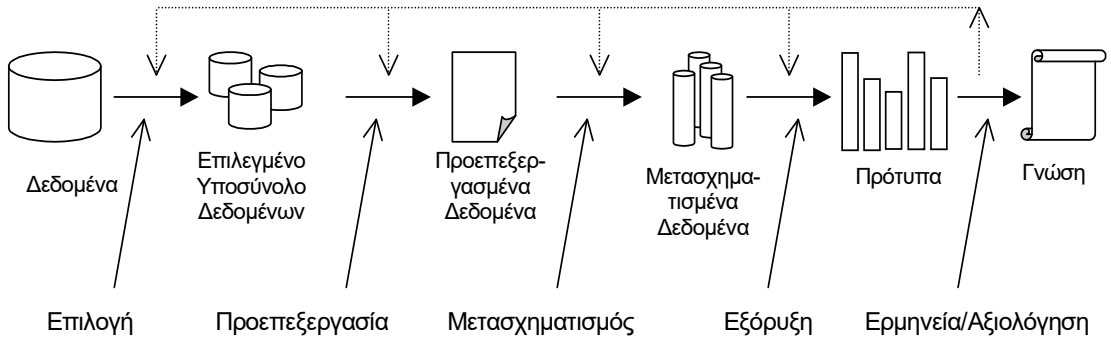
διαδικασία, στη διάρκεια των επιμέρους σταδίων της οποίας ο ειδικός σε θέματα ανακάλυψης γνώσης καλείται να πάρει συγκεκριμένες αποφάσεις.

Τυπικά μεγέθη του όγκου δεδομένων στα οποία εφαρμόζονται διαδικασίες ανακάλυψης γνώσης είναι χιλιάδες ή εκατοντάδες χιλιάδες δεδομένα (εγγραφές, συναλλαγές, κείμενα, κλπ) ή και εκατομμύρια (βιολογικά δεδομένα), και πολλές δεκάδες ΤΒ. Ενδεικτικά αναφέρεται ότι μια μεγάλη αλυσίδα super market μπορεί να καταγράφει χιλιάδες συναλλαγές κάθε μέρα, μια εταιρία κινητής τηλεφωνίας να καταγράφει εκατοντάδες χιλιάδες κλήσεις τη μέρα, ενώ μια εταιρία που αναζητά κοιτάσματα πετρελαίου να διαχειρίζεται εκατοντάδες ΤΒ δεδομένων.

20.1. Τα Στάδια της Ανακάλυψης Γνώσης

Τα επιμέρους στάδια στην ανακάλυψη ή εξόρυξη γνώσης απεικονίζονται στο Σχήμα 20.1 και περιγράφονται στη συνέχεια.



Σχήμα 20.1: Τα βασικά στάδια της διαδικασίας ανακάλυψης γνώσης.

Η ανακάλυψη γνώσης αρχίζει με την κατανόηση του τομέα στον οποίο θα εφαρμοστεί και τον προσδιορισμό του στόχου της από τη σκοπιά του χρήστη των αποτελεσμάτων. Ο ειδικός σε θέματα ανακάλυψης γνώσης πρέπει να συνεργαστεί με τον ειδικό του τομέα ώστε το πρόβλημα να καθοριστεί με αρκετή ακρίβεια, να είναι επιλύσιμο και τα αποτελέσματα να είναι μετρήσιμα και εφικτά (για παράδειγμα να προκύψουν σε αποδεκτά χρονικά όρια).

Πρέπει να σημειωθεί ότι τις περισσότερες φορές κάποια από τα επί μέρους βήματα είναι αναγκαίο να επαναληφθούν, καθώς στην πορεία ενδέχεται να προκύψουν προβλήματα που σχετίζονται με τις αρχικές επιλογές και τα οποία δεν ήταν δυνατό να εντοπιστούν αρχικά, όπως κακές επιλογές στην προεπεξεργασία των δεδομένων, κακή επιλογή παραμέτρων σε αλγόριθμους που επηρεάζουν την απόδοση του παραγόμενου μοντέλου ή προτύπου, κ.α.

