Πυκνωτική ομαδοποίηση

Ιωάννης Μαδεμλής

Πυκνωτική ομαδοποίηση

- Μία διαφορετική προσέγγιση στο πρόβλημα της ομαδοποίησης είναι η **πυκνωτική ομαδοποίηση**.
- Οι εν λόγω αλγόριθμοι είναι διαμεριστικοί και στηρίζονται στον πυκνωτικό ορισμό της ομάδας.
  - Δεν ορίζουν πρωτότυπα.
- Αναζητούν περιοχές υψηλής πυκνότητας στοιχείων, διαχωριζόμενες μεταξύ τους από περιοχές χαμηλής πυκνότητας στον χώρο των προτύπων.
  - Κάθε τέτοια περιοχή υψηλής πυκνότητας θεωρείται μία ομάδα.

Πυκνωτική ομαδοποίηση

DBSCAN

Πυκνωτική ομαδοποίηση

- Ο ορισμός της πυκνότητας είναι ένα κρίσιμο ζήτημα.
- Ο συνηθέστερος αλγόριθμος είναι ο **DBSCAN**.
- Υπολογίζει την πυκνότητα με βάση τα λεγόμενα **κεντρικά στοιχεία/πρότυπα** (core points).
- Η πυκνότητα ενός στοιχείου ορίζεται ως το πλήθος των προτύπων τα οποία απέχουν από αυτό λιγότερο από κάποιο κατώφλι ε.

Πυκνωτική ομαδοποίηση

DBSCAN

- Οι πυκνές περιοχές εμφανίζονται γύρω από κεντρικά στοιχεία στον χώρο των προτύπων.

- Τα κεντρικά στοιχεία γειτνιάζουν με κάποιο πλήθος τουλάχιστον n άλλων προτύπων, εντός μίας εμβέλειας ε από αυτά («γειτονιά»).
  - Το χρησιμοποιούμενο μέτρο εγγύτητας καθορίζεται ελεύθερα από τον χρήστη, αναλόγως με τα χαρακτηριστικά των δεδομένων.

- Το n και το ε είναι υπερπαράμετροι οι οποίες πρέπει να οριστούν χειροκίνητα, ώστε να διαμεριστεί ο χώρος των προτύπων σε πυκνές και αραιές περιοχές.

DBSCAN

- Ο DBSCAN διατρέχει ένα-ένα όλα τα στοιχεία του συνόλου δεδομένων και αναθέτει στο καθένα μία από τρεις δυνατές ετικέτες:
  - Κεντρικό (core).
  - Συνοριακό (border).
  - Θορύβου (noise).

Πυκνωτική ομαδοποίηση

- Αρχικώς, τα κεντρικά στοιχεία ορίζουν το καθένα από μία πυκνή περιοχή τουλάχιστον n στοιχείων.
  - Στη συνέχεια, οι μεταξύ τους επικαλυπτόμενες πυκνές περιοχές συγχωνεύονται σε μία ομάδα.

DBSCAN

Πυκνωτική ομαδοποίηση

- Τα συνοριακά πρότυπα δεν είναι κεντρικά, αλλά τοποθετούνται στην ομάδα πυκνότητας/γειτονιά κάποιου κεντρικού.
  - Ένα συνοριακό πρότυπο γειτνιάζει με ένα κεντρικό, αλλά το ίδιο έχει λιγότερους από n γείτονες.
  - Δύο πρότυπα γειτνιάζουν όταν η μεταξύ τους απόσταση είναι μικρότερη του ε.
- Στοιχεία τα οποία δεν ανήκουν σε καμία ομάδα/πυκνή περιοχή, αγνοούνται ως θόρυβος.
  - Έτσι, ο αλγόριθμος είναι εγγενώς ανθεκτικός στις ανωμαλίες και στον θόρυβο.
  - Εκτελεί *μερική ομαδοποίηση*, επιλέγοντας μόνος του ποια στοιχεία θα μείνουν εκτός των παραγόμενων ομάδων.

