

ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

- ❖ Μετασχηματισμός των δεδομένων σε μορφή κατάλληλη και αποδοτική για την επιλεγμένη μέθοδο μάθησης.

- ❖ Καλύτερη μέθοδος: trial & error
 - Επιλογή Χαρακτηριστικών
Μεγάλος αριθμός χαρακτηριστικών, περισσότερα εκ των οποίων είναι περιττά
Διαδικασία: α) Οπτική διερεύνηση, β) Αυτόματα (select attributes tab)

 - Διακριτοποίηση Χαρακτηριστικών
Απαιτείται από κάποιους αλγόριθμους, και γενικότερα προκύπτουν πιο καλά αποτελέσματα και ταχύτερα
Πραγματοποιείται μέσω φίλτρων

 - Μετασχηματισμός Δεδομένων
(μαθηματικοί μετασχηματισμοί, λογικοί μετασχηματισμοί, αλλαγή δομής/μορφής δεδομένων)

 - Καθαρισμός Δεδομένων
Απλές μέθοδοι οπτικοποίησης ή αυτοματοποιημένες μέθοδοι εντοπισμού τιμών προς εξαίρεση και ανωμαλιών)

ΕΦΑΡΜΟΓΗ ΣΤΟ WEKA

Για την επιλογή μεταβλητών πηγαίνουμε στο select attributes tab κι έπειτα επιλέγουμε τη μέθοδο με την οποία θα πραγματοποιηθεί η διαδικασία.

Για τον μετασχηματισμό χαρακτηριστικών (attributes) και δεδομένων (instances) και τον καθαρισμό δεδομένων ακολουθούμε τη διαδικασία: Preprocess > Choose > Filters > Supervised/Unsupervised > Attribute/Instance > επιλέγουμε το φίλτρο που θέλουμε να εφαρμόσουμε > Apply .

Τα φίλτρα με επίβλεψη (supervised) δε συνιστούν προεπεξεργασία. Οι διαμερίσεις των δεδομένων ελέγχου απαιτείται να μη λαμβάνουν υπόψη τις τιμές των τάξεων των υποδειγμάτων ελέγχου, καθώς αυτές υποτίθεται πως δεν είναι γνωστές.

- **Μετασχηματισμός χαρακτηριστικών:**

- **Discretize** An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.
- **AddExpression** An instance filter that creates a new attribute by applying a mathematical expression to existing attributes.
- **NominalToBinary** Converts all nominal attributes into binary numeric attributes.
- **Normalize** Normalizes all numeric values in the given dataset (apart from the class attribute, if set)
- **ChangeDateFormat** Changes the date format used by a date attribute.
- etc

- **Μετασχηματισμός / Καθαρισμός δεδομένων:**

- **Randomize** Randomly shuffles the order of instances passed through it.
- **Resample** Produces a random subsample of a dataset using either sampling with replacement or without replacement.
- **RemoveMissclassified** A filter that removes instances which are incorrectly classified.