μαζικά δεδομένα -Big Data» ορίζουμε τον τεράστιο όγκο δεδομένων ο οποίος με τις παραδοσιακές υπολογιστικές μεθόδους και τεχνικές είναι σχεδόν αδύνατο να αναλυθεί και να εξαχθούν αποτελέσματα.

Οι υπολογιστικές προκλήσεις αφορούν τη χωρητικότητα των μέσων αποθήκευσης δεδομένων, το εύρος ζώνης του δικτύου για τη μετακίνηση δεδομένων από μια συσκευή αποθήκευσης σε μια άλλη ή το πλήθος και το είδος των υπολογισμών που είναι απαραίτητοι για την επεξεργασία και την ανάλυση των δεδομένων. Επίσης, πολλές φορές τα υπολογιστικά αποτελέσματα που παράγονται συμβαίνει να υπερβαίνουν τον όγκο των ίδιων των δεδομένων.

Τα μαζικά δεδομένα χαρακτηρίζονται συχνά από τα τέσσερα V.

Το πρώτο V είναι ο όγκος (Volume) των δεδομένων. Αυτό το χαρακτηριστικό αναφέρεται στη χωρητικότητα των μέσων που απαιτείται για την αποθήκευση των δεδομένων και είναι ένα από τα πιο σημαντικά προβλήματα που αντιμετωπίζει κάποιος με τα μαζικά δεδομένα. Το δεύτερο V είναι η ταχύτητα (Velocity). Τα δεδομένα ίσως να παράγονται από κάποιο σύστημα ηλεκτρονικών μετρήσεων όπως για παράδειγμα μια φορητή συσκευή η οποία δημιουργεί τα δεδομένα πιο γρήγορα από ότι η υπολογιστική δομή που κάποιος διαθέτει μπορεί να διαχειριστεί και να αποθηκεύσει. Αυτό καθιστά αδύνατη τη μεταφορά και αποθήκευση δεδομένων από τη μια συσκευή στην άλλη. Το τρίτο V είναι η ποικιλία (Variety). Τα μαζικά δεδομένα δεν είναι πάντα ομοιογενή αλλά είναι συχνά ένας ετερογενής συνδυασμός τύπων δεδομένων που προέρχονται από διαφορετικές πηγές μέτρησης. Ένα σχετικό παράδειγμα αποτελούν οι ηλεκτρονικοί φάκελοι υγείας, γιατί περιέχουν δεδομένα διαφορετικής μορφής, όπως θα αναλύσουμε παρακάτω. Το τέταρτο V είναι η αξιοπιστία (Veracity), δηλαδή πόσο ακριβή και σαφή είναι τα δεδομένα ώστε να μπορεί να τα εμπιστευτεί κανείς ως αξιόπιστα. Τα μαζικά δεδομένα περιέχουν συχνά ασάφειες, λανθασμένες ή ελλιπείς τιμές. Αυτό δημιουργεί τεράστιες προκλήσεις στην προσπάθεια να υπάρξει τελικά ένα σύνολο δεδο-

μένων έτοιμο για ανάλυση.

Όλα τα παραπάνω χαρακτηριστικά θεωρούνται τα πιο συνηθισμένα στη βιβλιογραφία. Όπως, υπάρχουν τουλάχιστον δύο ακόμη που θα μπορούσαν να συμπεριληφθούν σε αυτά τα τέσσερα. Το πρώτο είναι η πολυπλοκότητα (Vexedness) ή μεταβλητότητα. Τα δεδομένα που είναι ιεραρχικά ή διαχρονικά προσθέτουν σύνθετες διαστάσεις όταν μάλιστα συνδυάζονται με όλα τα παραπάνω V.

Το δεύτερο, είναι η αξία (Value) των δεδομένων. Υπάρχει όπως προβληματισμός ορισμένες φορές για το αν τα δεδομένα που συλλέγονται είναι αξιόλογα δικιολογειται να τεθεί σε εφαρμογή μια διαδικασία για τη μαζική αποθήκευση δεδομένων και επεξεργασίας τους με υπολογιστές υψηλής απόδοσης. Τα μαζικά δεδομένα δεν είναι πάντα η βέλτιστη λύση, και όπως ορισμένοι ισχυρίζονται οι στοχευμένες προσεγγίσεις με ανάλυση μικρότερου όγκου δεδομένων ίσως να έχουν καλύτερα αποτελέσματα στην απάντηση ορισμένων επιστημονικών ερωτημάτων [10].

Ηλεκτρονικός Φάκελος Υγείας – ΗΦΥ

Ο ηλεκτρονικός φάκελος υγείας αποτελεί ένα σύγχρονο παράδειγμα σημαντικής πηγής μαζικών δεδομένων που έχει ευρεία εφαρμογή σε χιλιάδες ακαδημαϊκά και ιατρικά κέντρα παγκοσμίως. Λόγω των χαρακτηριστικών του έχει αποκτήσει ένα σπουδαίο ρόλο στην παρακολούθηση των δεδομένων και πληροφοριών των ασθενών αλλά και στη βελτίωση της τιμολόγησης των ιατρικών υπηρεσιών. Η εξέλιξη της τεχνολογίας της πληροφορικής επέτρεψε την υιοθέτηση του ΗΦΥ κυρίως λόγω της ανέξοδης αποθήκευσης δεδομένων και τις διαθέσιμες βάσεις δεδομένων που μπορούν να χειριστούν αποτελεσματικά μαζικά δεδομένα. Οι βάσεις δεδομένων που χρησιμοποιούνται για τους ΗΦΥ έχουν τεράστιες δυνατότητες. Πιο συγκεκριμένα, μπορούν να ενσωματώσουν ένα εντυπωσιακό εύρος διαφορετικών δεδομένων ασθενών, συμπεριλαμβανομένων δημογραφικών στοιχείων, εργαστηριακών εξετάσεων, απεικονίσεων, ιατρικού ιστορικού και της χρήσης οποιασδήποτε φαρμακευτικής αγωγής, καθώς και κλινικών σημειώσεων που περιλαμβάνουν σχόλια τόσο από τον ιατρό όσο και από τον ασθενή σε ελεύθερο κείμενο. Το τελευταίο διάστημα προστίθενται στα παραπάνω όλο και περισσότερα γονιδιωματικά δεδομένα αλλά και δεδομένα επιτήρησης υγείας που καταγράφονται από ενδυόμενες ή έξυπνες συσκευές.

Αυτές οι νέες πηγές μαζικών δεδομένων δημιουργούν προκλήσεις στη διαχείριση των δεδομένων αλλά και στη χρήση τους για τη λήψη κλινικών αποφάσεων. Με άλλα λόγια, η συσσώρευση πρωτογενών δεδομένων οδηγεί στην ανάγκη για ανάπτυξη μεθόδων διαχείρισης των ροών δεδομένων αλλά και προγραμμάτων μετατροπής τους σε μορφή αξιοποιήσιμη από την εξατομικευμένη ιατρική. Επιπλέον για κάθε ασθενή προκύπτουν νέα δεδομένα μέσω της χρήσης ιατρικών εφαρμογών αλλά και έξυπνων φορητών συσκευών. Σύμφωνα με μελέτες, εκτιμάται ότι κάθε ασθενής στο εγγύς μέλλον θα έχει πολλά terabytes ή ακόμα και petabytes δεδομένων και πληροφοριών που θα πρέπει να αποθηκευτούν και να επεξεργαστούν καταλλήλως ως μέρος της φροντίδας του. Αυτά τα δεδομένα θα χρησιμεύσουν ως πρώτη ύλη τόσο για εξατομικευμένη ιατρική όσο και για ιατρική ακρίβειας οι οποίες θα αναπτυχθούν με τη βοήθεια των μεθόδων που περιγράφονται παρακάτω.

Διαχείριση και Ενσωμάτωση Δεδομένων

Εάν υποθέσουμε ότι τα δεδομένα είναι οι δομικοί λίθοι για την έρευνα και εφαρμογή της εξατομικευμένης ιατρικής, τότε οι βάσεις δεδομένων είναι το κονίαμα που εξασφαλίζει την συνοχή τους. Οι βάσεις δεδομένων παρέχουν τη δομή στην οποία διατηρούνται και διατίθενται τα δεδομένα για μελλοντική χρήση. Τα συστήματα βάσεων δεδομένων παρέχουν τους υπολογιστικούς μηχανισμούς, που χρειάζονται για την αποθήκευση, επεξεργασία και ανάκτηση δεδομένων, συνήθως μέσω γραφικών διεπαφών και γλωσσών επερωτήσεων. Η κυρίαρχη αρχιτεκτονική των βιοϊατρικών βάσεων δεδομένων είναι σχεσιακή. Τα δεδομένα αποθηκεύονται σε πίνακες που αντιπροσωπεύουν μια συγκεκριμένη οντότητα, όπως τα δημογραφικά στοιχεία.

Το τελευταίο διάστημα παρατηρείται ένα αυξανόμενο ενδιαφέρον για τις βάσεις δεδομένων στις οποίες τα δεδομένα παρουσιάζονται ως κόμβοι σε ένα μη κατευθυνόμενο γράφημα και οι σχέσεις μεταξύ των κόμβων ως σύνδεσμοι [11,12]. Όποια αρχιτεκτονική και να έχει η βάση δεδομένων, ο χρήστης μπορεί να χειριστεί τα δεδομένα μέσω ερωτημάτων χρησιμοποιώντας μια συγκεκριμένη γλώσσα προσαρμοσμένη για τη βάση αυτή. Ανεξάρτητα από την αρχιτεκτονική της βάσης δεδομένων, όλες οι βάσεις δεδομένων σχεδιάζονται και υλοποιούνται χρησιμοποιώντας ένα μοντέλο δεδομένων, το οποίο καθορίζει τα χαρακτηριστικά της βάσης δεδομένων και παρέχει οδηγίες για τους εμπλεκόμενους.

