Ανίχνευση ανωμαλιών

Ιωάννης Μαδεμλής

Ανίχνευση ανωμαλιών

- **Ανίχνευση ανωμαλιών** λέγεται ο αυτόματος εντοπισμός ανωμαλιών (outliers).
  - Πρότυπα τα οποία διαφέρουν σημαντικά από τη μεγάλη πλειονότητα των υπολοίπων.
- Ορισμένες φορές οι ανωμαλίες είναι ενδιαφέρουσες και η εύρεσή τους είναι αυτοσκοπός. Π.χ.:
  - Ανίχνευση εισβολέων σε δίκτυα.
  - Ανίχνευση απάτης.
- Συνήθως, όμως, η ανίχνευση ανωμαλιών είναι στάδιο της προεπεξεργασίας των δεδομένων.
  - Αναγνωρίζουμε τις ανωμαλίες και τις απομακρύνουμε από το σύνολο δεδομένων, πριν εφαρμόσουμε κάποιον αλγόριθμο εξόρυξης γνώσης.
  - Ιδιαιτέρως σημαντικό όταν χρησιμοποιούμε αλγορίθμους ή μέτρα εγγύτητας χωρίς ανθεκτικότητα σε ανωμαλίες.

Ανίχνευση ανωμαλιών

Ανίχνευση ανωμαλιών

Αιτίες εμφάνισης ανωμαλιών

- Οι ανωμαλίες μπορεί να έχουν προκύψει για διαφορετικούς λόγους:
  - Ενδεχομένως να ανήκουν σε μία διαφορετική, άγνωστη κλάση, σε σχέση με τα υπόλοιπα πρότυπα.
  - Ίσως απλώς εμπίπτουν σε μία περιοχή του δειγματικού χώρου της υποκείμενης γεννήτριας κατανομής των δεδομένων μας με πολύ μικρή, αλλά όχι μηδενική πυκνότητα πιθανότητας.
  - Μπορεί να προκύπτουν από θόρυβο ή σφάλματα στις διαδικασίες μέτρησης ή/και συλλογής των δεδομένων.

Προσεγγίσεις ανίχνευσης ανωμαλιών

Ανίχνευση ανωμαλιών

- Τρεις είναι οι βασικές προσεγγίσεις ανίχνευσης ανωμαλιών:
  - **Μοντέλα δεδομένων**: Κατασκευάζουμε ένα μοντέλο του συνόλου δεδομένων (π.χ., μία ομαδοποίηση, μία παραμετρική εκτίμηση της γεννήτριας κατανομής, κλπ.) και εντοπίζουμε τα πρότυπα τα οποία δεν εξηγούνται καλά από το μοντέλο.
  - **Εγγύτητα**: Υπολογίζουμε έναν πίνακα εγγύτητας μεταξύ όλων των προτύπων και εντοπίζουμε τα δεδομένα τα οποία είναι περισσότερο απομακρυσμένα από τη μεγάλη πλειονότητα των προτύπων.
  - **Πυκνότητα**: Μετρούμε την πυκνότητα των προτύπων σε κάθε περιοχή του διανυσματικού τους χώρου και εντοπίζουμε ποια πρότυπα είναι σε περιοχή με μικρότερη τοπική πυκνότητα από τους περισσότερους γείτονές τους.
- Απαιτείται η επιλογή κάποιου μέτρου εγγύτητας ή/και κάποιες τιμές κατωφλίων (υπερπαράμετροι).

Προσεγγίσεις ανίχνευσης ανωμαλιών

- Μία εναλλακτική διάκριση είναι μεταξύ των *επιβλεπόμενων* και των *ανεπίβλεπτων* μεθόδων ανίχνευσης ανωμαλιών.

