Ιεραρχική ομαδοποίηση

Ιωάννης Μαδεμλής

Συσσωρευτικοί και διαιρετικοί αλγόριθμοι

Ιεραρχική Ομαδοποίηση

- Οι αλγόριθμοι ιεραρχικής ομαδοποίησης κατασκευάζουν ένα **δενδρόγραμμα**.
  - Αντιδιαστέλλονται με τους διαμεριστικούς.
  <
- Διακρίνονται σε συσσωρευτικούς (agglomerative) και σε διαιρετικούς (divisive).

- Στους **συσσωρευτικούς** αλγορίθμους, αρχικώς όλα τα στοιχεία είναι μεμονωμένες ομάδες και διαδοχικά συγχωνεύονται ώσπου τελικά να μείνει μόνο μία καθολική ομάδα.

- Στους **διαιρετικούς** αλγορίθμους, αρχικώς όλα τα στοιχεία περιέχονται σε μία ομάδα η οποία διαδοχικά διασπάται.

Ιεραρχική Ομαδοποίηση

Συσσωρευτικοί και διαιρετικοί αλγόριθμοι

- Κάθε βήμα κάποιου ιεραρχικού αλγορίθμου απεικονίζεται σε ένα επίπεδο δενδρογράμματος.

- Το τελικό δένδρο οπτικοποιεί:
  - Τις σχέσεις ομάδων-υποομάδων,
  - Την αλληλουχία με την οποία πραγματοποιήθηκαν οι συγχωνεύσεις/διασπάσεις.

- Και τα δύο αυτά χαρακτηριστικά είναι σημαντικά για την τελική ιεραρχία.

Ιεραρχική Ομαδοποίηση

Συσσωρευτική ομαδοποίηση

- Ο βασικός συσσωρευτικός αλγόριθμος στηρίζεται στον υπολογισμό και συνεχή ενημέρωση του **πίνακα εγγύτητας**.
- Πρόκειται περί ενός *τετραγωνικού* πίνακα ομοιοτήτων/ανομοιοτήτων.
  - Περιέχει την εγγύτητα *κάθε ομάδας σε κάθε άλλη ομάδα* του συνόλου δεδομένων.
- Στην έναρξη του αλγορίθμου, το πλήθος των ομάδων ισούται με το ολικό πλήθος N των προτύπων.
- Ο αλγόριθμος κατασκευάζει επαναληπτικά ένα δενδρόγραμμα, με φορά από τα φύλλα προς τη ρίζα.

Ιεραρχική Ομαδοποίηση

Συσσωρευτική ομαδοποίηση

- Ο βασικός συσσωρευτικός αλγόριθμος είναι ο εξής:
  - 1: Υπολογισμός του πίνακα εγγύτητας.
  - 2: Επανάλαβε
  - 3: { Συγχώνευση των δύο πλησιέστερων ομάδων σε μία νέα ομάδα.
  - 4: Ενημέρωση του πίνακα εγγύτητας, ώστε να περιέχει την εγγύτητα της νέας ομάδας σε όλες τις άλλες, αντί για την εγγύτητα των μεμονωμένων στοιχείων της.
  - 5: }
  - 6: Ώσπου να παραμείνει μόνο μία ομάδα.

Ιεραρχική Ομαδοποίηση

Εγγύτητα

- Ο ανωτέρω αλγόριθμος μπορεί να μετατραπεί σε διαιρετικό με τετριμμένες μόνο τροποποιήσεις.
  - Αντίστροφη φορά κατασκευής του δενδρογράμματος.
  - Μία διάσπαση αντί για μία συγχώνευση σε κάθε επανάληψη.
- Ο πιο κρίσιμος παράγοντας για τα αποτελέσματα του αλγορίθμου είναι ο καθορισμός της έννοιας της *εγγύτητας* των ομάδων.
  - Είτε κάποιο μέτρο ομοιότητας είτε κάποιο μέτρο ανομοιότητας.
- Σε περίπτωση ομαδοποίησης πρωτοτύπων ως εγγύτητα δύο ομάδων συνήθως ορίζεται η εγγύτητα των πρωτοτύπων τους.