Είναι αλήθεια ότι οι ειδικοί της επιστήμης δεδομένων διστάζουν να αναπτύξουν εξελιγμένες βάσεις δεδομένων για την αποθήκευση και ενσωμάτωση βιοϊατρικών δεδομένων λόγω της πολύπλοκης φύσης αυτών των δεδομένων. Μολαταύτα, οι ειδικοί διαθέτουν κάποιες υπολογιστικές δυνατότητες ώστε να ενσωματώνουν ή

να συνδέουν φαινομενικά ανόμοιες βάσεις δεδομένων για να προκύπτει μια ολοκληρωμένη εικόνα ενός κλινικού προβλήματος. Για παράδειγμα, στην εκτίμηση της αιτίας για την οποία το HbA1C (glycated haemoglobin) του διαβητικού ασθενούς είναι εκτός ελέγχου, τα δεδομένα που λαμβάνονται από τη μόνιμη παρακολούθηση της σωματικής δραστηριότητας με τη βοήθεια μιας φορητής προσωπικής συσκευής, όπως το Fitbit, θα μπορούσαν να προστεθούν στα κλινικά δεδομένα του ασθενούς. Αυτό θα βοηθούσε τον ιατρό να ποσοτικοποιήσει τη σωματική δραστηριότητα ενός συγκεκριμένου ασθενούς για να βελτιστοποιήσει το συνολικό πρόγραμμα διαβητικού ελέγχου αυτού. Η άνθιση νέων πηγών δεδομένων οδήγησε σε ένα αυξημένο ενδιαφέρον για ανάπτυξη νέων μεθόδων σύνδεσης και συσχέτισης των διαφόρων αρχείων, αναγνωρίζοντας παράλληλα τις δυσκολίες και προκλήσεις που αυτό συνεπάγεται [13,14].

Προκλήσεις

Πέρα από την ποικιλία τύπων και πηγών δεδομένων που περιέχει ο ΗΦΥ, υπάρχουν και άλλα στοιχεία ποικιλομορφίας στα βιοϊατρικά δεδομένα που δυσχεραίνουν την ενσωμάτωσή τους σε βάσεις δεδομένων. Η διαφορά μονάδων μέτρησης ανάμεσα στις βάσεις δεδομένων είναι, για παράδειγμα, ένα συχνό πρόβλημα που εμποδίζει την καθολική επεξεργασία των δεδομένων και την κανονικοποίησή τους. Αυτή η ποικιλομορφία στην παρουσίαση των δεδομένων αποτελεί μια σημαντική πρόκληση για την αποτελεσματική και ακριβή συσχέτιση και σύνδεση των δεδομένων καθώς και για την τελική ενσωμάτωσή τους στη βάση.

Για να αντιμετωπιστεί αυτή η πρόκληση, πρέπει να στραφούμε στις επιστημολογικές διαστάσεις των βιοϊατρικών δεδομένων με τον τρόπο που αυτές παρουσιάζονται στο συντακτικό τους μέρος και στη σημασιολογία τους. Αυτό είναι απαραίτητο για την εναρμόνιση των δεδομένων αλλά και για την ενσωμάτωση τους στη συνέχεια. Όλο και περισσότερο, οι επαγγελματίες της πληροφορικής στρέφονται στις οντολογίες για να πραγματοποιήσουν τη διαδικασία εναρμόνισης των δεδομένων, προκειμένου να αποδώσουν τις έννοιες και τις σχέσεις μεταξύ τους σε μορφή γραφημάτων- πχ η ενσωμάτωση των multi-omics δεδομένων στην υπάρχουσα βιολογική γνώση [15]. Ως εκ τούτου, οι οντολογίες μπορούν να θεωρηθούν ως τύπος μοντέλου δεδομένων.

Ένα άλλο πιεστικό ζήτημα στη διαχείριση και ενσωμάτωση δεδομένων είναι η διασφάλιση της ποιότητας αυτών. Ένας σοβαρός λόγος για την κακή ποιότητα των δεδομένων είναι ότι υπάρχουν ελλιπή στοιχεία, αφού είναι πιθανό εργαστηριακές αναφορές να μην αποθηκευτούν στον ιατρικό φάκελο ή τμήμα μιας φυσικής εξέτασης να μην ολοκληρωθεί. Προκειμένου να προκύψουν αποτελεσματικά συμπεράσματα από τα δεδομένα, πρέπει να υπάρχει ένα προκαθορισμένο πρωτόκολλο για τη διαχείριση των δεδομένων που λείπουν, τα οποία μπορεί να περιλαμβάνουν διάφορες μεθόδους καταλογισμού, όπως το deep learning [16]. Οι Kim και συνεργάτες εφάρμοσαν ένα νέο πλαίσιο ενσωμάτωσης για την πρόβλεψη τιμών ίδιας μορφής με τα δεδομένα που λείπουν [17].

Μια πιθανή λύση στα προβλήματα που προκύπτουν στη διαχείριση μαζικών κλινικών δεδομένων είναι οι αποθήκες κλινικών δεδομένων μεγάλης κλίμακας. Αυτές αποτελούν σημαντικούς πόρους δεδομένων, η δημιουργία και συντήρηση των οποίων εξυπηρετούν μια μεγάλη ποικιλία χρηστών και ενδιαφερομένων στον τομέα της εξατομικευμένης ιατρικής. Κατά τη διαδικασία εξαγωγής-μεταφοράς-φόρτωσης που είναι ο κύριος ρόλος του παραδείγματος χρήσης μιας αποθήκης δεδομένων, μπορούν να εφαρμοστούν διαδικασίες που εξασφαλίζουν την ποιότητα των δεδομένων και ενσωματώνουν δεδομένα από διάφορες πηγές με αποτέλεσμα την παροχή μιας ασφαλούς πλατφόρμας για την ανάλυση των δεδομένων που δεν έχουν ταυτοποιηθεί. Αυτές οι αποθήκες θα μπορούσαν να προσφέρουν στους ερευνητές τη δυνατότητα ταυτοποίησης και εξαγωγής επιθυμητού κοόρτιου και στους ειδικούς της ιατρικής ακριβείας τα μέσα για την αντιμετώπιση ενός ασθενούς με ένα συγκεκριμένο φαινότυπο την ώρα της περίθαλψης.

Στατιστική Ανάλυση

Η στατιστική ανάλυση είναι ο κλάδος των μαθηματικών που χρησιμοποιεί μοντέλα για να συνοψίσει τα δεδομένα και να εξαχθούν συμπεράσματα. Το όριο μεταξύ της στατιστικής ανάλυσης και της μηχανικής εκμάθησης (machine learning, το πεδίο της τεχνητής νοημοσύνης που επιτρέπει στα συστήματα ηλεκτρονικών υπολογιστών να μαθαίνουν από τα δεδομένα) είναι απροσδιόριστο και αποτελεί κοινό θέμα συζήτησης. Κάθε ένας από τους δύο αυτούς κλάδους επικεντρώνεται σε μια διαφορετική πτυχή της διαδικασίας εξαγωγής συμπερασμάτων από τα δεδομένα. Η μηχανική εκμάθηση επικεντρώνεται στην πραγματοποίηση προβλέψεων ακριβείας από τα δεδομένα, ενώ η στατιστική ανάλυση επικεντρώνεται στην αξιολόγηση της εγκυρότητας ενός μοντέλου για τα δεδομένα για την εξαγωγή συμπερασμάτων. Έτσι, η στατιστική ανάλυση και η εκμη-

χανική μάθηση αλληλοκαλύπτουν τα μεταξύ τους κενά ξεπερνώντας τους υπάρχοντες περιορισμούς. Αρκετοί μάλιστα πιστεύουν ότι, καθώς τα επιστημονικά πεδία των εφαρμοσμένων μαθηματικών και της επιστήμης των υπολογιστών συνεχίζουν να αλληλεπιδρούν, οι δύο κλάδοι θα συγχωνευθούν σε ένα μόνο πεδίο, που μερικές φορές αποκαλείται στατιστική εκμάθηση [18].

Η σημασία της στατιστικής ανάλυσης στις βιοϊατρικές εφαρμογές είναι αξιοσημείωτη. Εστιάζοντας το αντικείμενο της μελέτης στο μοντέλο των δεδομένων και καθιστώντας σαφείς τις παραδοχές της μοντελοποίησης, η στατιστική ανάλυση επιτρέπει ερμηνευτικές και αιτιολογημένες υποθέσεις χάρη στη χρήση των πιθανοτήτων και του στατιστικού συμπεράσματος. Αυτά τα δυο αποτελούν θεμελιώδη στοιχεία της στατιστικής ανάλυσης, γιατί επιτρέπουν στους βιοϊατρικούς ερευνητές να ποσοτικοποιούν με ακρίβεια υποθέσεις σχετικά με τα δεδομένα, όσον αφορά τις πιθανότητες ή τις συχνότητες εμφάνισης σε σχέση με το στατιστικό μοντέλο που χρησιμοποιούν. Αυτό θεωρείται κρίσιμο σε κλινικές εφαρμογές, όπου οι αποφάσεις που επηρεάζουν την υγεία των ασθενών πρέπει να δικαιολογούνται με ακριβείς, ορθολογικούς και εύχρηστους τρόπους για την εκπλήρωση των κοινών δεοντολογικών και νομικών απαιτήσεων.

Χρησιμοποιώντας τα εργαλεία των στατιστικών συμπερασμάτων, οι ερευνητές μπορούν να δοκιμάσουν πολλαπλά μοντέλα ή υποθέσεις και να προσδιορίσουν εκείνα που ερμηνεύουν καλύτερα τα δεδομένα. Με τη διαμόρφωση της διαδικασίας συμπερασμάτων όσον αφορά τις κατανομές πιθανοτήτων, μπορούν να ποσοτικοποιήσουν την αβεβαιότητα στα συμπεράσματά τους ως συνέπεια των στοχαστικών παραγόντων, όπως τα σφάλματα μέτρησης και τα ελλειπή δεδομένα.