20.1.1 Επιλογή

Στο *στάδιο της επιλογής* (*selection*) δημιουργείται το σύνολο δεδομένων στο οποίο θα εφαρμοστούν οι αλγόριθμοι ανακάλυψης γνώσης. Επειδή συχνά τα δεδομένα είναι οργανωμένα για άλλη χρήση και οι αλγόριθμοι που εκτελούν την ανακάλυψη γνώσης συνήθως δεν μπορούν να εφαρμοστούν σε πολλαπλούς πίνακες δεδομένων, απαιτείται

η εξαγωγή των δεδομένων από αυτούς και η οργάνωσή τους σε απλούστερες δομές. Συνήθως αυτή η απαίτηση καλύπτεται από τα *συστήματα αποθήκευσης δεδομένων* (*data warehouse*) τα οποία παρέχουν στους αλγόριθμους ανακάλυψης γνώσης μία ευκολότερα προσβάσιμη *όψη* (*view*) των δεδομένων.

Επίσης, σε ορισμένες περιπτώσεις όπου τα δεδομένα είναι πάρα πολλά και οι αλγόριθμοι ανακάλυψης γνώσης έχουν πρόβλημα με το χειρισμό του όγκου τους πρέπει να γίνει επιλογή ενός υποσυνόλου τους μέσω δειγματοληψίας.

20.1.2 Προεπεξεργασία

Στο στάδιο της *προεπεξεργασίας* (*preprocessing*) των δεδομένων αντιμετωπίζονται περιπτώσεις ελλιπών δεδομένων (για παράδειγμα, άδεια πεδία), πεδίων με τιμές που ουσιαστικά τα καθιστούν κενά (για παράδειγμα, Οδός = Άγνωστο), πεδίων με τιμές που υπονοούν (κατά σύμβαση) κάτι άλλο (για παράδειγμα, καταχώριση της ημερομηνίας "1/1/1970" ως προκαθορισμένης (default) τιμής σε πεδίο ημερομηνίας γέννησης που δεν συμπληρωνόταν, κτλ.) Λόγω της φύσης των εργασιών που πραγματοποιούνται, το στάδιο αυτό ονομάζεται και *στάδιο καθαρισμού* των δεδομένων (*data cleaning*). Είναι σημαντικό και συνήθως χρονοβόρο στάδιο που δύσκολα παραβλέπεται. Αν τα δεδομένα δεν είναι "καθαρά" θα επηρεαστεί σημαντικά η ποιότητα των αποτελεσμάτων.

Είναι χαρακτηριστικό ότι το καθάρισμα και η οργάνωση των δεδομένων είναι το πιο χρονοβόρο βήμα, καταναλώνοντας γύρω στο 60% του χρόνου της πλήρους διαδικασίας. Ταυτόχρονα είναι και το λιγότερο ευχάριστο στάδιο για τουλάχιστον 57% των εργαζόμενων σε τέτοιες διεργασίες[38].

20.1.3 Μετασχηματισμός

Στο στάδιο του *μετασχηματισμού* (*transformation*) τα δεδομένα μετασχηματίζονται ώστε να διευκολύνουν την ανακάλυψη γνώσης. Τέτοιοι μετασχηματισμοί μπορεί να περιλαμβάνουν:

- τη συνάθροιση και ομοιόμορφη κωδικοποίηση της ποιοτικά ίδιας πληροφορίας, όπως για παράδειγμα η ενοποίηση ενός πεδίου με τίτλο *salary* από έναν πίνακα με το πεδίο *payment* ενός άλλου πίνακα,
- τη μείωση του αριθμού των υπό εξέταση ανεξάρτητων μεταβλητών ή *χαρακτηριστικών* (*dimensionality reduction*) με επιλογή ορισμένων εξ' αυτών (*feature selection ή attribute selection*), που σχετίζονται ισχυρότερα με την εξαρτημένη μεταβλητή,
- τη *διακριτοποίηση* (*discretization*), δηλαδή τη μετατροπή συνεχών αριθμητικών τιμών σε διακριτές, επειδή για παράδειγμα έτσι το απαιτεί κάποιος αλγόριθμος είτε γιατί μας διευκολύνει στην κατανόηση,
- την κανονικοποίηση τους, δηλαδή τη μετατροπή των αριθμητικών δεδομένων σε ένα συγκεκριμένο και περιορισμένο εύρος τιμών, γιατί πιθανώς βοηθά κάποιον αλγόριθμο να "μάθει" καλύτερα,

[38] Από σχετική αναφορά του KDNuggets.com, 2015.

