Τμήμα Πληροφορικής με Εφαρμογές στη Βιοϊατρική
Σχολή Θετικών Επιστημών
Πανεπιστήμιο Θεσσαλίας

ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ

Ομαδοποίηση - Μέτρα Εγκύτητας

Αριστείδης Γ. Βραχάτης, Dipl-Ing, M.Sc, PhD
Adjunct Lecturer

Εγκυρότητα και αποτίμηση συσταδοποίησης

- Η αποτίμηση συσταδοποίησης επιδιώκει να αποτιμήσει την καταλληλότητα ή ποιότητα της συσταδοποίησης· η σταθερότητα συσταδοποίησης έχει ως στόχο
  - να κατανοήσει την ευαισθησία που εμφανίζει το αποτέλεσμα της συσταδοποίησης σε διάφορες αλγοριθμικές παραμέτρους, π.χ. το πλήθος των συστάδων· και
  - η τάση συσταδοποίησης αξιολογεί το κατά πόσο θα έπρεπε εξαρχής να εφαρμοστεί η συσταδοποίηση, δηλαδή το αν τα δεδομένα εμφανίζουν οποιαδήποτε εγγενή δομή ομαδοποίησης.

- Τα μέτρα της εγκυρότητας μπορούν να χωριστούν σε τρεις κύριους τύπους:
  - ***Εξωτερικά***: Τα εξωτερικά μέτρα εγκυρότητας χρησιμοποιούν κριτήρια που δεν είναι εγγενή για το σύνολο δεδομένων, π.χ. ετικέτες κατηγορίας.

  - ***Εσωτερικά***: Τα εσωτερικά μέτρα εγκυρότητας στηρίζονται σε κριτήρια που προκύπτουν από τα ίδια τα δεδομένα, π.χ. μετρικές απόστασης εντός της ίδιας συστάδας ή μεταξύ διαφορετικών συστάδων.

  - ***Σχετικά***: Τα σχετικά μέτρα εγκυρότητας επιδιώκουν να συγκρίνουν ευθέως διαφορετικές συσταδοποιήσεις, συνήθως εκείνες που προκύπτουν από διαφορετικές ρυθμίσεις των παραμέτρων του ίδιου αλγορίθμου.

Εξωτερικά μέτρα

- Τα εξωτερικά μέτρα υποθέτουν ότι είναι γνωστή εκ των προτέρων η σωστή συσταδοποίηση (δηλαδή εκείνη που αντιστοιχεί στη δεδομένη αλήθεια), η οποία και χρησιμοποιείται για την αποτίμηση μιας δεδομένης συσταδοποίησης.

- Έστω ότι $D = \{X_i\}_{i=1}^{n}$ είναι ένα σύνολο δεδομένων που αποτελείται από n σημεία σε έναν d-διάστατο χώρο, το οποίο έχει διαμεριστεί σε k συστάδες.

- Έστω ότι συμβολίζουμε με $y_i \in \{1,2,\dots,k\}$ τις πληροφορίες συμμετοχής στις συστάδες (ή των ετικετών των συστάδων) που αντιστοιχούν στη δεδομένη αλήθεια για κάθε σημείο.

- Η συσταδοποίηση που αντιστοιχεί στη δεδομένη αλήθεια ορίζεται ως $T = \{T_1, T_2, \dots, T_k\}$, όπου η συστάδα T_j αποτελείται από όλα τα σημεία με ετικέτα j, δηλαδή $T_j = \{x_i \in D | y_i = j\}$.

- Για λόγους σαφήνειας, θα αναφερόμαστε στη συσταδοποίηση T ως τον διαμερισμό που αντιστοιχεί στη δεδομένη αλήθεια, και σε κάθε συστάδα T_i ως διαμέριση.

Εξωτερικά μέτρα

– Τα εξωτερικά μέτρα αποτίμησης προσπαθούν να αποτυπώσουν τον βαθμό στον οποίο τα σημεία από την ίδια διαμέριση εμφανίζονται στην ίδια συστάδα, καθώς και τον βαθμό στον οποίο τα σημεία από διαφορετικές διαμερίσεις ομαδοποιούνται σε διαφορετικές συστάδες.

– Όλα τα εξωτερικά μέτρα στηρίζονται στον πίνακα συνάφειας N, διαστάσεων r x k, ο οποίος επάγεται από μια συσταδοποίηση C και τον διαμερισμό T που αντιστοιχεί στη δεδομένη αλήθεια. Ο πίνακας συνάφειας ορίζεται ως εξής:

$$\mathbf{N}(i, j) = n_{ij} = \left| C_i \cap T_j \right|$$

– Η καταμέτρηση n_{ij} αναπαριστά το πλήθος των σημείων που ανήκουν τόσο στη συστάδα C_i όσο και στη διαμέριση T_j της δεδομένης αλήθειας.