Στην εποχή των μαζικών δεδομένων, αυτές οι πτυχές αποκτούν πρωταρχική σημασία, καθώς οι δυσκολίες που απαντώνται στον έλεγχο μεγάλου αριθμού υποθέσεων ή στη βελτιστοποίηση σύνθετων μοντέλων μπορούν να οδηγήσουν σε εσφαλμένη ερμηνεία των δεδομένων. Η στατιστική ανάλυση συμπληρώνει με αυτόν τον τρόπο τις προσεγγίσεις της μηχανικής εκμάθησης και της τεχνητής νοημοσύνης στην ανάλυση βιοϊατρικών δεδομένων με ουσιαστικούς τρόπους, όπως θα δούμε παρακάτω. Επιπρόσθετα, είναι κρίσιμο η επόμενη γενιά επιστημόνων να είναι εφοδιασμένη με ένα στιβαρό υπόβαθρο στη στατιστική ανάλυση ώστε να είναι ικανή να προβεί σε αξιόπιστες προβλέψεις από μεγάλα σύνολα βιοϊατρικών δεδομένων.

Επιστήμη Δεδομένων

Η επιστήμη των δεδομένων αναφέρεται γενικότερα στην ενσωμάτωση στατιστικών και υπολογιστικών τεχνικών, για την εξόρυξη γνώσεων από μαζικά δεδομένα. Ως ένας κλάδος που βασίζεται στα δεδομένα, η επιστήμη αυτή είναι σε θέση να απευθύνει προκαθορισμένες ερωτήσεις, καθώς και να διατυπώνει νέες υποθέσεις αμερόληπτα. Στην περίπτωση των βιοϊατρικών δεδομένων, η επιστήμη των δεδομένων μπορεί να εφαρμοστεί για απόκτηση νέων γνώσεων, βιολογικά χρηστικών ως προς τη βελτίωση της διάγνωσης, θεραπείας και πρόληψης μιας ασθένειας.

Η στατιστική, όπως περιγράφεται στην προηγούμενη ενότητα, αποτελεί θεμέλιο της επιστήμης των δεδομένων. Ωστόσο, η γνώση μόνο της στατιστικής θεωρίας δεν επαρκεί για την ανάλυση μεγάλων και πραγματικών συνόλων δεδομένων. Οι δεξιότητες πληροφορικής είναι ζωτικής σημασίας για την επιστήμη των δεδομένων, καθώς ο όγκος των συνόλων δεδομένων έχει αυξηθεί, κάτι το οποίο απαιτεί τη χρήση υπολογιστών για την αποτελεσματική αποθήκευση, αναζήτηση και ανάλυση τους. Η καλή γνώση των υπολογιστών που χρειάζεται για να είναι αποτελεσματική η επιστήμη των δεδομένων απαιτείται στην πράξη και όχι θεωρητικά.

Για το λόγο αυτό, ο όρος "γνώση των υπολογιστών" συχνά αντικαθίσταται από τον όρο "δεξιότητες χακαρίσματος" ή στα αγγλικά “hacking skills”. Είναι αλήθεια ότι για κάποιον ειδικευμένο στην επιστήμη των δεδομένων είναι απαραίτητο ένα υπόβαθρο στην επιστήμη των υπολογιστών με την προϋπόθεση να διαθέτει ικανότητες στον προγραμματισμό για να αντιληφθεί πλήρως και να αναλύσει τα δεδομένα. Υπάρχουν πολλές γλώσσες προγραμματισμού για ανάλυση δεδομένων αλλά οι επιστήμονες χρησιμοποιούν πιο συχνά τις Python και R. Στην περίπτωση των πιο μεγάλων συνόλων δεδομένων, χρησιμοποιούνται διάφορα άλλα προγράμματα όπου η χρήση της C είναι απαραίτητη, ενώ η αποθήκευση τους απαιτεί επιπλέον τη χρήση βάσεων δεδομένων όπως SQL.

Για να διασφαλίσει κάποιος ότι τα ερωτήματα που τίθενται σχετικά με τα δεδομένα είναι λογικά και οδηγούν στην ερμηνεία των αποτελεσμάτων πρέπει να διαθέτει μια εξειδίκευση στο πεδίο της επιστήμης δεδομένων. Οι επιστήμονες δεδομένων δεν αναλύουν παθητικά και αυτόματα τα δεδομένα. Αντίθετα, καλούνται να κάνουν διαλογή μεταβλητών κατά τη διάρκεια της διαδικασίας επιλογής ή / και μετατροπής αυτών, να επιλέξουν τις καταλληλότερες μεθόδους για την απάντηση συγκεκριμένων ερωτήσεων και κατά συνέπεια τον καλύτερο τρόπο επικοινωνίας και ερμηνείας των ευρημάτων χρησιμοποιώντας αποτελεσματικές τεχνικές απεικόνισης.

Η σύζευξη υψηλής τεχνικής κατάρτιση στη στατιστι-

κή, την επιστήμη των υπολογιστών, την πληροφορική παρέχει πρακτικές δεξιότητες που είναι απαραίτητες για την επιστήμη των δεδομένων. Σε αυτές περιλαμβάνονται οι γνώσεις μεθόδων ανάκτησης και επεξεργασίας δεδομένων, διεξαγωγής διερευνητικών αναλύσεων, δημιουργίας μοντέλων για την απάντηση σε επιστημονικά ερωτήματα και ενημερωτικών και οπτικά ελκυστικών παρουσιάσεων αποτελεσμάτων. Η διαδικασία της ανάλυσης δεδομένων δεν είναι γραμμική. Πολλοί επαναληπτικοί κύκλοι επεξεργασίας πρέπει γίνουν πριν ληφθούν τα "τελικά" αποτελέσματα. Δηλαδή, μετά τη διεξαγωγή διερευνητικών αναλύσεων ή κατασκευής αρχικών μοντέλων, ενδέχεται να χρειαστεί προσαρμογή των χαρακτηριστικών των δεδομένων ή / και μοντέλων που αξιοποιούνται σύμφωνα με το υπό εξέταση θέμα. Σε διάφορες περιπτώσεις, οι επιστήμονες δεδομένων αναπτύσσουν τα δικά τους εργαλεία και μεθόδους, προσαρμοσμένα στις ανάγκες που προκύπτουν κατά τις αναλύσεις πραγματικών συνόλων δεδομένων.

Για να διασφαλιστεί ότι τα αποτελέσματα της ερευνητικής διαδικασίας θα αποφέρουν το μέγιστο όφελος, πολλοί επιστήμονες δεδομένων έχουν αναπτύξει ερευνητικά πρωτόκολλα με υψηλή αναπαραγωγιμότητα. Αυτά περιλαμβάνουν τη δημιουργία πακέτων λογισμικού ανοιχτού κώδικα που διατίθενται δωρεάν, τις δημοσιεύσεις προτεινόμενων βημάτων για τη λήψη αποτελεσμάτων και την ανταλλαγή δεδομένων που είναι απαραίτητα για την αναπαραγωγή των ευρημάτων [19-22]. Με αυτό τον τρόπο διατίθενται διάφορες τεχνολογίες με ταυτόχρονη προώθηση τόσο της διαφάνειας όσο και της αναπαραγωγιμότητας των μεθόδων που εφαρμόζονται σε μεγάλα σύνολα δεδομένων. Το RStudio είναι ένα ολοκληρωμένο περιβάλλον ανάπτυξης που ενίσχυσε σημαντικά την χρηστικότητα και τη δημοτικότητα της γλώσσας προγραμματισμού R από τους επιστήμονες δεδομένων, κυρίως για λόγους που αφορούν στη βελτίωση της ροής εργασιών και στη διευκόλυνση της δημιουργίας εγγράφων R Markdown τα οποία μπορούν εύκολα να μετατραπούν σε διάφορες άλλες μορφές (π.χ. HTML, PDF) ώστε να παρουσιαστούν τα τελικά αποτελέσματα [22].

Μια άλλη χρήσιμη κατηγορία εφαρμογών είναι τα εργαστηριακά σημειωματάρια, όπως το Jupyter και το Apache Zepplin. Αυτά παρέχουν διαδραστικά περιβάλλοντα υπολογιστών που βασίζονται σε τεχνολογίες διαδικτύου και υποστηρίζουν τη χρήση λογισμικού ανοικτού κώδικα και γλωσσών προγραμματισμού, όπως οι Python, R, Scala, Groovy και SQL. Οι επιστήμονες δεδομένων μπορούν να εκμεταλλευτούν αυτά τα λογισμικά για την ανταλλαγή, την ανάλυση και την απεικόνιση των δεδομένων, κάτι που επιτρέπει την εύκολη συνεργασία μεταξύ επιστημόνων. Για παράδειγμα, ένας επιστήμονας δεδομένων που χρησιμοποιεί Python μπορεί να επεξεργαστεί τα δεδομένα με τη χρήση της βιβλιοθήκης «pandas», να αναλύσει δεδομένα με το scikit-learn και να απεικονίσει δεδομένα χρησιμοποιώντας τη βιβλιοθήκη Altair. Επίσης, τα παραπάνω λογισμικά υποστηρίζουν επεξεργασία μαζικών δεδομένων και τεχνολογίες υπολογιστών όπως οι Hadoop, Spark και Hive. Τα συστήματα παρακολούθησης των διαδοχικών εκδόσεων λογισμικών , όπως το Git, παρέχουν αποτελεσματικά μέσα για την διαχρονική παρακολούθηση της εξέλιξης μεγάλων έργων.