- τη δημιουργία νέων χαρακτηριστικών από τα ήδη υπάρχοντα, όπως για παράδειγμα την απομόνωση του ταχυδρομικού κώδικα από μια διεύθυνση.

Μείωση διαστάσεων

Ένα πρόβλημα πολλών διαστάσεων (δηλ. με πολλά χαρακτηριστικά) πιθανώς να είναι πιο εύκολο να επιλυθεί εάν αναχθεί σε ένα πρόβλημα λιγότερων διαστάσεων. Για παράδειγμα, μια λογιστική παρεμβολή μπορεί να αποδίδει καλύτερα χρησιμοποιώντας τις τρεις από τις εφτά διαθέσιμες διαστάσεις στα δεδομένα και επομένως αυτές οι τρεις διαστάσεις να είναι πιο "σημαντικές" για τον εκτιμητή από τις υπόλοιπες.

Η μείωση διαστάσεων (dimensionality reduction) μπορεί να γίνει με 2 τρόπους:

- *Επιλογή χαρακτηριστικών* (*feature selection*), επιλέγοντας ένα υποσύνολο χαρακτηριστικών, χωρίς μετασχηματισμό των δεδομένων. Υπάρχει εμφανής απώλεια πληροφορίας. Μια αριθμητική/στατιστική μέθοδος είναι η Lasso όπου οι συντελεστές/βάρη των χαρακτηριστικών μειώνονται και επομένως αυτοί που είναι κοντά στο μηδέν μπορούν να απαλειφθούν τελείως. Επειδή η επιλογή χαρακτηριστικών είναι μια σημαντική διαδικασία σε κάθε πρόβλημα ανακάλυψης γνώσης από δεδομένα, περιγράφεται αναλυτικά στην επόμενη υπο-ενότητα.
- *Προβολή χαρακτηριστικών* (*feature projection*), με μετασχηματισμό/προβολή υπαρχόντων δεδομένων σε μικρότερη διάσταση (π.χ. μέθοδος PCA). Υπάρχει ελάχιστη ή καθόλου απώλεια πληροφορίας.

Δημοφιλείς μέθοδοι μείωσης διαστάσεων με προβολή είναι οι ακόλουθες:

- *Ανάλυση πίνακα σε ιδιάζουσες τιμές - singular value decomposition* (*SVD*) που χρησιμοποιεί αλγεβρικές πράξεις πινάκων για να μειώσει το βαθμό (rank) του πίνακα δεδομένων.
- *Ανάλυση κύριων συνιστωσών - principal component analysis* (*PCA*) που μειώνει τις διαστάσεις των δεδομένων με προβολή τους σε έναν χώρο λιγότερων, αλλά διαφορετικών, διαστάσεων (δεν απαλείφει τα χαρακτηριστικά αλλά τα μετασχηματίζει). Τα μετασχηματισμένα χαρακτηριστικά είναι ανεξάρτητα μεταξύ τους. Οι μετασχηματισμοί όμως που εφαρμόζει η PCA συνεπάγονται λιγότερο ερμηνεύσιμα αποτελέσματα.
- *Γραμμική διακριτική ανάλυση - linear discriminant analysis* (*LDA*) που σε προβλήματα ταξινόμησης προβάλλει (μετασχηματίζει) τα δεδομένα σε έναν υποχώρο μικρότερων (λιγότερων από τις αρχικές) διαστάσεων, ο οποίος μεγιστοποιεί την απόσταση μεταξύ των κλάσεων και τις διαχωρίζει καλύτερα (έχει κάποια κοινά χαρακτηριστικά με την PCA).

Επιλογή χαρακτηριστικών

Η μάθηση ενός μοντέλου στη μάθηση με επίβλεψη προϋποθέτει ότι ορισμένες ή όλες οι ανεξάρτητες μεταβλητές ή χαρακτηριστικά (παράμετροι εισόδου) παρουσιάζουν στατιστική συσχέτιση με την εξαρτημένη μεταβλητή (παράμετρο εξόδου). Έτσι, όταν κατασκευάζεται ένα σύνολο εκπαίδευσης, υπάρχει η τάση να συμπεριλαμβάνονται όσο δυνατόν περισσότερα χαρακτηριστικά, γιατί αυτά είναι πιθανό να σχετίζονται με την