– Έστω ότι το $n_i = |C_i|$ είναι το πλήθος των σημείων που ανήκουν στη συστάδα C_i, και το $m_j = |T_j|$ είναι το πλήθος των σημείων που ανήκουν στη διαμέριση T_j.

– Ο πίνακας συνάφειας μπορεί να υπολογιστεί από τον διαμερισμό T και τη συσταδοποίηση C:
  – Για κάθε σημείο $x_i \in D$ εξετάζεται η ετικέτα y_i της διαμέρισης και η ετικέτα της συστάδας

Μέτρα που βασίζονται στο ταίριασμα: Καθαρότητα

– Η καθαρότητα ποσοτικοποιεί τον βαθμό στον οποίο μια συστάδα C_i περιέχει οντότητες από μία μόνο διαμέριση:

$$purity_i = \frac{1}{n_i} \max_{k}^{j=1} \{n_{ij}\}$$

– Η καθαρότητα της συσταδοποίησης C ορίζεται ως το σταθμισμένο άθροισμα των τιμών καθαρότητας ανά συστάδα:

$$purity = \sum_{i=1}^{r} \frac{n_i}{n} purity_i = \frac{1}{n} \sum_{i=1}^{r} \max_{k}^{j=1} \{n_{ij}\}$$

– όπου ο λόγος n_i/n αντιπροσωπεύει το ποσοστό των σημείων που ανήκουν στη συστάδα C_i.

Μέτρα που βασίζονται στο ταίριασμα: Μέγιστο ταίριασμα

- Το μέτρο του μέγιστου ταιριάσματος επιλέγει εκείνη την αντιστοίχιση μεταξύ συστάδων και διαμερίσεων για την οποία μεγιστοποιείται το άθροισμα του πλήθους των κοινών σημείων (n_{ij}), με την προϋπόθεση ότι μόνο μία συστάδα μπορεί να ταιριάζει με μια καθορισμένη διαμέριση.

- Έστω ότι G είναι ένα διχοτομήσιμο γράφημα για το σύνολο κορυφών $V = C \cup T$, και έστω ότι το σύνολο ακμών είναι $\mathrm{E} = \{(C_i, T_j)\}$ με βάρη $\mathrm{w}(C_i, T_j) = n_{ij}$.

- Ένα ταίριασμα M στο γράφημα G είναι υποσύνολο του E, τέτοιο ώστε οι ακμές του M να είναι μη γειτονικές ανά ζεύγη, δηλαδή να μην έχουν κοινή κορυφή.

- Το ταίριασμα μέγιστου βάρους στο γράφημα G ορίζεται ως εξής:

$$match = \arg\max_{M} \left\{ \frac{w(M)}{n} \right\}$$

- όπου το βάρος ενός ταιριάσματος M είναι απλώς το άθροισμα των βαρών όλων των ακμών του M, και δίνεται από τη σχέση

$$w(M) = \sum_{e \in M} w(e)$$

Μέτρα που βασίζονται στο ταίριασμα: F-μέτρο

– Για μια συστάδα C_i, έστω ότι συμβολίζουμε με j_i τη διαμέριση που περιέχει το μέγιστο πλήθος σημείων από τη C_i, δηλαδή

$$j_i = \max_{j=1}^{k} \{n_{ij}\}$$

– Η ακρίβεια μιας συστάδας C_i είναι ίδια με την καθαρότητά της:

$$prec_i = \frac{1}{n_i} \max_{k}^{j=1} \left\{ n_{ij} \right\} = \frac{n_{ij_i}}{n_i}$$

– Η ανάκληση της συστάδας Ci ορίζεται ως

$$recall_i = \frac{n_{ij_i}}{|T_{j_i}|} = \frac{n_{ij_i}}{m_{j_i}}$$

– όπου $m_{ji} = |T_{ji}|$.