Κριτήριο του Ward

- Εναλλακτικά, μπορεί να χρησιμοποιηθεί το κριτήριο του Ward ως μέτρο ανομοιότητας μεταξύ δύο ομάδων.
- Η ανομοιότητα Ward δύο ομάδων ορίζεται ως η αύξηση στην ολική αθροιστική διασπορά (ολικό SSE) αν οι δύο ομάδες συγχωνευτούν.
- Επομένως, αν η επίμαχη αύξηση είναι μεγάλη και θετική, τότε οι δύο εν λόγω ομάδες είναι σημαντικά ανόμοιες.
- Η χρήση του κριτηρίου του Ward είναι το ιεραρχικό ανάλογο των k-μέσων.
  - Δεν αναζητείται όμως κάποιο τοπικό ελάχιστο του ολικού SSE.
  - Τα πρωτότυπα αξιοποιούνται μόνον εμμέσως, κατά τον υπολογισμό του SSE.

Ιεραρχική Ομαδοποίηση

Εγγύτητα ομάδων χωρίς πρωτότυπα

- Αν δεν θεωρούμε ομάδες πρωτοτύπων, δεν μπορεί να μετρηθεί η απόσταση μεταξύ δύο ομάδων ως η απόσταση των πρωτοτύπων τους.
- Τότε υπάρχουν τρεις δυνατές εκδοχές για τον ορισμό της εγγύτητας μεταξύ δύο ομάδων:
  - Η εγγύτητα των πλησιέστερων στοιχείων τους (προσέγγιση **single-link**).
  - Η εγγύτητα των πιο απομακρυσμένων τους στοιχείων (**complete-link**).
  - Ο μέσος όρος της εγγύτητας όλων των στοιχείων κάθε ομάδας με όλα τα στοιχεία της άλλης ομάδας ανά δύο (**group-average**).
- Σε κάθε περίπτωση, στο Βήμα 3 του αλγορίθμου συγχωνεύονται οι δύο πιο όμοιες ομάδες.

Ιεραρχική Ομαδοποίηση

Ιεραρχική ομαδοποίηση χωρίς πρωτότυπα

- Η ιεραρχική ομαδοποίηση είναι κατάλληλη για προβλήματα όπου το σύνολο δεδομένων παρουσιάζει μία εγγενή ιεραρχική διαμεριστική δομή.
  - Η εκδοχή single-link είναι κατάλληλη για δεδομένα με φυσικό διαμερισμό σε ομάδες αυθαίρετου σχήματος (*μη ελλειψοειδείς*).
  - Η εκδοχή complete-link είναι λιγότερο ευάλωτη σε ανωμαλίες, αλλά κατάλληλη περισσότερο για δεδομένα με φυσικό διαμερισμό σε *υπερσφαιρικές ομάδες*.
  - Όταν ως μέτρο εγγύτητας χρησιμοποιείται η τετραγωνική ευκλείδεια απόσταση, η εκδοχή group-average αποδεικνύεται ότι ισοδυναμεί με *χρήση του κριτηρίου του Ward σε ομάδες πρωτοτύπων*.

Ιεραρχική Ομαδοποίηση

Πλεονεκτήματα

- Χαρακτηριστικά της ιεραρχικής ομαδοποίησης είναι πως **αποφεύγεται** η διατύπωση του προβλήματος ως προβλήματος βελτιστοποίησης και οι πως σχετικοί αλγόριθμοι είναι **άπληστοι**.
  - Προβαίνουν σε τοπικά βέλτιστες ως προς τη συνοχή των ομάδων διασπάσεις/συγχωνεύσεις ανά επανάληψη.
- Έτσι, η ιεραρχική ομαδοποίηση:
  - Είναι απλή και αποτελεσματική.
  - Δεν επιχειρεί να λύσει προσεγγιστικά κάποιο δυσεπίλυτο πρόβλημα βελτιστοποίησης.
  - Δεν μπορεί να παγιδευτεί σε τοπικά ελάχιστα κάποιας συνάρτησης κόστους.
  - Δεν εξαρτάται από την αρχικοποίηση, όπως οι επαναληπτικοί αλγόριθμοι βελτιστοποίησης.