- Στην περίπτωση επιβλεπόμενων μεθόδων, έχουμε στη διάθεσή μας ετικέτες οι οποίες διακρίνουν μία κλάση *κανονικών* και μία κλάση *ανώμαλων* προτύπων.
  - Συνήθως λύνουμε το πρόβλημα εκπαιδεύοντας έναν δυαδικό ταξινομητή.
  - Πρόκειται για ειδική περίπτωση μεθόδου βασισμένης σε μοντέλο δεδομένων (υπερεπιφάνεια απόφασης).
  - Η κλάση των ανωμαλιών έχει εξ ορισμού πολύ λιγότερα πρότυπα από την κανονική κλάση.

- Κάθε μέθοδος η οποία δεν αξιοποιεί ετικέτες κλάσης, είναι ανεπίβλεπτη.

Ανίχνευση ανωμαλιών

Προβλήματα ανίχνευσης ανωμαλιών

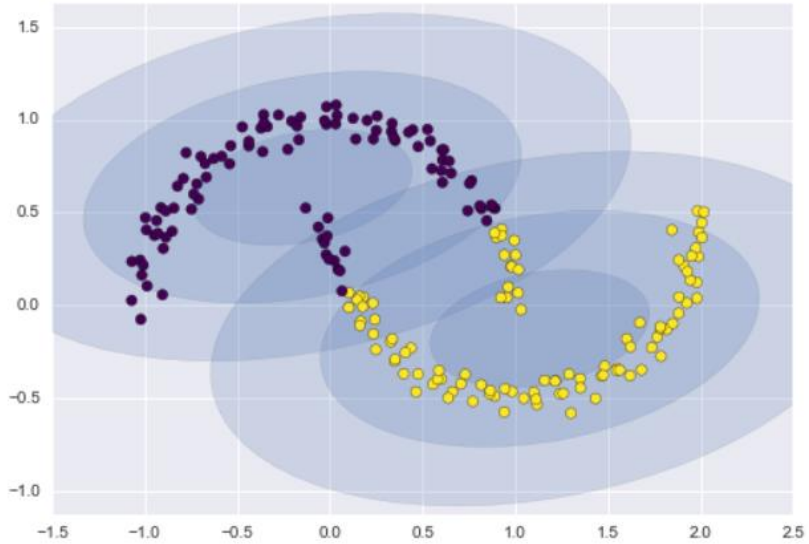
- Συνηθισμένα ζητήματα στην ανίχνευση ανωμαλιών είναι τα παρακάτω:
  - Κάποια πρότυπα είναι ανώμαλα μόνο όσον αφορά ορισμένα γνωρίσματά τους και όχι άλλα.
  - Κάποια πρότυπα μπορεί να μοιάζουν ανώμαλα σε σχέση με το σύνολο των δεδομένων, αλλά όχι σε σχέση με τα γειτονικά τους.
  - Η κατάσταση της ανωμαλίας είναι ένα συνεχές φάσμα (πολύ ανώμαλα έως καθόλου ανώμαλα πρότυπα), οπότε ίσως απαιτείται η απόδοση ενός σκορ ανωμαλίας.
  - Στις ανεπίβλεπτες μεθόδους μοντέλου δεδομένων, οι ανωμαλίες παραμορφώνουν το ίδιο το μοντέλο.
  - Η αξιολόγηση των αλγορίθμων σε ένα σύνολο δεδομένων δεν είναι πάντα εύκολη.
    - Στην περίπτωση επιβλεπόμενης ανίχνευσης, μπορούμε να χρησιμοποιήσουμε μετρικές αξιολόγησης ταξινομητών.
    - Στην περίπτωση ανεπίβλεπτων μεθόδων μοντέλου δεδομένων, μπορούμε να εκτιμήσουμε την ποιότητα του μοντέλου πριν και μετά την αφαίρεση των ανιχνευθέντων ανωμαλιών (π.χ., πιθανοφάνεια κατανομής).