Μέτρα που βασίζονται στο ταίριασμα: F-μέτρο

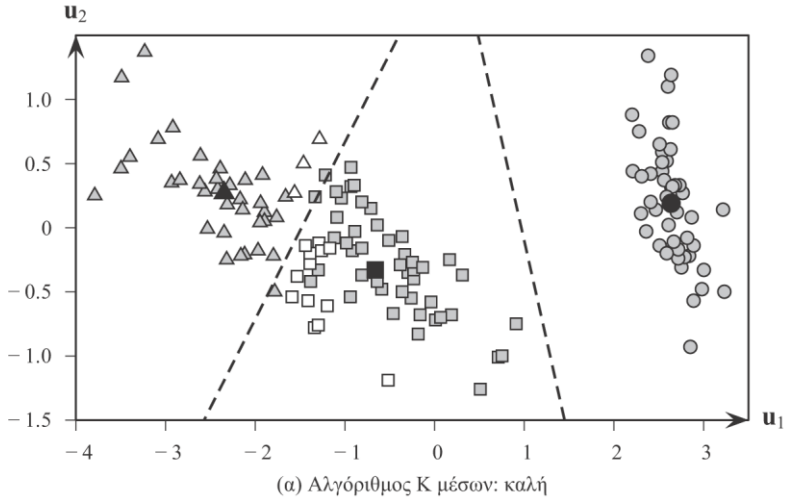
– Το F-μέτρο είναι ο αρμονικός μέσος των τιμών ακρίβειας και ανάκλησης για κάθε συστάδα C_i

$$F_i = \frac{2}{\frac{1}{prec_i} + \frac{1}{recall_i}} = \frac{2 \cdot prec_i \cdot recall_i}{prec_i + recall_i} = \frac{2\, n_{ij_i}}{n_i + m_{j_i}}$$

– Το F-μέτρο για τη συσταδοποίηση C είναι ο μέσος των τιμών του F-μέτρου ανά συστάδα

$$F = \frac{1}{r} \sum_{i=1}^{r} F_i$$

Αλγόριθμος K-means



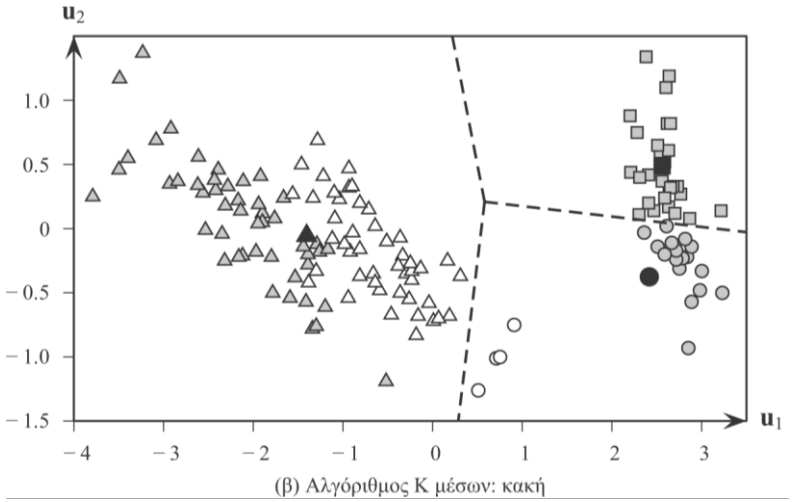
(α) Αλγόριθμος K μέσων: καλή

– Πίνακας συνάφειας:

	iris-setosa T_1	iris-versicolor T_2	iris-virginica T_3	n_i
C_1 (τετράγωνα)	0	47	14	61
C_2 (κύκλοι)	50	0	0	50
C_3 (τρίγωνα)	0	3	36	39
m_j	50	50	50	$n = 150$

purity = ???, match = ???, F = ???

Αλγόριθμος K-means



(β) Αλγόριθμος K μέσων: κακή

– Πίνακας συνάφειας:

	iris-setosa T_1	iris-versicolor T_2	iris-virginica T_3	n_i
C_1 (τετράγωνα)	30	0	0	30
C_2 (κύκλοι)	20	4	0	24
C_3 (τρίγωνα)	0	46	50	96
m_j	50	50	50	$n = 150$

purity = ??? , match = ???, F = ???

Μέτρα ανά ζεύγη

- Δίνεται η συσταδοποίηση C και ο διαμερισμός T που αντιστοιχεί στη δεδομένη αλήθεια· έστω ότι xi, xj ∈ D είναι δύο οποιαδήποτε σημεία, με i ≠ j.

- Αν τόσο το xi όσο και το xj ανήκουν στην ίδια συστάδα, περιγράφουμε αυτή την κατάσταση με τον όρο θετικό συμβάν

- Αν δεν ανήκουν στην ίδια συστάδα, χρησιμοποιούμε τον όρο αρνητικό συμβάν.