Ιεραρχική Ομαδοποίηση

Μειονεκτήματα

- Έχει υψηλές υπολογιστικές απαιτήσεις σε χρόνο και μνήμη, εξαιτίας της δημιουργίας και συνεχούς ενημέρωσης του πίνακα εγγύτητας.
- Το γεγονός ότι η ομαδοποίηση σε κάποιο επίπεδο του δενδρογράμματος δεν μπορεί να αλλάξει στο επόμενο επίπεδο και θεωρείται δεδομένη, σημαίνει πως όλες οι επιμέρους αποφάσεις ομαδοποίησης είναι οριστικές.
  - Προβληματικό σε ειδικές περιπτώσεις (π.χ., υψηλός θόρυβος, πολλαπλές ανωμαλίες).

Μέτρα αξιολόγησης

Εσωτερική αξιολόγηση

- Στην περίπτωση της ιεραρχικής ομαδοποίησης, μία συνηθισμένη μετρική εσωτερικής αξιολόγησης είναι η **συφαινετική συσχέτιση** (cophenetic correlation).

- Στηρίζεται στην έννοια της **συφαινετικής απόστασης** μεταξύ δύο προτύπων $\mathbf{x}_i$ και $\mathbf{x}_j$.
  - Είναι μία τιμή ανομοιότητας μεταξύ δύο διαφορετικών ομάδων, όπου ανήκουν τα $\mathbf{x}_i$ και $\mathbf{x}_j$.
  - Είναι η τιμή ανομοιότητας κατά τη στιγμή όπου ένας συσσωρευτικός ιεραρχικός αλγόριθμος τοποθετεί τα $\mathbf{x}_i$ και $\mathbf{x}_j$ στην ίδια ομάδα (μέσω συγχώνευσης των δύο εν λόγω ομάδων) για πρώτη φορά, κατά τη διαδικασία κατασκευής του δενδρογράμματος.

Μέτρα αξιολόγησης

Εσωτερική αξιολόγηση

- Επομένως, τι είναι η συφαινετική απόσταση μεταξύ ενός ετερογενούς ζεύγους προτύπων δύο ομάδων, οι οποίες συγχωνεύονται σε κάποια επανάληψη του αλγορίθμου;
  - Η ελάχιστη απόσταση μεταξύ οποιωνδήποτε δύο στοιχείων των δύο ομάδων.
  - Άρα όλα τα στοιχεία της μίας ομάδας έχουν την ίδια συφαινετική απόσταση με όλα τα στοιχεία της άλλης ομάδας.
    - Είναι η απόσταση των δύο μεταξύ τους πλησιέστερων προτύπων των δύο ομάδων.

- Οι αποστάσεις αυτές εξαρτώνται από τη συγκεκριμένη **διαδοχή συγχωνεύσεων** κατά την κατασκευή του δενδρογράμματος.

Μέτρα αξιολόγησης

Εσωτερική αξιολόγηση

- Αν υπολογιστούν οι εν λόγω αποστάσεις για όλα τα N πρότυπα του συνόλου δεδομένων μας, σχηματίζουν έναν τετραγωνικό πίνακα διάστασης N x N ο οποίος καλείται **πίνακας συφαινετικών αποστάσεων**.
- Παρομοίως, έχουμε ήδη υπολογισμένο τον απλό πίνακα ανομοιοτήτων μεταξύ όλων των στοιχείων του συνόλου δεδομένων μας.
  - Επίσης διάστασης N x N.
- Διανυσματοποιούμε αυτούς τους δύο πίνακες.
- Ο βαθμωτός **συντελεστής συφαινετικής συσχέτισης** της τελικής ιεραρχικής ομαδοποίησης είναι η συσχέτιση μεταξύ των δύο διανυσμάτων.
- Μεγαλύτερη τιμή συσχέτισης υποδηλώνει περισσότερο εύλογη ομαδοποίηση.

Thank you for your attention!

Q & A

Contact: imademlis@aueb.gr