Ανισορροπία στα δεδομένα
Class Imbalance

Κάτια Κερμανίδου
kerman@ionio.gr

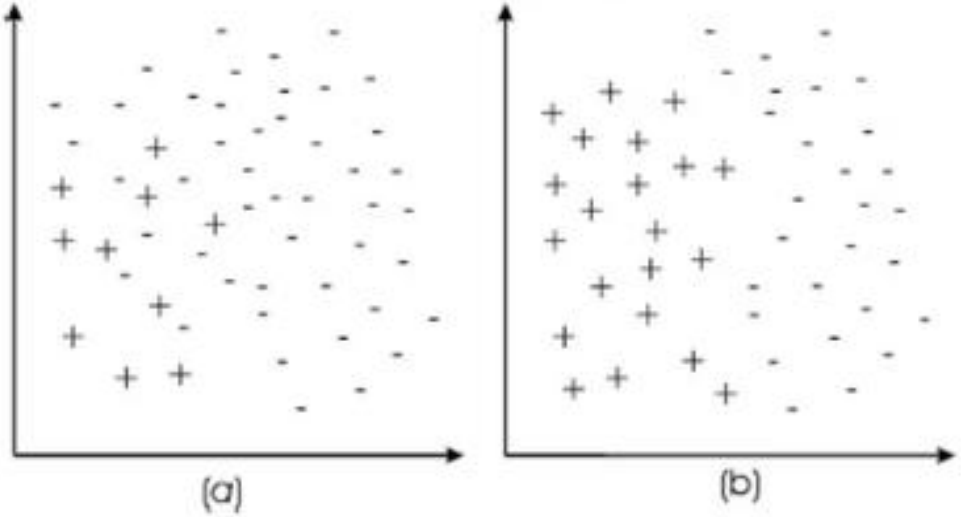
Class Imbalance

- Το πρόβλημα: Τα παραδείγματα εκπαίδευσης που ανήκουν σε μια τιμή της εξόδου είναι πολύ περισσότερα από τα παραδείγματα που ανήκουν στην άλλη τιμή.
- Οι περισσότεροι αλγόριθμοι θεωρούν ότι τα δεδομένα εκπαίδευσης είναι ισορροπημένα
- Αποτέλεσμα: η απόδοση των αλγορίθμων στην πρόβλεψη της τιμής που μειοψηφεί είναι χαμηλή

Γιατί είναι δύσκολη η μάθηση από ανισόρροπα σώματα εκπαίδευσης;

Τα σπάνια παραδείγματα της μειοψηφικής κλάσης μπορεί να παραπλανήσουν έναν ταξινομητή όπως ο k-Nearest Neighbor

Figure 1: Many negative cases against some spare positive cases (a) balanced data set with well-defined clusters (b).



https://www.slideserve.com/vlien/a-study-of-the-behavior-of-several-methods-for-balancing-machine-learning-training-data-powerpoint-ppt-presentation

Βαθμός ανισορροπίας (Degree of class imbalance)

- Imbalanced Ratio

- IR = Πλήθος παραδειγμάτων πλειοψηφικής κλάσης / Πλήθος παραδειγμάτων μειοψηφικής κλάσης

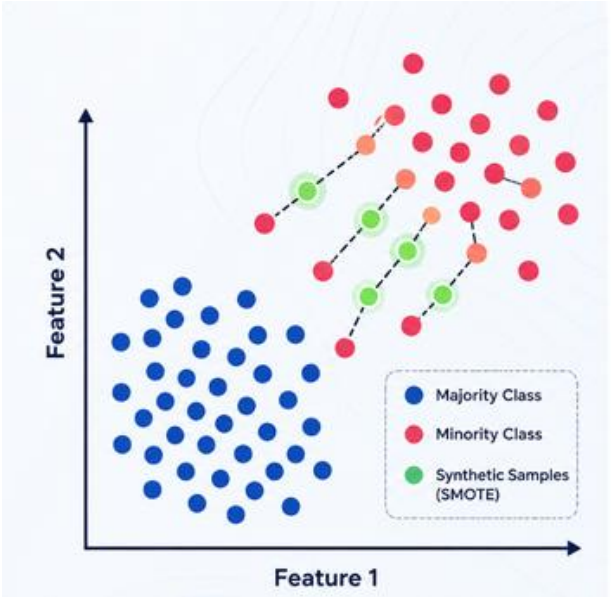
Λύση

- Εξισορρόπηση των παραδειγμάτων
- Απομάκρυνση των παραδειγμάτων τα οποία είναι θορυβώδη (noisy), δηλ. βρίσκονται στη λάθος πλευρά του ορίου απόφασης (decision border)
- Μέθοδοι
  - Υπερδειγματοληψία (Oversampling)
    - Τυχαία (Random) – τυχαία αντιγραφή παραδειγμάτων που μειοψηφούν
      - » Μπορεί να οδηγήσει σε υπερπροσαρμογή
    - SMOTE (Synthetic Minority Oversampling)
  - Υποδειγματοληψία (Undersampling)
    - Τυχαία (Random) – τυχαία διαγραφή παραδειγμάτων της κλάσης που πλειοψηφεί
      - » Μπορεί να οδηγήσει σε απώλεια παραδειγμάτων που είναι χρήσιμα για την διαδικασία της επαγωγής
    - Edited Nearest Neighbors (ENN)
    - Tomek Links
    - Condensed Nearest Neighbors (CNN)
    - One-sided sampling