- Ανάλογα με το αν οι ετικέτες των συστάδων συμφωνούν με τις ετικέτες των διαμερίσεων, υπάρχουν τέσσερα ενδεχόμενα που πρέπει να ληφθούν υπόψη:

Μέτρα ανά ζεύγη

- Αληθώς θετικά (TP): Τα σημεία xi και xj ανήκουν στην ίδια διαμέριση του διαμερισμού T, αλλά και στην ίδια συστάδα της συσταδοποίησης C.

- Ψευδώς αρνητικά (FN): Τα σημεία xi και xj ανήκουν στην ίδια διαμέριση του T, αλλά όχι και στην ίδια συστάδα της C.

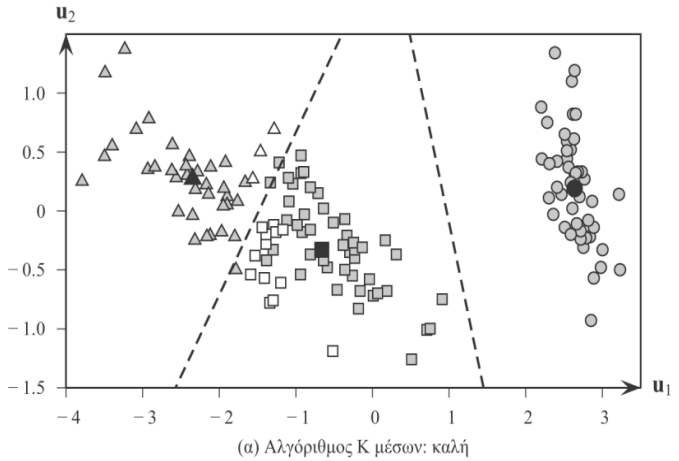
- Ψευδώς θετικά (FP): Τα σημεία xi και xj δεν ανήκουν στην ίδια διαμέριση του T, αλλά ανήκουν στην ίδια συστάδα της C.

- Αληθώς αρνητικά (TN): Τα σημεία xi και xj δεν ανήκουν ούτε στην ίδια διαμέριση του T, ούτε στην ίδια συστάδα της C.

- Επειδή υπάρχουν $N = \binom{n}{2} = \frac{n(n-1)}{2}$ ζεύγη σημείων, προκύπτει η ακόλουθη ισότητα

$$N = TP + FN + FP + TN$$

Αλγόριθμος Κ μέσων:



(α) Αλγόριθμος Κ μέσων: καλή

Πίνακας συνάφειας:

	iris-setosa	**iris-versicolor**	**iris-virginica**
	T_1	T_2	T_3
C_1	0	47	14
C_2	50	0	0
C_3	0	3	36

- Το πλήθος των αληθώς θετικών ζευγών είναι: $TP = \binom{47}{2} + \binom{14}{2} + \binom{50}{2} + \binom{3}{2} + \binom{36}{2} = 3030$

- Ομοίως, βρίσκουμε ότι FN = ???, FP = ???, TN = ???

Μέτρα ανά ζεύγη: Συντελεστής Jaccard, στατιστικό Rand, μέτρο FM

– Συντελεστής Jaccard: Μετρά το ποσοστό των αληθώς θετικών ζευγών (σημείων), αφού όμως πρώτα αγνοήσει τα αληθώς αρνητικά ζεύγη.

$$Jaccard = \frac{TP}{TP + FN + FP}$$

– Στατιστικό Rand: Μετρά το ποσοστό των αληθώς θετικών και αληθώς αρνητικών ζευγών για όλα τα ζεύγη σημείων.

$$Rand = \frac{TP + TN}{N}$$

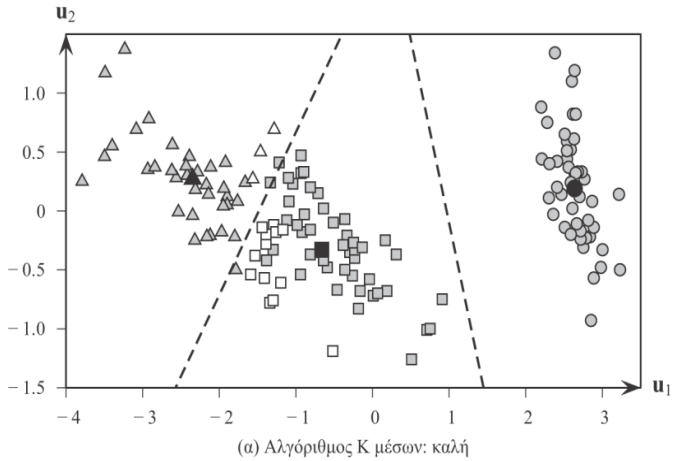
– Μέτρο των Fowlkes-Mallows: Ορίζουμε τη συνολική ακρίβεια ανά ζεύγη και ανάκληση ανά ζεύγη για μια συσταδοποίηση C, όπως φαίνεται παρακάτω