DBSCAN

Πυκνωτική ομαδοποίηση

- Ο βασικός αλγόριθμος είναι ο εξής:
  - 1: **Κατηγοριοποίηση** όλων των στοιχείων ως κεντρικών, συνοριακών ή θορύβου.
  - 2: **Εξάλειψη** σημείων θορύβου.
  - 3: **Συγχώνευση** των ομάδων που ορίζονται από κεντρικά σημεία τοποθετημένα το ένα στη γειτονιά του άλλου.
  - 4: **Ανάθεση** κάθε συνοριακού στοιχείου στην ομάδα της γειτονιάς όπου τοποθετείται.
- Στο βήμα 4, τι συμβαίνει αν ένα συνοριακό στοιχείο εμπεριέχεται ταυτοχρόνως σε γειτονιές διαφορετικών κεντρικών προτύπων;
  - Ανάθεσή του στην ομάδα του περισσότερο όμοιου με αυτό κεντρικού στοιχείου.

Πυκνωτική ομαδοποίηση

DBSCAN

- Σε μία αφελή υλοποίηση, η υπολογιστική πολυπλοκότητα του DBSCAN είναι O(N^2), όπου N το ολικό πλήθος προτύπων.

- Ωστόσο, για δεδομένα χαμηλής διαστατικότητας, βελτιστοποιημένες υλοποιήσεις και δομές δεδομένων μειώνουν την πολυπλοκότητα σε O($N \log N$).

- Βασικό πλεονέκτημα: ο αλγόριθμος δεν απαιτεί να ορίσουμε άμεσα το πλήθος k των απαιτούμενων ομάδων!

Πυκνωτική ομαδοποίηση

DBSCAN

- Ωστόσο, οι δύο βασικές υπερπαράμετροι n και ε καθορίζουν *εμμέσως* το τι πυκνότητας ομάδες προσπαθεί να ανακαλύψει ο αλγόριθμος.

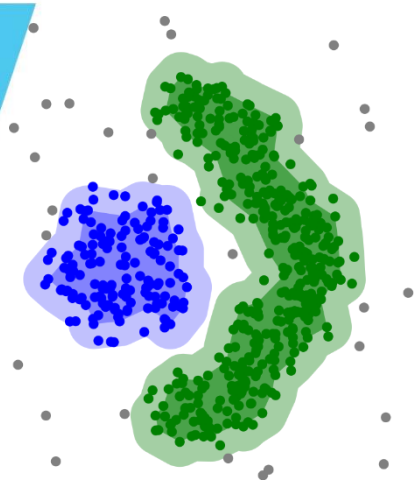
- Επομένως, από τις τιμές των υπερπαραμέτρων αυτών εξαρτάται το τελικό πλήθος των ανακαλυφθέντων ομάδων.

- Ως αποτέλεσμα, ο DBSCAN δεν μπορεί να εντοπίσει εύκολα/ικανοποιητικά ομάδες διαφορετικής πυκνότητας.
  - Αυτές θα απαιτούσαν διαφορετικές τιμές υπερπαραμέτρων για να ανακαλυφθούν.

DBSCAN

- Επίσης, δεν είναι αποδοτικός σε δεδομένα υψηλής διαστατικότητας.
  - Η πυκνότητα προτύπων ενός διανυσματικού χώρου (πλήθος προτύπων ανά μονάδα όγκου) τείνει εκθετικά προς το μηδέν όσο αυξάνεται η διαστατικότητα του χώρου.
- Ωστόσο, ο DBSCAN:
  - Είναι **ντετερμινιστικός** στη βασική του υλοποίηση.
  - Είναι εγγενώς **ανθεκτικός** στον θόρυβο και σε ανωμαλίες.
  - Μπορεί να χειριστεί με άνεση **ομάδες διαφορετικών μεγεθών και αυθαίρετων σχημάτων** (μη ελλειψοειδείς, μη υπερσφαιρικές).

Πυκνωτική ομαδοποίηση



Πηγή: *Wikipedia*.

Πυκνωτική ομαδοποίηση

DBSCAN

- Τι συμβαίνει αν δώσουμε στην υπερπαράμετρο ε μία τιμή η οποία δεν ταιριάζει στη φυσική διασπορά και διαμέριση των προτύπων;
  - Ένα ε υπερβολικά *χαμηλό* για το συγκεκριμένο σύνολο δεδομένων, οδηγεί σε αγνόηση των περισσότερων προτύπων ως θόρυβο/ανωμαλίες.
  - Ένα ε υπερβολικά *υψηλό* για το συγκεκριμένο σύνολο δεδομένων, οδηγεί σε συγχώνευση των περισσότερων ομάδων σε μία.
- Νεότερες παραλλαγές του DBSCAN προσπαθούν να λύσουν αυτά τα ζητήματα.
  - Π.χ., OPTICS.

Thank you for your attention!

Q & A

Contact: imademlis@aueb.gr