Ανεπίβλεπτες μέθοδοι στατιστικών μοντέλων

- Η συνηθέστερη οικογένεια ανεπίβλεπτων μεθόδων ανίχνευσης ανωμαλιών είναι τα στατιστικά μοντέλα.
  - Συνήθως, εκτιμούμε την υποκείμενη γεννήτρια κατανομή των προτύπων.
  - Θεωρούμε κάποιο παραμετρικό στατιστικό μοντέλο (π.χ., γκαουσιανή κατανομή) και εκτιμούμε τις παραμέτρους του, έτσι ώστε να μεγιστοποιείται η πιθανοφάνεια.
  - Όποιο από τα πρότυπα έχει μικρή πυκνότητα πιθανότητας με βάση το προκύπτον μοντέλο, θεωρείται ανωμαλία.
  - Άρα, η πυκνότητα πιθανότητας λειτουργεί ως σκορ ανωμαλίας το οποίο μπορούμε να κατωφλιώσουμε σε ένα κατώφλι-υπερπαράμετρο.
    - Για γκαουσιανό μοντέλο, η απόσταση Mahalanobis από τον μέσο της κατανομής είναι μία εναλλακτική τιμή προς κατωφλίωση.
  - Γεννήτριες κατανομές αυθαίρετου σχήματος (στον χώρο των προτύπων) μπορούν να προσεγγιστούν με Γκαουσιανές Μεικτές Κατανομές (Gaussian Mixture Models, GMMs).
    - Απαιτείται να οριστεί ως υπερπαράμετρος το μέγεθος της μείξης (το πλήθος k των επιμέρους κατανομών).

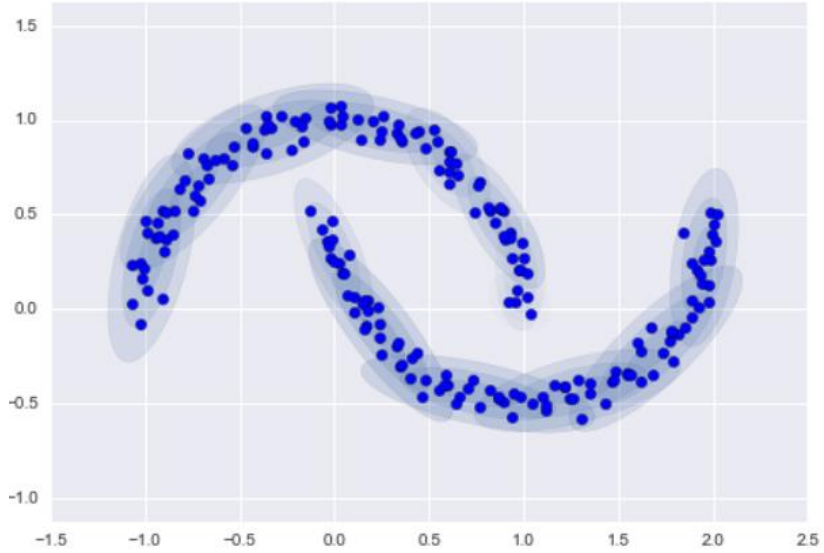
Ανεπίβλεπτες μέθοδοι στατιστικών μοντέλων

Ανίχνευση ανωμαλιών



Ομαδοποίηση μη ελλειψοειδών διδιάστατων δεδομένων με GMM και $k = 2$.

Ομαδοποίηση μη ελλειψοειδών διδιάστατων δεδομένων με GMM και $k = 16$.



Πηγή: J. VanderPlas, "*Python Data Science Handbook*", 2017.