$$prec = \frac{TP}{TP + FP} \qquad recall = \frac{TP}{TP + FN}$$

– Το μέτρο των Fowlkes-Mallows (FM) ορίζεται ως ο γεωμετρικός μέσος της ακρίβειας ανά ζεύγη και της ανάκλησης ανά ζεύγη

$$FM = \sqrt{prec \cdot recall} = \frac{TP}{\sqrt{(TP + FN)(TP + FP)}}$$

Αλγόριθμος Κ μέσων:



(α) Αλγόριθμος Κ μέσων: καλή

Πίνακας συνάφειας:

	iris-setosa	**iris-versicolor**	**iris-virginica**
	T_1	T_2	T_3
C_1	0	47	14
C_2	50	0	0
C_3	0	3	36

- TP = 3030, FN = 645, FP = 766, TN = 6734
- Jaccard = ???
- Rand = ???
- FM = ???

Εσωτερικά μέτρα

- Τα εσωτερικά μέτρα αποτίμησης δεν καταφεύγουν στον διαμερισμό που αντιστοιχεί στη δεδομένη αλήθεια.

- Επομένως, για να αποτιμήσουν την ποιότητα της συσταδοποίησης, πρέπει να στηριχθούν σε έννοιες της «ενδοσυσταδικής» ομοιότητας ή πυκνότητας, σε αντίθεση με έννοιες της «διασυσταδικής» απόστασης
  - συνήθως, πρέπει να επιτευχθεί ένα συμβιβασμός όσον αφορά τη μεγιστοποίηση αυτών των δύο στόχων.

- Τα εσωτερικά μέτρα βασίζονται στη μήτρα αποστάσεων (που είναι επίσης γνωστή ως μήτρα εγγύτητας), διαστάσεων n x n, όλων των αποστάσεων ανά ζεύγη για τα n σημεία:

$$\mathbf{W} = \left\{ \delta(\mathbf{x}_i, \mathbf{x}_j) \right\}_{i,j=1}^{n}$$

Συντελεστής περιγράμματος

– Ορίζουμε τον συντελεστή περιγράμματος ή σιλουέτας ενός σημείου x_i ως

$$s_i = \frac{\mu_{out}^{\min}(\mathbf{x}_i) - \mu_{in}(\mathbf{x}_i)}{\max\left\{\mu_{out}^{\min}(\mathbf{x}_i), \mu_{in}(\mathbf{x}_i)\right\}}$$

– όπου $\mu_{in}(x_i)$ είναι η μέση απόσταση του x_i από σημεία της συστάδας $\hat{y}_i$ στην οποία ανήκει:

$$\mu_{in}(\mathbf{x}_i) = \frac{\sum_{\mathbf{x}_j \in C_{\hat{y}_i}, j \neq i} \delta(\mathbf{x}_i, \mathbf{x}_j)}{n_{\hat{y}_i} - 1}$$

– και $\mu_{out}^{min}(x_i)$ είναι ο μέσος των αποστάσεων του x_i από σημεία της πλησιέστερης συστάδας:

$$\mu_{out}^{\min}(\mathbf{x}_i) = \min_{j \neq \hat{y}_i} \left\{ \frac{\sum_{\mathbf{y} \in C_j} \delta(\mathbf{x}_i, \mathbf{y})}{n_j} \right\}$$

Συντελεστής περιγράμματος

- Η τιμή si ενός σημείου ανήκει στο διάστημα [−1, +1].
- Μια τιμή πλησίον του +1 υποδεικνύει ότι το x_i βρίσκεται πολύ πιο κοντά σε σημεία της δικής του συστάδας· μια τιμή πλησίον του μηδενός δείχνει ότι το x_i βρίσκεται κοντά στο όριο μεταξύ δύο συστάδων
- Μια τιμή πλησίον του −1 υποδεικνύει ότι το x_i βρίσκεται πολύ πιο κοντά σε κάποια άλλη συστάδα, με συνέπεια να είναι ορατό το ενδεχόμενο να αντιστοιχιστεί σε λάθος συστάδα.
- Ο συντελεστής περιγράμματος ορίζεται ως η μέση τιμή SC για όλα τα σημεία.

$$SC = \frac{1}{n}\sum_{i=1}^{n} s_i$$

- Μια τιμή πλησίον του +1 αποτελεί ένδειξη καλής συσταδοποίησης.