Synthetic Minority Oversampling Technique (SMOTE)

N.V.Chawla, K.W.Bowyer, L.O.Hall and W.P.Kegelmeyer, "SMOTE: synthetic minority over-sampling technique". Journal of Artificial Intelligence Research, vol. 16, pp.321-357, 2002

- Η τεχνική SMOTE προσθέτει συνθετικά παραδείγματα της μειοψηφικής κλάσης
- Επιλέγεται τυχαία ένα παράδειγμα της μειοψηφικής κλάσης
- Βρίσκονται οι πχ 5 κοντινότεροι γείτονές του μειοψηφικής κλάσης.
- Επιλέγεται τυχαία ένας γείτονας από τους πέντε
- Δημιουργείται μια ευθεία ανάμεσα στο αρχικό παράδειγμα και τον επιλεγμένο γείτονα
- Πάνω στην ευθεία αυτή τοποθετείται σε τυχαίο σημείο ένα καινούριο παράδειγμα
- $X_{καινούριο} = x_{αρχικό} + δ * (x_{γείτονα} - x_{αρχικό})$, όπου δ ανήκει [0,1]



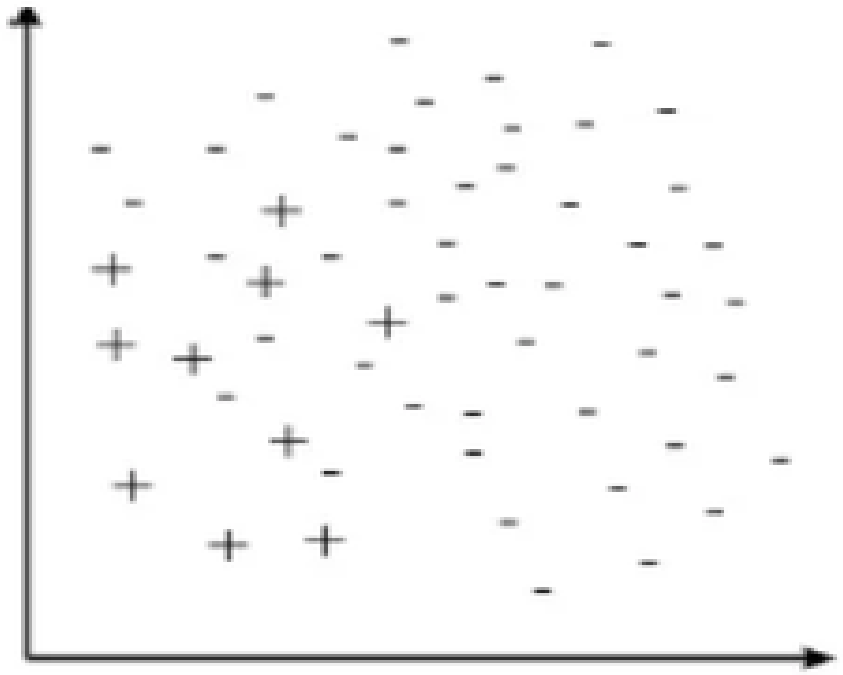
Edited Nearest Neighbors (ENN)

- Τα παραδείγματα που ανήκουν στην πλειοψηφική κλάση αφαιρούνται βάσει των k πλησιέστερων γειτόνων τους.
- Πχ για k=3, το παράδειγμα πλειοψηφικής κλάσης θα συγκριθεί με τα 3 πλησιέστερα γειτονικά του παραδείγματα.
- Εάν η πλειοψηφία των τριών γειτόνων ανήκει στην μειοψηφική κλάση, τότε το παράδειγμα πλειοψηφικής κλάσης αφαιρείται.

Tomek Links

Υπάρχουν τέσσερις ομάδες παραδειγμάτων της κλάσης που πλειοψηφεί

- Παραδείγματα θορύβου (Noise examples)
- Παραδείγματα στο όριο (Borderline examples)
- Πλεονάζοντα παραδείγματα (Redundant examples)
- Ασφαλή παραδείγματα (Safe examples)



Τα παραδείγματα που είναι στο όριο μπορεί πολύ εύκολα με την εισαγωγή μικρής ποσότητας θορύβου να πέσουν στην λάθος πλευρά του ορίου απόφασης, και να γίνουν θόρυβος.