Ανίχνευση ανωμαλιών

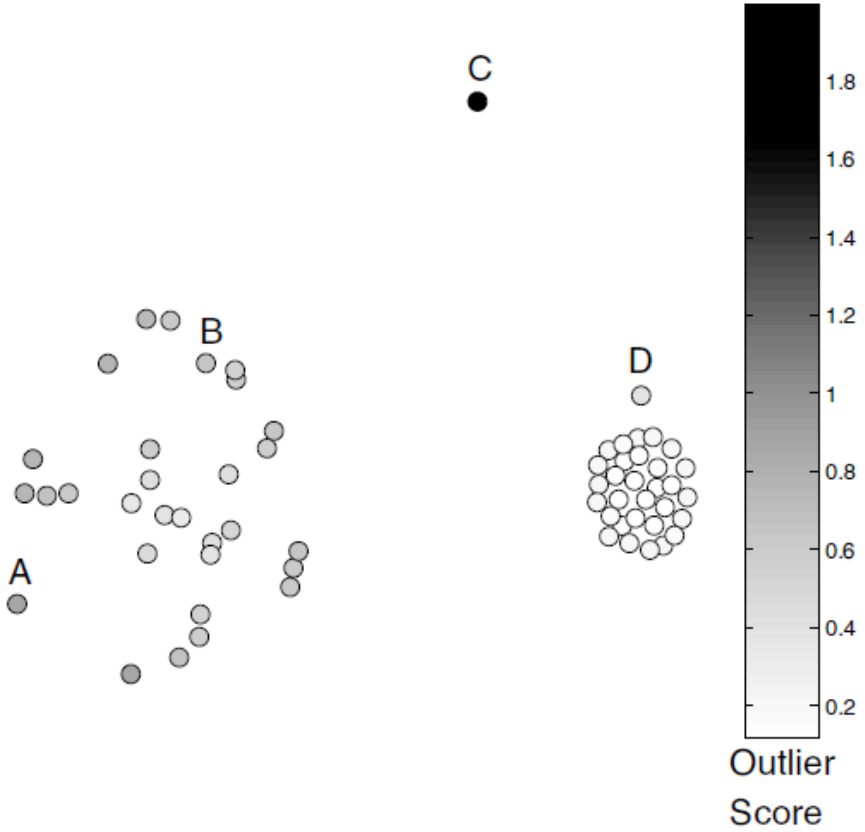
Μέθοδοι εγγύτητας

- Η συνηθέστερη μέθοδος εγγύτητας είναι ο υπολογισμός της απόστασης κάθε προτύπου από τον k-οστό εγγύτερό του γείτονα.
  - Το k είναι υπερπαράμετρος με τιμή ορισμένη από εμάς.
  - Κατωφλιώνουμε αυτή την απόσταση, ξεχωριστά για κάθε πρότυπο, σε ένα κατώφλι-υπερπαράμετρο.
  - Θεωρούμε ανωμαλία όποιο έχει απόσταση μεγαλύτερη από το κατώφλι.
- Ένα πολύ μικρό k ίσως οδηγήσει στη θεώρηση ορισμένων ανωμαλιών γειτονικών μεταξύ τους ως κανονικών δεδομένων.
- Ένα πολύ μεγάλο k ίσως οδηγήσει στη θεώρηση ως ανωμαλιών των στοιχείων μίας ομάδας με λίγα (λιγότερα από k) κανονικά πρότυπα.
  - Το ορθό k εξαρτάται από τα δεδομένα.
  - Η μέθοδος είναι προβληματική όταν υπάρχουν ομάδες διαφορετικής πυκνότητας.

Μέθοδοι εγγύτητας

Ανίχνευση ανωμαλιών

Παράδειγμα με διδιάστατα δεδομένα. Το πρότυπο C θα ανιχνευθεί εύκολα και ορθά ως ανωμαλία. Το D όμως δεν θα ανιχνευθεί, εκτός και αν τεθεί ένα κατώφλι στο σκορ ανωμαλίας το οποίο θα οδηγήσει σε εσφαλμένη σήμανση ως ανωμαλιών και των A, B.