Γι' αυτό το σκοπό έχει αναπτυχθεί η GitHub, μια πλατφόρμα που φιλοξενεί έργα που χρησιμοποιούν το Git για έλεγχο της έκδοσης, η οποία έχει γίνει ένα ευρέως χρησιμοποιούμενο αποθετήριο όπου μπορεί κάποιος να μοιραστεί κώδικες, μικρά σύνολα δεδομένων και αποτελέσματα αναλύσεων. Πιο πρόσφατα, οι πλατφόρμες όπως το Docker και το Singularity παρέχουν ένα φιλικό προς το χρήστη μέσο διανομής κώδικα με προεγκατεστημένα εργαλεία λογισμικού και διαδικασίες ελεγχόμενες από τους χρήστες, οι οποίες μπορούν επίσης να βοηθήσουν στην αναπαραγωγιμότητα [23]. Παράλληλα έχει αναπτυχθεί εξειδικευμένο λογισμικό (π.χ. Kubernetes, OpenShift) το οποίο υποστηρίζει τη δημιουργία, τη διαχείριση και την ανάπτυξη επεκτάσιμων εφαρμογών σε κατανεμημένα συστήματα. Η αύξηση της πρόσβασης σε cloud computing έχει επιτρέψει πολλά από αυτά τα εργαλεία και τεχνολογικές προσεγγίσεις να αξιοποιηθούν και για μαζικά δεδομένα [24]. Η ακαδημαϊκή κοινότητα αλλά και η βιομηχανία αξιοποιούν αυτές τις τεχνολογίες για να στηρίξουν τις προσπάθειες της επιστήμης των δεδομένων που έχουν ως στόχο τη βελτίωση της υγείας.

Τεχνητή Νοημοσύνη

Βασικές Αρχές

Ο όρος «τεχνητή νοημοσύνη» (AI) έχει εξελιχθεί και έχει πλέον αποκτήσει ένα πιο γενικό, διεπιστημονικό και περιεκτικό νόημα σε σχέση με την πρώτη χρήση του. Ως μια υποκατηγορία της επιστήμης των υπολογιστών, η τεχνητή νοημοσύνη χρησιμοποιείται συχνά ως εναλλακτικός όρος της "μηχανικής εκμάθησης". Ορθότερα όμως η «μηχανική εκμάθηση» είναι ουσιαστικά ένα παρακλάδι της AI το οποίο ασχολείται με την ευρύτερη έννοια της επαγωγικής λογικής. Ωστόσο, ένα μεγάλο μέρος προαπαιτούμενων κομβικών ζητημάτων, που επικεντρώνονται στην αφαιρετική λογική,

ισχύουν στο μεγαλύτερο μέρος των εφαρμογών βιοϊατρικής πληροφορικής, που χρησιμοποιούνται σήμερα ενεργά.

Οι θεμελιώδεις αρχές της τεχνητής νοημοσύνης σε συνδυασμό με βιοϊατρικές εφαρμογές πληροφορικής [25] είναι απαραίτητες για όσους επιθυμούν να εκμεταλλευτούν πλήρως τα μαζικά δεδομένα για εξατομικευμένη ιατρική αλλά και για άλλες εφαρμογές στην υγειονομική περίθαλψη. Αυτές οι αρχές είναι πολύ χρήσιμες και για τη βαθύτερη κατανόηση της μηχανικής εκμάθησης αλλά και για το μέλλον της έρευνας στην AI. Συνοπτικά, οι βασικές αρχές της τεχνητής νοημοσύνης επικεντρώνονται στο πώς μπορούν να οργανώνονται, να παρουσιάζονται, να ερμηνεύονται, να διερευνώνται και να εφαρμόζονται βιοϊατρικά δεδομένα με σκοπό να αντλούνται γνώσεις, να λαμβάνονται αποφάσεις και, τελικά, να γίνονται προβλέψεις.

Η κατάρτιση θα πρέπει να ξεκινά με μια ιστορική ανασκόπηση της εξέλιξης της τεχνητής νοημοσύνης, με τον προσδιορισμό των ορισμών, με αναφορά σε σημαντικές εξελίξεις σε εφαρμογές και ηθικά ζητήματα [26]. Αυτό θα πρέπει να συνδυάζεται με θέματα που σχετίζονται με τη λογική (δηλ. Πρόταση και λογική πρώτης τάξης) που περιγράφουν την κοινή επίσημη γλώσσα για δεδομένα και γνώσεις που επιτρέπουν την διεπαφή μεταξύ προσώπου και μηχανής. Υπάρχουν συγκεκριμένες προδιαγραφές για την αναπαράσταση των δεδομένων, συμπεριλαμβανομένων των πλαισίων, των κανόνων, των δεντρογραμμάτων, των οντολογιών και των σημασιολογικών δικτύων ενώ η παρουσίαση είναι ένα καίριο θέμα που συνδέει τόσο την αφαιρετική όσο και την επαγωγική λογική.

Είναι επίσης σημαντικό να κατανοηθεί ο ρόλος ενός παράγοντα ως παραδοσιακού δομικού στοιχείου ενός συστήματος AI, το οποίο αντιλαμβάνεται το περιβάλλον του μέσω αισθητήρων και επενεργεί σε αυτό με ενεργοποιητές. Ένα άλλο κρίσιμο θέμα περιλαμβάνει την εισαγωγή στα βασικά της επίλυσης προβλημάτων μέσω αλγορίθμων αναζήτησης, συμπεριλαμβανομένων των ομοιόμορφων αναζητήσεων (π.χ. εύρους ή βάθους κατά προτεραιότητα) και της ευρετικής αναζήτησης (π.χ. άπληστη μέθοδος ή A * αναζήτηση). Η αναζήτηση είναι συναφής με τις συνήθεις προκλήσεις στην πρόσβαση στις βιοϊατρικές πληροφορίες και είναι απαραίτητη για τη βελτιστοποίηση των περιορισμών. Πολλές φορές, πρέπει να ικανοποιηθούν κάποιοι περιορισμοί ορισμένων μεταβλητών για να επιτευχθεί η λύση.

Η κατάρτιση στις βασικές αρχές της τεχνητής νοημοσύνης επεκτείνεται επίσης στην κατανόηση της αιτιολόγησης με αβεβαιότητα και του τρόπου που αυτή συνδέεται με την πιθανοτική βιοϊατρική γνώση. Αυτό οδηγεί σε άλλα υπολογιστικά θέματα των υπό προϋποθέσεις πιθανοτήτων, της εντροπίας, της συμπερασματολογίας κατά Bayes. Η ενσωμάτωσή τους σε συστήματα βασισμένα στη γνώση, συμπεριλαμβανομένων των συμπερασμάτων βάσει κανόνων, των εξειδικευμένων συστημάτων και των σύγχρονων συστημάτων υποστήριξης των κλινικών αποφάσεων είναι απόρροια των παραπάνω και έχει ιδιαίτερη σημασία για την υγειονομική περίθαλψη [25]. Υπάρχουν ακόμη αρκετά προηγμένα θέματα τεχνητής νοημοσύνης τα οποία θα μπορούσαν να βρουν εφαρμογή στο πεδίο της βιοϊατρικής πληροφορικής στο μέλλον.

Μηχανική εκμάθηση

Όπως προαναφέρθηκε, η μηχανική εκμάθηση αποτελεί υποπεδίο της τεχνητής νοημοσύνης που ασχολείται με την ευρύτερη έννοια της επαγωγικής λογικής. Πιο συγκεκριμένα, αποτελείται από ένα σύνολο μεθόδων που μπορούν να αναγνωρίζουν και εξάγουν μοτίβα από ακατέργαστα δεδομένα ώστε να τα χρησιμοποιήσουν για την πρόβλεψη μελλοντικών δεδομένων ή για τη λήψη αποφάσεων [25,27,28]. Χωρίζεται δε σε δυο μεγάλες κατηγορίες: την επιτηρούμενη και μη επιτηρούμενη εκμάθηση.

Στην επιτηρούμενη εκμάθηση, χρησιμοποιείται μια συνάρτηση που αντιστοιχίζει μια εισακτέα τιμή που διαθέτει ένα σύνολο χαρακτηριστικών, σε ένα αποτέλεσμα. Αν η τιμή του αποτελέσματος είναι κατηγορική τότε η παραπάνω διαδικασία ονομάζεται ταξινόμηση ή κατηγοριοποίηση ενώ αν είναι συνεχής ονομάζεται παλινδρόμηση.

Στη μη επιτηρούμενη ή περιγραφική εκμάθηση, το μόνο που είναι γνωστό είναι οι εισακτέες τιμές βάσει των οποίων εντοπίζονται ενδιαφέροντα μοτίβα όπως ομάδες, ανωμαλίες και λανθάνοντες παράγοντες. Η ανάλυση κατά ομάδες στοχεύει στην ομαδοποίηση παρόμοιων αντικειμένων σε ομάδες ενώ η ανίχνευση ανωμαλιών στοχεύει στον εντοπισμό των αποκλίσεων στα δεδομένα. Τέλος, η εύρεση των λανθανόντων παραγόντων είναι πολύ σημαντική αφού συμβάλλει στην εξαγωγή συμπαγών αναπαραστάσεων των δεδομένων ή άλλων ενημερωτικών χαρακτηριστικών. Δεδομένου ότι πολλά βιοϊατρικά προβλήματα μπορούν να διατυπωθούν ή να αποδοθούν μέσω των παραπάνω διεργασιών φαίνεται ότι η μηχανική εκμάθηση προσφέρει μια σειρά ισχυρών εργαλείων για την επίλυση των προβλημάτων της επιστήμης των δεδομένων στη βιοϊατρική.