Τα παραδείγματα θορύβου και τα παραδείγματα στο όριο πρέπει να διαγραφούν.
Πώς; Με **Tomek Links**.

Tomek Links (συνέχεια)

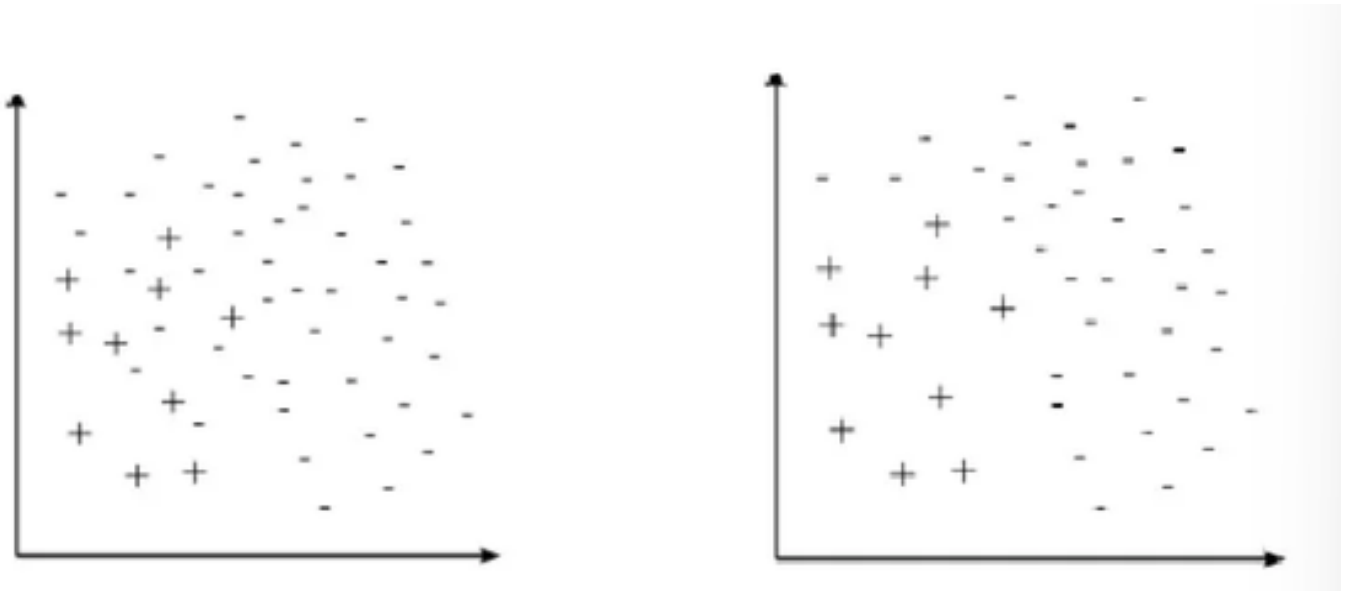
Έστω ότι τα παραδείγματα Ei και Ej ανήκουν σε διαφορετική τιμή της εξόδου, και d(Ei, Ej) η απόσταση μεταξύ τους.

Αν δεν υπάρχει κανένα άλλο παράδειγμα Ek, τέτοιο ώστε

- d(Ei, Ek) < d(Ei, Ej) ή
- d(Ej, Ek) < d(Ei, Ej)

τότε το ζεύγος παραδειγμάτων Ei και Ej ονομάζεται **Tomek Link.**

Το παράδειγμα το πλειοψηφικό του ζεύγους αφαιρείται.



Πώς αλλιώς μπορώ να διώξω παραδείγματα κοντά στο όριο; Ο Condensed Nearest Neighbor Rule

- Βρες ένα συνεπές (consistent) υποσύνολο Ε' του αρχικού σετ εκπαίδευσης Ε
  - Το Ε' είναι συνεπές υποσύνολο του Ε εάν, χρησιμοποιώντας τον 1-πλησιέστερο γείτονα (1-NN), το Ε' ταξινομεί σωστά όλα τα παραδείγματα του Ε.
- Αλγόριθμος
  - Σχημάτισε ένα υποσύνολο Esub του Ε που να περιέχει όλα τα παραδείγματα της μειοψηφικής κλάσης, και ένα τυχαίο παράδειγμα της πλειοψηφικής
  - Ταξινόμησε με το Esub χρησιμοποιώντας τον 1-NN όλα τα παραδείγματα του Ε
  - Μετάφερε στο Esub όλα τα λάθος ταξινομημένα παραδείγματα
  - Επανάλαβε τα δυο προηγούμενα βήματα μέχρι να μην υπάρχουν λάθος ταξινομήσεις
  - Το τελικό Esub είναι το πιο συνεκτικό σώμα εκπαίδευσης