Πηγή: Tan et al., *Introduction to Data Mining, 2006.*

Ανίχνευση ανωμαλιών

Μέθοδοι πυκνότητας

- Οι μέθοδοι πυκνότητας συνήθως αναθέτουν σε κάθε πρότυπο ένα σκορ ανωμαλίας, υπολογισμένο ως τον αντίστροφο της τοπικής πυκνότητας προτύπων.
  - Έτσι, ανωμαλίες θεωρούνται όσα πρότυπα τοποθετούνται σε περιοχή μικρής πυκνότητας.
  - Το σκορ μπορεί να κατωφλιωθεί σε ένα κατώφλι-υπερπαράμετρο.
- Η πυκνότητα εκφράζει το πλήθος προτύπων ανά μονάδα όγκου στον χώρο των προτύπων.
  - Πώς μπορεί όμως να μετρηθεί με ακρίβεια και ξεχωριστά στην τοπική περιοχή κάθε προτύπου;
- Ένας τρόπος είναι να ποσοτικοποιηθεί η τοπική πυκνότητα της περιοχής ενός προτύπου $\mathbf{x}_i$ ως ο αντίστροφος της μέσης απόστασης του $\mathbf{x}_i$ από τους k εγγύτερους γείτονές του.
  - Πρέπει φυσικά να επιλεχθεί κατάλληλο μέτρο απόστασης και κατάλληλο k.

Μέθοδοι πυκνότητας

- Με τον βασικό ορισμό της πυκνότητας εμφανίζεται ξανά μία αδυναμία χειρισμού περιοχών διαφορετικής πυκνότητας προτύπων.
  - Απαιτούν διαφορετικές τιμές υπερπαραμέτρων για να τις χειριστεί ορθά ο αλγόριθμος.

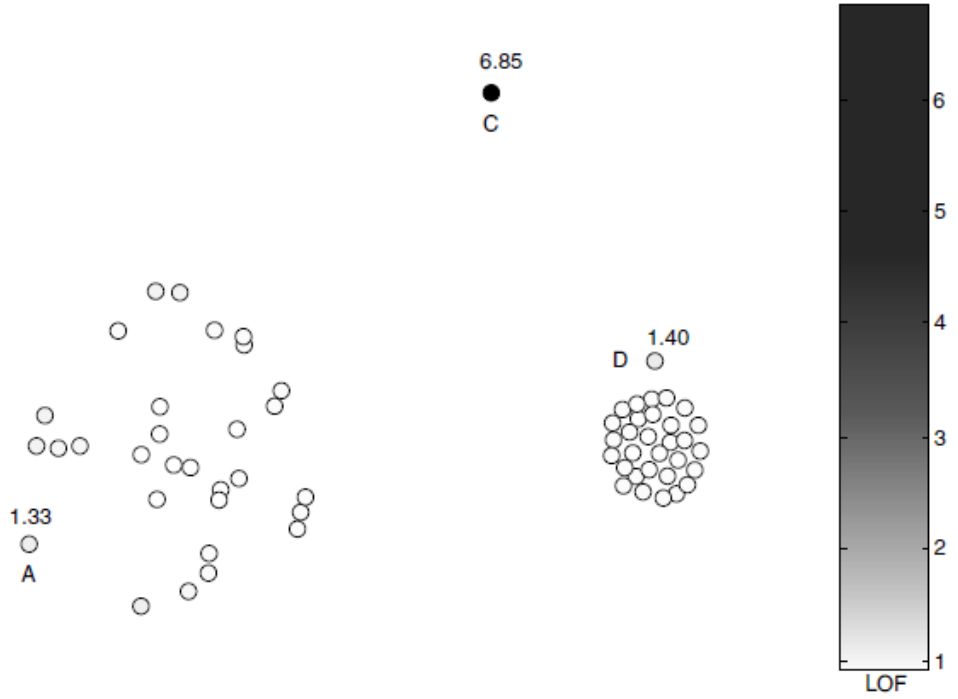
- Μία λύση είναι να τροποποιήσουμε την υπολογισμένη τοπική πυκνότητα ενός προτύπου $\mathbf{x}_i$, ώστε να λαμβάνει υπόψη και την πυκνότητα των γειτόνων του.
  - Διαιρούμε την προκαταρκτική τοπική πυκνότητα του $\mathbf{x}_i$ με τη μέση πυκνότητα των k εγγύτερων γειτόνων του.
  - **Σχετική πυκνότητα**.

- Έτσι η προκαταρκτική τοπική πυκνότητα του $\mathbf{x}_i$ υφίσταται αυτόματη διόρθωση:
  - Αυξάνεται σε αραιές περιοχές.
  - Μειώνεται σε πυκνές περιοχές.