Για το λόγο αυτό έχουν αναπτυχθεί πολλές επιτυχημένες εφαρμογές που αφορούν τη διάγνωση ασθενειών,

την ανακάλυψη βιολογικών δεικτών και φαρμάκων, τη μελέτη ομικών πεδίων, την πρόβλεψη κλινικών αποτελεσμάτων και την παρακολούθηση των ασθενών, την εξατομικευμένη θεραπεία αλλά κα την έγκαιρη πρόβλεψη επιδημιών. Επιπρόσθετα, αυτές οι εφαρμογές συμβάλλουν στη δημιουργία ηλεκτρονικών φακέλων υγείας, στην ανάπτυξη έξυπνων φορητών συσκευών που καταγράφουν την κατάσταση της υγείας του ατόμου ενώ καθοριστικό ρόλο διαδραματίζουν στην υποστήριξη της λήψης αποφάσεων στους τομείς της ακτινολογίας, δερματολογίας, οφθαλμολογίας και παθολογίας. Οι περισσότερες από αυτές στηρίζονται στις πιο γνωστές γλώσσες προγραμματισμού που χρησιμοποιούνται ευρέως σε εφαρμογές μηχανικής εκμάθησης, όπως Python, Java, R, C ++, C, JavaScript, Scala και Julia.

Μέθοδοι και αλγόριθμοι

Στον πεδίο της μηχανικής εκμάθησης αξιοποιείται ένα ευρύ φάσμα μεθόδων και αλγορίθμων το οποίο διαφοροποιείται ανάλογα με την κατηγορία εκμάθησης. Μια κλασσική μέθοδος της επιτηρούμενης εκμάθησης ονομάζεται εκμάθηση του δέντρου αποφάσεων. Σε αυτήν κάθε εσωτερικός κόμβος περιγράφει μια δοκιμή σε ένα χαρακτηριστικό, κάθε κλάδος αντιστοιχεί σε ένα αποτέλεσμα και κάθε κόμβος φύλλου αντιπροσωπεύει μια τελική τιμή. Αν και τα δέντρα αποφάσεων θεωρούνται μια καλή και εύκολα ερμηνεύσιμη μέθοδος ωστόσο αποτελούν εκτιμητές μεγάλων διακυμάνσεων, αφού ελαφρώς διαφορετικές εισακτέες τιμές μπορούν να οδηγήσουν σε πολύ διαφορετικές δομές δέντρων.

Για να ξεπεραστεί αυτή η αστάθεια, προτάθηκε η έννοια του τυχαίου δάσους το οποίο είναι πολύ πιο ισχυρό από το δέντρο αποφάσεων. Αυτό προκύπτει όταν συγκεντρώνονται πολλά δέντρα αποφάσεων βασισμένα σε τυχαία υποσύνολα δεδομένων και χαρακτηριστικών.

Άλλα παραδείγματα κλασσικών μεθόδων επιτηρούμενης μάθησης περιλαμβάνουν Μηχανές Διανυσμάτων Υποστήριξης (SVM), γραμμική παλινδρόμηση (linear regression), μοντέλο λογιστικής παλινδρόμησης/logit (logistic regression), αφελείς Μπαεσιανοί Ταξινομητές (naive Bayes classifiers), γραμμική διακρίνουσα ανάλυση (linear discriminant analysis, LDA) και αλγόριθμος των k-πλησιέστερων γειτόνων (k-nearest neighbors algorithm, k-NN). Σε αυτές τις κλασικές μεθόδους, τα χαρακτηριστικά που αντιπροσωπεύουν ένα αντικείμενο είναι σχεδιασμένα με το χέρι και δεν είναι βελτιστοποιημένα για τη μαθησιακή εργασία.

Το τελευταίο διάστημα η μηχανική εκμάθηση έχει αρχίσει πλέον να εφαρμόζεται όχι μόνο για την χαρτογράφηση αντικειμένων αλλά για την ανακάλυψη των αντικειμένων και την ίδια την αναπαράσταση τους. Ένα χαρακτηριστικό παράδειγμα αυτής της κατηγορίας είναι οι μέθοδοι εκμάθησης των νευρικών δικτύων [27] που έχουν αποδειχθεί εξαιρετικά επιτυχείς σε πολλούς τομείς εφαρμογών της μηχανικής εκμάθησης συμπεριλαμβανομένης της επιστήμης των βιοϊατρικών δεδομένων [29-33].

Προκλήσεις

Δεδομένης της πολύ μεγάλης κλίμακας και πολυπλοκότητας των μαζικών δεδομένων της βιοϊατρικής, η μηχανική εκμάθηση αντιμετωπίζει σημαντικές υπολογιστικές και μεθοδολογικές προκλήσεις, ώστε η τελική επιλογή μιας μεθόδου να είναι αρκετά δύσκολη και προβληματική. Αυτές είναι:

α. το ζήτημα της υπερεκπαίδευσης που προκύπτει όταν προσθέτουμε σε ένα μαθησιακό μοντέλο πολλές παραμέτρους που δεν μπορούν να δικαιολογηθούν από τα δεδομένα

β. η αδυναμία επιλογής του καταλληλότερου μοντέλου μεταξύ πολλών μοντέλων με διαφορετικές πολυπλοκότητες,

γ. η αναζήτηση της βέλτιστης στρατηγικής όταν δεν έχουμε μια κλειστού τύπου λύση,

δ. η βελτιστοποίηση των υπερπαραμέτρων όταν υπάρχουν πάρα πολλές παράμετροι

ε. η δυσκολία της βιοϊατρικής ερμηνείας των αποτελεσμάτων όταν προβλέπονται πολλά υποσχόμενα αποτελέσματα από πολύπλοκα μοντέλα

στ. η έλλειψη μιας μεθόδου μηχανικής εκμάθησης που να είναι κατάλληλη για όλους τους τύπους δεδομένων [34,35].

Για να αντιμετωπιστούν αυτές οι προκλήσεις και να καταστεί η μηχανική εκμάθηση πιο φιλική στους χρήστες, ειδικά στους μη επαγγελματίες, έχουν γίνει σοβαρές προσπάθειες στην ανάπτυξη του τομέα της αυτοματοποιημένης μηχανικής εκμάθησης (AutoML). Αυτός προβλέπει την αυτοματοποίηση της διαδικασίας εφαρμογής μηχανικής εκμάθησης σε πραγματικά προβλήματα. Τα υπάρχοντα συστήματα AutoML (π.χ. AutoWeka [36], AutoSklearn [37], TPOT [38] και PennAI [39]) έχουν σχεδιαστεί για την αυτοματοποίηση ενός ή περισσοτέρων βημάτων της διαδικασίας, όπως η προετοιμασία των δεδομένων, η αναζήτηση καθηκόντων, η τροποποίηση των χαρακτηριστικών, η επιλογή μοντέλου, η βελτιστοποίηση υπερπαραμέτρων, κ.ά.

Συμπερασματικά, γίνεται αντιληπτό ότι η μηχανική εκμάθηση είναι ένα αξιοσημείωτο και αρκετά αποτελεσματικό εργαλείο για τη διαχείριση μαζικών δεδομένων στον κλάδο της υγείας. Κατά συνέπεια, θα ήταν απαραί-

τητο να συμπεριληφθεί σε ένα πρόγραμμα σπουδών για την κατάρτιση της επόμενης γενιάς επιστημόνων στους κλάδους της βιοϊατρικής πληροφορικής και επιστήμης δεδομένων.

Επεξεργασία φυσικής γλώσσας και εξόρυξη κειμένου

Στόχοι

Κάθε χρόνο εκατοντάδες χιλιάδες νέες επιστημονικές δημοσιεύσεις και άρθρα εκδίδονται και αποθηκεύονται σε διάφορα αποθετήρια βιβλιογραφίας όπως η PubMed. Παράλληλα, εκατοντάδες χιλιάδες γραπτά κείμενα και ιατρικές σημειώσεις για ασθενείς καταγράφονται κάθε χρόνο και ενσωματώνονται στον ΗΦΥ. Όλα αυτά τα γραπτά αρχεία έχουν τεράστια αξία και σημασία αφού αποτελούν πρωτογενή δεδομένα. Συνεπώς, υπάρχει ανάγκη για ενσωμάτωση όλων αυτών σε μια βάση δεδομένων και εναρμόνισής τους με άλλα σετ δεδομένων.

Αυτό μπορεί να επιτευχθεί μέσω της επεξεργασίας φυσικής γλώσσας (NLP) που αποτελεί ένα ακόμη υποπεδίο της τεχνητής νοημοσύνης. Ο στόχος αυτής της λειτουργίας είναι να αυτοματοποιεί την επιμέλεια των εγγράφων τόσο από την επιστημονική βιβλιογραφία όσο και από τις κλινικές σημειώσεις για να δώσει τελικά μια κατανόηση του περιεχομένου τους. Με άλλα λόγια, η μέθοδος αυτή επιτρέπει την αυτόματη εξαγωγή λέξεων-κλειδιών και φράσεων από έγγραφα με σκοπό το σχολιασμό του νοήματός τους. Το περιεχόμενο που εξάγεται μπορεί στη συνέχεια να μετατραπεί σε δομημένα δεδομένα που μπορούν να ενσωματωθούν με άλλα δεδομένα είτε σε μια βάση δεδομένων.

Η αυτοματοποίηση της εξαγωγής δεδομένων από γραπτά κείμενα δεν είναι μια εύκολη διαδικασία αλλά είναι ένας εξελισσόμενος και πολλά υποσχόμενος ερευνητικός τομέας. Η δυσκολία του έγκειται στο γεγονός ότι ο υπολογιστής πρέπει να είναι σωστά "εκπαιδευμένος" στο να αναγνωρίζει, να καταγράφει και να αναδεικνύει τις σχέσεις σε λέξεις, όρους και ονόματα οντοτήτων των γραπτών κειμένων.

Όσον αφορά την αυτοματοποίηση στην αναγνώριση λέξεων και όρων, μια ενδιαφέρουσα εφαρμογή είναι η εύρεση και αναγνώριση των φαρμακευτικών σκευασμάτων που αναφέρονται σε μια σειρά κλινικών σημειώσεων. Στην συγκεκριμένη περίπτωση, ο υπολογιστής πρέπει να γνωρίζει τις διαφορετικές ονομασίες φαρμάκων, τις συντομογραφίες τους, τα ακρωνύμιά τους αλλά και την ανθρώπινη στενογραφία. Ένας ανθρώπινος εγκέφαλος μπορεί να το κάνει αυτό αλλά ένας υπολογιστής πρέπει να "εκπαιδευτεί" πάρα πολύ σκληρά για να το επιτύχει. Σύμφωνα με τον Hobbs, ο υπολογιστής μπορεί να επιτύχει σε ένα επίπεδο 60% τη σωστή και έγκυρη εξαγωγή δεδομένων χωρίς δυσκολία [40] αλλά η επίτευξη ενός μεγαλύτερου ποσοστού σε επίπεδο 90% απαιτεί την αναγνώριση σπάνιων όρων, συντμήσεων κλπ. Η βελτίωση του αλγορίθμου θα μπορούσε να αντιμετωπίσει το πρόβλημα αλλά αυτό χρειάζεται πολύ χρόνο και μελέτη.

- Στον πυρήνα της αναζήτησης μιας αυτοματοποιημένης διαδικασίας, υπάρχει μια μέθοδος γνωστή ως «αναγνώριση γνωστών οντοτήτων». Το συχνότερο πρόβλημα στην εφαρμογή της είναι η εύρεση αναφορών σε οντότητες όπως γονίδια, πρωτεΐνες, ασθένειες, φάρμακα, ονόματα οργανισμών, σε κείμενο φυσικής γλώσσας και οι αντιστοίχισή τους με τη θέση και τον τύπο τους. Η εύρεση και αντιστοίχιση των οντοτήτων είναι το δομικό χαρακτηριστικό για όλες τις εργασίες εξόρυξης κειμένου. Έχουν αναπτυχθεί διάφορα προγράμματα ανοιχτού κώδικα για το σκοπό αυτό, όπως το BANNER που είναι ένα σύστημα για την επισήμανση ονομάτων γονιδίων στη βιβλιογραφία [43].

Τα τελευταία χρόνια έχει παρατηρηθεί μια εξέλιξη στις μεθόδους αυτοματοποίησης για την εξαγωγή κειμένου στον τομέα της βιοϊατρικής αλλά οι προκλήσεις παραμένουν-. Αν και η ταχεία αλλαγή και ασυνέπεια των όρων και συντομογραφιών που χρησιμοποιούνται σε κείμενα βιοϊατρικής μειώνει σημαντικά την ακρίβεια της διαδικασίας, η συχνή χρήση μακροσκελών ονομάτων σε αυτά τα επιστημονικά κείμενα διευκολύνει τον υπολογιστή στο να προσδιορίσει αν υπάρχει όνομα οντότητας χωρίς απαραίτητα να καθορίσει τα ακριβή όριά του [41,42].

Τέλος, το επόμενο βήμα για την ολοκλήρωση της διαδικασίας εξαγωγής κειμένου, είναι η ανάδειξη των σχέσεων ανάμεσα σε δυο ή περισσότερες οντότητες. Έχουν σχεδιαστεί διάφορα συστήματα εξαγωγής σχέσεων γι' αυτό το σκοπό. Παράδειγμα αυτών αποτελεί το Pharmspresso το οποίο βρίσκει τις αναφορές σε γονίδια και φάρμακα αλλά και τη σχέση τους σε επιστημονικές δημοσιεύσεις [44]. Ο στόχος του είναι να βοηθήσει τον υπολογιστή να καταλάβει τη σχέση γονιδίου-φαρμάκου όπως τη θέτει ο συγγραφέας του άρθρου. Η πληροφορία που προκύπτει έχει τεράστια σημασία για τη βιοϊατρική έρευνα, διότι χρησιμεύει για τον επαναπροσδιορισμό φαρμάκων και την εύρεση νέων ενδείξεων σε φαρμακευτικά σκευάσματα που είναι ήδη σε κυκλοφορία [45].

Ρόλος και εφαρμογές

Η γεφύρωση της επεξεργασίας φυσικής γλώσσας

(NLP) με τις υπόλοιπες μεθόδους που περιγράφηκαν παραπάνω είναι το τελικό βασικό στοιχείο για την ενσωμάτωση δεδομένων σε ένα βιολογικό πλαίσιο. Η διαδικασία αυτή είναι γνωστή ως κανονικοποίηση και σχετίζεται με την δημιουργία μιας μοναδικής ονομασίας ταυτοποίησης μιας οντότητας [46] ώστε να αναγνωρίζεται άμεσα σε όλα τα είδη κειμένου. Ένα απλό παράδειγμα είναι η ονομασία των γονιδίων και η εύρεσή τους. Όταν ένα γονίδιο αναφέρεται σε ένα κείμενο, θα πρέπει να είναι γραμμένο με κατάλληλη ονομασία ώστε να μπορεί να αντιστοιχιστεί με το αναγνωριστικό που έχει στην πλατφόρμα Entrez Gene. Η κανονικοποίηση σαν διεργασία αποτελείται από τέσσερα βασικά βήματα. Το πρώτο είναι επιλογή ενός λεξικού στο οποίο θα αντιστοιχίζονται οι αναφορές του κειμένου. Το δεύτερο είναι εύρεση και η ταυτοποίηση των αναφορών του κειμένου που μας ενδιαφέρουν. Αυτό ίσως να περιλαμβάνει και τη διαχείριση των προθεμάτων, επιθεμάτων και λίστες των οντοτήτων. Το τρίτο είναι αντιστοίχιση των αναφορών του κειμένου με αυτές που περιλαμβάνει το λεξικό ενώ το τελευταίο περιλαμβάνει την μετα-επεξεργασία των αναφορών με σκοπό να απομακρυνθούν τα ψευδώς θετικά αποτελέσματα λόγω ασάφειας.

Αν και φαίνεται ότι τα βήματα είναι καλά μελετημένα και διορθώνουν τυχόν λάθη ή ασάφειες, στα βιοϊατρικά και κλινικά κείμενα απαιτούνται επιπρόσθετα μέτρα για την μείωση των ψευδώς θετικών αποτελεσμάτων. όπως για παράδειγμα η εφαρμογή μιας μεθόδου χαρακτηρισμού του περιεχομένου του κειμένου στο οποίο αναφέρεται η οντότητα. Αυτή η μέθοδος ονομάζεται ισχυρισμός και βασίζεται στην ιδέα ότι οι οντότητες και τα περιεχόμενά τους περιγράφονται με διαφορετικό τρόπο στα βιοϊατρικά και στα κλινικά κείμενα [48,49]. Φυσικά, είναι δυνατόν το περιεχόμενο μιας οντότητας να είναι γενικό ή ειδικό αναλόγως το κείμενο. Για παράδειγμα, σε ένα κλινικό κείμενο τα συμπτώματα αναφέρονται ως επιβεβαιωμένα ή μη.

Συνοψίζοντας, η αυτοματοποίηση της εξόρυξης κειμένου αλλά και της επεξεργασίας φυσικής γλώσσας παίζει καθοριστικό ρόλο στη σωστή ενσωμάτωση των στοιχείων σε βάσεις δεδομένων καθώς επίσης και στα συστήματα εξαγωγής πληροφοριών και επεξεργασίας βάσεων δεδομένων. Η ακριβής αναπαράσταση των οντοτήτων, οι σχέσεις τους και οι ισχυρισμοί τους μπορεί να έχουν κρίσιμες συνέπειες για την κατανόηση του τρόπου με τον οποίο εξηγούνται τα βιολογικά μονοπάτια για τα κλινικά προφίλ των ασθενών και τελικά για την παραγωγή γνώσης από πληθυσμιακές μελέτες [49].

Μελλοντικές προοπτικές

Η πρόοδος στην εξατομικευμένη ιατρική και την ιατρική ακριβείας εξαρτάται από την ικανότητά μας να καθορίσουμε τα μοναδικά χαρακτηριστικά των ατόμων ή μικρών ομάδων ατόμων όπου απαιτούνται συγκεκριμένες στρατηγικές πρόληψης και θεραπείας ασθενειών. Πριν εφαρμοστεί κάτι τέτοιο στην κλινική πρακτική, είναι αναγκαίο να καθοριστούν ποια είναι τα σημαντικά χαρακτηριστικά για κάθε ασθένεια. Για να γίνει αυτό εφικτό, είναι απαραίτητη η μέτρηση ενός μεγάλου αριθμού εσωτερικών και εξωτερικών βιολογικών διεργασιών.

Αυτό έχει σαν αποτέλεσμα, τη συγκέντρωση όλο και μεγαλύτερου όγκου βιοϊατρικών δεδομένων η οποία απαιτεί την ανάπτυξη και εφαρμογή ειδικών μεθόδων τεχνολογίας και πληροφορικής για την αποθήκευση, διαχείριση, ανάλυση και ερμηνεία τους.

Στην παρούσα εργασία έχουμε κάνει μια σύντομη ανασκόπηση στους βασικότερους επιστημονικούς τομείς που είναι απαραίτητο να έχει εκπαιδευτεί η νέα γενιά επιστημόνων ώστε να είναι εξοικειωμένοι με την εξατομικευμένη ιατρική, δίνοντας μεγαλύτερη έμφαση στην τεχνητή νοημοσύνη συμπεριλαμβανομένης της μηχανικής εκμάθησης και την επεξεργασία φυσικής γλώσσας στα μαζικά δεδομένα. Όλες αυτές οι υπολογιστικές μέθοδοι διαδραματίζουν ένα κομβικό ρόλο στην εξαγωγή χρήσιμων πληροφοριών από περίπλοκα μοτίβα σε μαζικά δεδομένα ενώ απαιτείται συμπληρωματική εκπαίδευση στη διαχείριση και ενσωμάτωση δεδομένων, στη στατιστική και στην επιστήμη των δεδομένων.