Μέθοδοι εγγύτητας

Ανίχνευση ανωμαλιών

Με χρήση σχετικής πυκνότητας (LOF) τα σκορ ανωμαλίας είναι τώρα εύλογα για τα πρότυπα A, C και D. Ένα μόνο κοινό κατώφλι αρκεί για να σημανθούν ορθώς ως ανωμαλίες και το C και το D, χωρίς να θεωρηθεί εσφαλμένα ως ανωμαλία και το A.



Πηγή: Tan et al., *Introduction to Data Mining, 2006.*

Μέθοδοι ομαδοποίησης

Ανίχνευση ανωμαλιών

- Αλγόριθμοι ομαδοποίησης μπορούν επίσης να αξιοποιηθούν για ανίχνευση ανωμαλιών.
  - Πρόκειται για ειδική περίπτωση ανεπίβλεπτων μεθόδων μοντέλου δεδομένων.
- Η τετριμμένη λύση είναι η εύρεση μίας ομαδοποίησης και η σήμανση των ομάδων μικρού μεγέθους ως ανωμαλιών.
  - Απαιτείται ένα κατώφλι πλήθους προτύπων ανά ομάδα.
- Εναλλακτικά, μπορεί να κατωφλιωθεί ένα μέτρο αξιολόγησης του κατά πόσον κάθε πρότυπο $\mathbf{x}_i$ είναι κοντά στην ομάδα του.
  - Παραδείγματα: απόσταση από πρωτότυπο, πιθανότητα σε περίπτωση ασαφούς ομαδοποίησης, βελτίωση σε μία αντικειμενική συνάρτηση (π.χ., ολικό SSE) αν το $\mathbf{x}_i$ αφαιρεθεί από το σύνολο δεδομένων.
  - Η απόσταση από το πρωτότυπο μπορεί να κανονικοποιηθεί, μέσω διαίρεσης με τη διασπορά της ομάδας, σε περίπτωση ομάδων διαφορετικής πυκνότητας.

Μέθοδοι ομαδοποίησης

- Μέθοδοι ομαδοποίησης όπως ο DBSCAN έχουν ενσωματωμένη τη λειτουργικότητα ανίχνευσης ανωμαλιών.
  - Ανωμαλίες είναι όσα πρότυπα δεν εντάσσονται σε καμία ομάδα.

- Αλγόριθμοι όπως ο GMM μπορούν να αξιοποιηθούν για ομαδοποίηση.
  - Κάθε επιμέρους κατανομή της μείξης είναι μία ομάδα.
    - **Παραγωγική ομαδοποίηση**, αντί για ιεραρχική ή διαμεριστική.
  - Η πυκνότητα πιθανότητας κάθε προτύπου αξιοποιείται ως αντίστροφο σκορ ανωμαλίας.

- Σε αλγορίθμους όπως οι k-μέσοι, το ζητούμενο πλήθος ομάδων επηρεάζει πολύ την ποιότητα της ανίχνευσης ανωμαλιών.

- Η ίδια η παρουσία των ανωμαλιών επηρεάζει την ομαδοποίηση.

Ανίχνευση ανωμαλιών

Ανίχνευση ανωμαλιών

Μέθοδοι ομαδοποίησης

- Απλές λύσεις:
  - Επαναλαμβάνουμε την ομαδοποίηση εκ νέου για διαφορετικά k.
  - Χρησιμοποιούμε υψηλό k, ώστε να βρεθούν πολλές, μικρές, συμπαγείς ομάδες και να ξεχωρίζουν πιο ευδιάκριτα οι ανωμαλίες.
  - Ομαδοποιούμε το σύνολο δεδομένων, εντοπίζουμε και αφαιρούμε τις ανωμαλίες, ομαδοποιούμε εκ νέου και αξιολογούμε ξανά τις ήδη αφαιρεθείσες ανωμαλίες.

Thank you for your attention!

Q & A

Contact: imademlis@aueb.gr