Για τους λόγους αυτούς προτείνεται η δημιουργία ενός προγράμματος σπουδών για την εκπαίδευση της νέας γενιάς επιστημόνων οι οποίοι προέρχονται είτε από τον κλάδο της υγείας είτε της πληροφορικής στα μαζικά βιοϊατρικά δεδομένα. Αυτό το πρόγραμμα θα απευθύνεται σε φοιτητές που επιθυμούν να αποκτήσουν μια περαιτέρω εξειδίκευση και θα περιλαμβάνει θεματικές ενότητες ή κύκλους κατάρτισης σχετικά με:

α. Τις διαδικασίες και τις μεθόδους διαχείρισης και ενσωμάτωσης δεδομένων σε βάσεις, καθώς και τις γνώσεις πλήρους χειρισμού των βάσεων αυτών.

β. Τις βασικές έννοιες και τις μεθόδους στατιστικής ανάλυσης και των πιθανοτήτων, που απαιτούνται για την εξαγωγή αποτελεσμάτων και συμπερασμάτων.

γ. Την επιστήμη των δεδομένων, συμπεριλαμβανομένου του προγραμματισμού ηλεκτρονικών υπολογιστών και των μεθόδων για τη βελτίωση της αναπαραγωγιμότητας

δ. Την εισαγωγή στις θεμελιώδεις αρχές της τεχνητής νοημοσύνης

ε. Τη σχέση μηχανικής εκμάθησης και τεχνητής νοημοσύνης για την επεξεργασία φυσικής γλώσσας σε μη δομημένα δεδομένα.

Τέλος, στις παραπάνω θεματικές ενότητες θα ήταν ωφέλιμο να συμπεριληφθεί και μια ενότητα η οποία θα αναφέρεται στην εξατομικευμένη ιατρική και στην ιατρική ακριβείας ώστε να δώσει κίνητρο για τη χρήση μαζικών δεδομένων στις βιοϊατρικές επιστήμες.

Οι παραπάνω θεματικές ενότητες θα μπορούσε να προσφέρονται ως επίσημα μαθήματα στο πλαίσιο ενός μεταπτυχιακού προγράμματος είτε δια ζώσης είτε ακόμη εξ αποστάσεως. Θα μπορούσε ακόμη να περιλαμβάνονται σε ένα εντατικό βραχυχρόνιο πρόγραμμα κατάρτισης (σεμινάριο) το οποίο θα προσέφερε μια θεμελιώδη κατανόηση των θεματικών ενοτήτων και στοιχειώδη πρακτική εμπειρία. Αδιαμφισβήτητα, η βέλτιστη πρόταση είναι η δημιουργία μεταπτυχιακού κύκλου μαθημάτων με δια ζώσης διδασκαλία σε φυσική αίθουσα, ώστε αφενός να μπορεί να καλυφθεί σε βάθος η ύλη σε όλες οι ενότητες αλλά και να υπάρχει άμεση αλληλεπίδραση φοιτητών και εκπαιδευτών.

Επιπλέον, για την πλήρη κάλυψη των θεματικών ενοτήτων και την καλύτερη κατάρτιση των φοιτητών, προτείνεται η δημιουργία ενός προγράμματος σπουδών που θα περιλαμβάνει μαθήματα που θα συμπληρώνει το ένα το άλλο, μέσω ορολογίας και παραδειγμάτων. Ακόμη η ύπαρξη μιας ολοκληρωμένης υπολογιστικής πλατφόρμας όπου οι φοιτητές θα μπορούν να εργαστούν με μαζικά δεδομένα, να πραγματοποιούν αναλύσεις, να αξιολογούν τα αποτελέσματα και να σχεδιάζουν εργαλεία υποστήριξης εικονικών κλινικών αποφάσεων για εξατομικευμένη φροντίδα θα ήταν ιδανική.

Όλα τα παραπάνω θα συνέβαλαν στην εμπέδωση της ύλης, στην εξοικείωση όλων των φοιτητών με τέτοιου τύπου δεδομένα αλλά και θα οδηγούσαν, πιθανώς, στο σχεδιασμό νέων εικονικών εργαλείων και προγραμμάτων για τη λήψη κλινικών αποφάσεων στην εξατομικευμένη ιατρική. Οι φοιτητές θα έχουν ελεύθερη πρόσβαση σε όλα τα λογισμικά προγράμματα που απαιτούνται όπως για παράδειγμα το OpenMRS [50], ένα πρόγραμμα διαχείρισης δεδομένων ασθενών. Οι εργασίες τους μπορεί να είναι διαθέσιμες αρχικά σε ένα τοπικό επίπεδο και στη συνέχεια να ενσωματωθούν σε μεγαλύτερες βάσεις δεδομένων συμβάλλοντας έτσι στην επιστημονική κοινότητα. Η πρόταση αυτή είναι αρκετά προχωρημένη και απαιτεί την αναθεώρηση πολλών ζητημάτων αλλά η υλοποίηση της δεν είναι ανέφικτη.

Οι συμμετέχοντες στο παραπάνω πρόγραμμα θα αποκτήσουν ένα επιπρόσθετο επαγγελματικό εφόδιο στον τομέα της βιοϊατρικής επιστήμης που θα τους δώσει ένα ανταγωνιστικό πλεονέκτημα στην αγορά εργασίας. Πριν από μερικά χρόνια οι ειδικοί των υπολογιστών, οι επιστήμονες δεδομένων, και οι στατιστικολόγοι συμμετείχαν ως σύμβουλοι ή συνεργάτες σε έργα όπου ήταν απαραίτητη η διαχείριση και ανάλυση δεδομένων με άμεση συμμετοχή στο σχεδιασμό της μελέτης.

Σήμερα, είναι ανάγκη να ισχυροποιηθεί η θέση πολλών νέων επιστημόνων οι οποίοι θα μπορούν να διατυπώνουν επιστημονικά ερωτήματα και να πραγματοποιούν τη δική τους έρευνα χωρίς να εμπλέκονται σ' αυτή τη διαδικασία άλλοι αναλυτές δεδομένων [51]. Ο συνδυασμός των σπουδών τους και των δεξιοτήτων που θα έχουν αναπτύξει κατά τη διάρκεια αυτών είναι ιδιαίτερα χρήσιμη στη σημερινή εποχή. Αυτή η νέα προσέγγιση στη βιοϊατρική έρευνα έχει ονομαστεί «σκέψη χωρίς όρια».

Συμπεράσματα

Ζούμε πλέον στην εποχή των μαζικών βιοϊατρικών δεδομένων. Κάθε μέρα ο αριθμός των βιοϊατρικών δεδομένων όλο και αυξάνεται αφού αποκτούμε νέες πληροφορίες για τον οργανισμό μας. Η συνεχής αύξηση των διαθέσιμων βιοϊατρικών δεδομένων αποτελεί μια κινητήρια δύναμη για την ενίσχυση του ρόλου και των εφαρμογών της εξατομικευμένης ιατρικής στην περίθαλψη. Για να εδραιωθεί όμως, αυτή η προσέγγιση περίθαλψης απαιτείται κατάλληλα εκπαιδευμένο και καταρτισμένο επιστημονικό και ερευνητικό προσωπικό.

Σε αυτήν την εργασία, προτείνεται η δημιουργία ενός προγράμματος κατάρτισης νέων επιστημόνων από τους κλάδους της πληροφορικής και των βιοϊατρικών επιστημών ώστε να αποκτήσουν κάποιες βασικές αλλά σημαντικές γνώσεις σε υπολογιστικές εφαρμογές που αξιοποιούνται στην εξαγωγή και διαχείριση μαζικών βιοϊατρικών δεδομένων. Καθώς οι κλάδοι των εφαρμοσμένων μαθηματικών και της επιστήμης των υπολογιστών συνεχίζουν να αλληλεπιδρούν, στο μέλλον αυτοί θα συγχωνευθούν σε ένα μόνο πεδίο, που ήδη μερικοί το αποκαλούν στατιστική μάθηση [18]. Είναι ευθύνη της ακαδημαϊκής κοινότητας να εκπαιδεύει και να προετοιμάζει τη νέα γενιά επιστημόνων οι οποίοι θα οδηγήσουν σ'αυτή την αλλαγή, διδάσκοντάς τους τη χρήση σύγχρονων εργαλείων και μεθόδων της στατιστικής ανάλυσης με τρόπους που να την ενθαρρύνουν. Η συμμετοχή σε ένα τέτοιο πρόγραμμα θα συνέβαλε στην επαγγελματική και ακαδημαϊκή ανέλιξη του νέου επιστήμονα ενώ θα εδραίωνε ακόμη περισσότερο το ρόλο της εξατομικευμένης ιατρικής.

Συνοψίζοντας, η χρήση βιοϊατρικών δεδομένων σε

συνδυασμό με την εκπαίδευση επιστημόνων πληροφορικής αλλά και επαγγελματιών υγείας, όπως περιεγράφηκε σε αυτήν την εργασία θα αποφέρει σημαντικά αποτελέσματα στην εξέλιξη και επίδραση της εξατομικευμένης ιατρικής στον τομέα της υγείας.

ABSTRACT

Personalized Medicine and big biomedical data

Margarita-Ioanna Koufaki[1], Ioannis G. Hatzis[2], George P. Patrinos[1]*

[1]University of Patras School of Health Sciences, Department of Pharmacy, Patras, Greece

[2]Public Institute of Vocational Training of Messolonghi, Messolonghi, Greece

Personalized medicine depends on our ability to measure and process biological and environmental information about patients so as to enact individually optimized treatment. Much of this data is being stored in electronic health records yielding big data that pose challenges for management and analysis. We review here several areas of knowledge that are necessary for next-generation scientists to fully realize the potential of biomedical big data. We begin with an overview of big data and its storage and management. We then review statistics and data science, the very foundations for artificial intelligence, machine learning, and natural language processing needed to develop predictive models for clinical decision making. We conclude with some specific training recommendations for preparing next-generation scientists for biomedical big data.

KEY WORDS: Personalized Medicine, biomedical big data, artificial intelligence, machine learning, electronic healthcare record

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Collins FS, Varmus H. A new initiative on precision medicine. N. Engl. J. Med. 372(9), 793–795 (2015).
2. Council NR. Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease [Internet]. Available from: https://www.nap.edu/catalog/13284/toward-precision-medicine-building-a-knowledge-network-for-biomedical-research.
3. Nimmesgern E, Benediktsson I, Norstedt I. Personalized Medicine in Europe. Clin Transl Sci. 10(2), 61–63 (2017).
4. Erlewyn-Lajeunesse M, Brathwaite N, Lucas JSA, Warner JO. Recommendations for the administration of influenza vaccine in children allergic to egg. BMJ. 339, b3680 (2009).
5. Tran T, Luo W, Phung D, et al. A framework for feature extraction from hospital medical data with applications in risk prediction. BMC Bioinformatics. 15(1), 425 (2014).
6. Albers DJ, Elhadad N, Tabak E, Perotte A, Hripcsak G. Dynamical Phenotyping: Using Temporal Analysis of Clinically Collected Physiologic Data to Stratify Populations. PLOS ONE. 9(6), e96443 (2014).
7. Westphal JF. Macrolide – induced clinically relevant drug interactions with cytochrome P-450A (CYP) 3A4: an update focused on clarithromycin, azithromycin and dirithromycin. British Journal of Clinical Pharmacology. 50(4), 285–295 (2000).
8. Phillips HS, Kharbanda S, Chen R, et al. Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. Cancer Cell. 9(3), 157–173 (2006).
9. Warner JL, Denny JC, Kreda DA, Alterovitz G. Seeing the forest through the trees: uncovering phe-

ΒΙΒΛΙΟΓΡΑΦΙΑ

nomic complexity through interactive network visualization. J Am Med Inform Assoc. 22(2), 324–329 (2015).

10. Huang X, Jennings SF, Bruce B, et al. Big data - a 21st century science Maginot Line? No-boundary thinking: shifting from the big data paradigm. BioData Min. 8, 7 (2015).
11. Angles R, Gutierrez C. Survey of Graph Database Models. ACM Comput. Surv. 40(1), 1:1–1:39 (2008).
12. Robinson I, Webber J, Eifrem E. Graph Databases. 1 edition. O'Reilly Media, Beijing ; Sebastopol, CA.
13. Harron K, Dibben C, Boyd J, et al. Challenges in administrative data linkage for research. Big Data Soc. 4(2), 205395171774567 8 (2017).
14. Mamun A-A, Aseltine R, Rajasekaran S. Efficient Record Linkage Algorithms Using Complete Linkage Clustering. PLoS ONE. 11(4), e0154446 (2016).
15. Kim D, Joung J-G, Sohn K-A, et al. Knowledge boosting: a graph-based integration approach with multi-omics data and genomic knowledge for cancer clinical outcome prediction. J Am Med Inform Assoc. 22(1), 109–120 (2015).
16. Beaulieu-Jones BK, Moore JH. MISSING DATA IMPUTATION IN THE ELECTRONIC HEALTH RECORD USING DEEPLY LEARNED AUTOENCODERS. Pac Symp Biocomput. 22, 207–218 (2017).
17. Kim D, Li R, Dudek SM, Ritchie MD. Predicting censored survival data based on the interactions between meta-dimensional omics data in breast cancer. J Biomed Inform. 56, 220–228 (2015).
18. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition [Internet]. 2nd ed. Springer-Verlag, New York Available from: //www.springer.com/us/book/9780387848570.
19. Greene CS, Garmire LX, Gilbert JA, Ritchie MD, Hunter LE. Celebrating parasites. Nat. Genet. 49(4), 483–484 (2017).
20. Huber W, Carey VJ, Gentleman R, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat. Methods. 12(2), 115–121 (2015).
21. Peng RD. Reproducible research in computational science. Science. 334(6060), 1226–1227 (2011).
22. Wickham H, Grolemund G. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. 1 edition. O'Reilly Media, Sebastopol, CA.
23. Beaulieu-Jones BK, Greene CS. Reproducibility of computational workflows is automated using continuous analysis. Nat. Biotechnol. 35(4), 342–346 (2017).
24. Cole BS, Moore JH. Eleven quick tips for architecting biomedical informatics workflows with cloud computing. PLoS Comput. Biol. 14(3), e1005994 (2018).
25. Russell S, Norvig P. Artificial Intelligence: A Modern Approach. 3rd ed. Prentice Hall Press, Upper Saddle River, NJ, USA.
26. Ertel W. Introduction to Artificial Intelligence [Internet]. 2nd ed. Springer International Publishing Available from: //www.springer.com/us/book/9783319584867.
27. Goodfellow I, Bengio Y, Courville A. Deep Learning. The MIT Press.
28. Murphy KP. Machine Learning: A Probabilistic Perspective. The MIT Press.
29. Baldi P. Deep Learning in Biomedical Data Science. Annu. Rev. Biomed. Data Sci. 1(1), 181–205 (2018).
30. Cao C, Liu F, Tan H, et al. Deep Learning and Its Applications in Biomedicine. Genomics Proteomics Bioinformatics. 16(1), 17–32 (2018).
31. Ching T, Himmelstein DS, Beaulieu-Jones BK, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface. 15(141) (2018).
32. Ravi D, Wong C, Deligianni F, et al. Deep Learning for Health Informatics. IEEE J Biomed Health Inform. 21(1), 4–21 (2017).
33. Wainberg M, Merico D, Delong A, Frey BJ. Deep learning in biomedicine. Nat. Biotechnol. 36(9), 829–838 (2018).
34. Olson RS, La Cava W, Orzechowski P, Urbanowicz RJ, Moore JH. PMLB: a large benchmark suite for machine learning evaluation and comparison. BioData Min. 10, 36 (2017).
35. Olson RS, Cava WL, Mustahsan Z, Varik A, Moore JH. Data-driven advice for applying machine learning to bioinformatics problems. Pac Symp Biocomput. 23, 192–203 (2018).
36. Nantasenamat C, Worachartcheewan A, Jamsak S, et al. AutoWeka: toward an automated data

ΒΙΒΛΙΟΓΡΑΦΙΑ

mining software for QSAR and QSPR studies. Methods Mol. Biol. 1260, 119–147 (2015).

37. Feurer M, Klein A, Eggensperger K, Springenberg J, Blum M, Hutter F. Efficient and Robust Automated Machine Learning [Internet]. In: Advances in Neural Information Processing Systems 28. Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R (Eds.). . Curran Associates, Inc., 2962–2970 (2015) [cited 2018 Mar 2]. Available from: http://papers.nips.cc/paper/5872-efficient-and-robust-automated-machine-learning.pdf.
38. Olson RS, Bartley N, Urbanowicz RJ, Moore JH. Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science [Internet]. In: Proceedings of the Genetic and Evolutionary Computation Conference 2016. ACM, New York, NY, USA, 485–492 (2016) [cited 2018 Mar 2]. Available from: http://doi.acm.org/10.1145/2908812.2908918.
39. Olson RS, Sipper M, Cava WL, et al. A System for Accessible Artificial Intelligence. In: Genetic Programming Theory and Practice XV. Banzhaf W, Olson RS, Tozier W, Riolo R (Eds.). . Springer International Publishing, 121–134 (2018).
40. Hobbs JR, Appelt D, Bear J, et al. FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text. arXiv:cmp-lg/9705013 [Internet]. (1997). Available from: http://arxiv.org/abs/cmp-lg/9705013.
41. Leser U, Hakenberg J. What makes a gene name? Named entity recognition in the biomedical literature. Brief. Bioinformatics. 6(4), 357–369 (2005).
42. Yeh A, Morgan A, Colosimo M, Hirschman L. BioCreAtIvE task 1A: gene mention finding evaluation. BMC Bioinformatics. 6 Suppl 1, S2 (2005).
43. Leaman R, Gonzalez G. BANNER: an executable survey of advances in biomedical named entity recognition. Pac Symp Biocomput. , 652–663 (2008).
44. Garten Y, Altman RB. Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. BMC Bioinformatics. 10 Suppl 2, S6 (2009).
45. Yang H-T, Ju J-H, Wong Y-T, Shmulevich I, Chiang J-H. Literature-based discovery of new candidates for drug repurposing. Brief. Bioinformatics. 18(3), 488–497 (2017).
46. Crim J, McDonald R, Pereira F. Automatically annotating documents with normalized gene lists. BMC Bioinformatics. 6 Suppl 1, S13 (2005).
47. Chen L, Liu H, Friedman C. Gene name ambiguity of eukaryotic nomenclatures. Bioinformatics. 21(2), 248–256 (2005).
48. Miwa M, Thompson P, McNaught J, Kell DB, Ananiadou S. Extracting semantically enriched events from biomedical literature. BMC Bioinformatics. 13, 108 (2012).
49. Velupillai S, Mowery D, South BR, Kvist M, Dalianis H. Recent Advances in Clinical Natural Language Processing in Support of Semantic Analysis. Yearb Med Inform. 10(1), 183–193 (2015).
50. Wolfe BA, Mamlin BW, Biondich PG, et al. The OpenMRS system: collaborating toward an open source EMR for developing countries. AMIA Annu Symp Proc. , 1146 (2006).
51. Huang X, Bruce B, Buchan A, et al. No-boundary thinking in bioinformatics research. BioData Min. 6(1), 19 (2013).